RESEARCH INSTITUTE OF NATURAL SCIENCES

제 32 호

2025 자연과학연구소 연구결과백서

2025. 03 – 2026. 02

숙명여자대학교
자연과학연구소

RESEARCH INSTITUTE OF NATURAL SCIENCES

2025 자연과학연구소 연구결과백서

2025. 03 – 2026. 02

제 32 호

숙명여자대학교
자연과학연구소

목차

(가나다 순)

수학과	서검교 교수	6
	신용현 교수	33
	신진우 교수	137
	최진원 교수	155
통계학과	김별 교수	210
	김양진 교수	237
	박호영 교수	269
	서범석 교수	291
	여인권 교수	359
화학과	김효진 교수	387
	이영준 교수	401
	이한길 교수	409

RESEARCH INSTITUTE OF NATURAL SCIENCES

2025 자연과학연구소 연구결과백서 제32호

발행인 숙명여자대학교 자연과학연구소
발행일 2026.02.28
발행처 숙명여자대학교 자연과학연구소
04310 서울 용산구 청파로 47길 100(청파동2가) 과학관 610호
RESEARCH INSTITUTE OF NATURAL SCIENCES
SookMyung Women's University 100, Cheongpa-ro 47-gil, Yongsan-gu, Seoul, Republic of Korea
Tel : 02) 710-9727
인 쇄 피오디컴퍼니, 서울 용산구 청파로47나길 7, 3F
Tel: 02) 715-4857

ISSN 2713-7929

RESEARCH INSTITUTE OF NATURAL SCIENCES

2025 자연과학연구소 연구결과백서

수학과

서검교 교수

논문

- E. Lee, K. Seo, An overdetermined Steklov eigenvalue problem on Riemannian manifolds with nonnegative Ricci curvature, Results Math. 80 (2025), no.4, Paper No. 102, 15 pp.

- D. Park, K. Seo, Harmonic minimal graphs in product spaces, J. Geom. Anal. 36 (2026), no. 2, Paper No. 59, 11 pp.

Results Math (2025) 80:102
© 2025 The Author(s)
1422-6383/25/040001-15
published online May 14, 2025
https://doi.org/10.1007/s00025-025-02422-5

Results in Mathematics

An Overdetermined Steklov Eigenvalue Problem on Riemannian Manifolds with Nonnegative Ricci Curvature

Eunjoo Lee and Keomkyo Seo

Abstract. We consider an overdetermined Steklov eigenvalue problem on a domain in a Riemannian manifold with nonnegative Ricci curvature. We prove that, given a compact connected domain Ω with nonnegative Gaussian curvature with C^2 boundary, if the first Steklov eigenfunction is a solution to the overdetermined problem, then the domain Ω is flat and the boundary $\partial\Omega$ consists of geodesics or geodesic circles. This can be regarded as a generalization of the result by Payne-Philippin [21], where they assumed that Ω is a simply-connected domain in $\mathbb{R}^2$. We also obtain a similar result for the same overdetermined problem on a higher-dimensional compact connected manifold with C^2 boundary with nonnegative Ricci curvature.

Mathematics Subject Classification. 35N25, 53C24, 58C40.

Keywords. Overdetermined problem, Steklov eigenvalue problem, Ricci curvature.

1. Introduction

The famous Serrin's theorem [28] for the overdetermined problem states as follows. Let $\Omega \subset \mathbb{R}^n$ be a smooth bounded domain. Assume that u is a smooth solution to the overdetermined problem

$$\begin{cases} \Delta u = -1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \\ \frac{\partial u}{\partial \nu} = const. = c & \text{on } \partial\Omega. \end{cases}$$

Then Ω is a ball and the solution u is radial. Here ν denotes the outward unit normal vector to $\partial\Omega$. This beautiful result has been extensively generalized in various settings (see [2,4,5,10,15–18,20,22–27] for example and references therein).

On the other hand, given an n-dimensional compact Riemannian manifold Ω with smooth boundary $\partial\Omega$, we consider the following Steklov eigenvalue problem

$$\begin{cases} \Delta u = 0 & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = \sigma u & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $\nu = \nu(x)$ denotes the outward unit normal to $\partial\Omega$. This problem has also been actively studied recently in the literature (see [7–9,11–13] for instance). One of the well-known facts is that Steklov eigenvalues form an unbounded sequence $0 = \sigma_0 < \sigma_1 \leq \sigma_2 \leq \cdots \to \infty$, where each eigenvalue σ_i is repeated according to its multiplicity. We refer the readers to two nice survey articles [14] by Girouard-Polterovich and [6] by Colbois-Girouard-Gordon-Sher for some recent results and developments on the Steklov eigenvalue problem.

In [21], Payne-Philippin considered the following overdetermined Steklov eigenvalue problem on a simply-connected domain $\Omega \subset \mathbb{R}^2$ with C^2 boundary

$$\begin{cases} \Delta u = 0 & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = \sigma u & \text{on } \partial\Omega, \\ |\nabla u| = const. = c & \text{on } \partial\Omega, \end{cases} \tag{2}$$

where $\nu = \nu(x)$ denotes the outward unit normal to $\partial\Omega$. They proved that if the first eigenfunction u corresponding to the first nonzero eigenvalue σ_1 satisfies this overdetermined Steklov eigenvalue problem, then Ω is a disk of radius $r = \frac{1}{\sigma_1}$ in $\mathbb{R}^2$. They conjectured that, for any Steklov eigenvalue σ and for $\mathbb{R}^n (n \geq 3)$, the assumption that the corresponding Steklov eigenfunction u is a solution to the overdetermined problem (2) would imply that Ω is a ball. However, Alessandrini-Magnanini [1] constructed examples of non-symmetric planar domains for which a solution to the problem (2) exists. This shows the Payne-Philippin's conjecture does not hold in general.

The overdetermined condition in (2) given by Payne-Philippin

$$|\nabla u| = c \quad \text{on } \partial\Omega$$

can be modified in the following way. Since $\langle \nabla u, \nabla u \rangle = c^2$ along $\partial\Omega$, we differentiate both sides of this equation for any tangent vector $X \in T_p(\partial\Omega)$ at $p \in \partial\Omega$ to get

$$\operatorname{Hess} u(\nabla u, X) = 0,$$

which implies that

$$\operatorname{Hess} u(\nabla u) = f\nu$$

for some function $f \in C^\infty(\partial\Omega)$. With this observation in mind, we study an overdetermined Steklov eigenvalue problem on a Riemannian manifold Ω with C^2 boundary and with nonnegative Ricci curvature as follows:

$$\begin{cases} \Delta u = 0 & \text{in } \Omega, \quad (1.1) \\ \frac{\partial u}{\partial \nu} = \sigma u & \text{on } \partial\Omega, \ (1.2) \\ \mathrm{Hess}\, u(\nabla u) = f\nu & \text{on } \partial\Omega \quad (1.3) \end{cases}$$

where $\nu = \nu(x)$ denotes the outward unit normal to $\partial\Omega$ and $f \in C^\infty(\partial\Omega)$ denotes any nonpositive function. In Section 2, we obtain a Serrin-type theorem for an overdetermined Steklov eigenvalue problem (1.1)-(1.3). To be precise, let Ω be a compact connected surface with nonnegative Gaussian curvature and with C^2 boundary and let u be a first Steklov eigenfunction with the corresponding eigenvalue σ_1. If u is a solution to the problem (1.1)-(1.3), then the Gaussian curvature of Ω is zero and $\partial\Omega$ is the disjoint union of geodesics and geodesic circles of radius $\frac{1}{\sigma_1}$ (see Theorem 2.3). In particular, if Ω is simply-connected, the existence of a solution u implies that Ω is a geodesic ball (see Corollary 2.4). In Section 3, we obtain a similar Serrin-type symmetry result for the overdetermined problem (1.1)-(1.3) on a higher-dimensional compact connected Riemannian manifold with nonnegative Ricci curvature and with C^2 boundary (see Theorem 3.1).

2. Compact surfaces with nonnegative Gaussian curvature

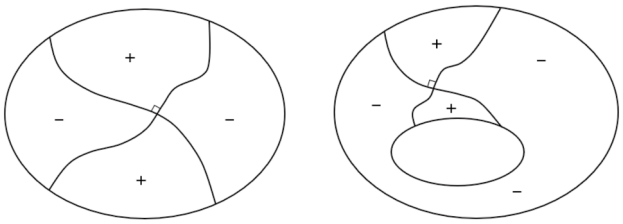
Before we state our Serrin-type theorem for an overdetermined Steklov eigenvalue problem on a compact surface with nonnegative Gaussian curvature, we demonstrate some useful lemmas. Payne-Philippin [21] proved that the first Steklov eigenfunction on a simply-connected domain in $\mathbb{R}^2$ has no critical point there. We offer a slight modification of their result as follows. The main idea of the proof remains essentially the same as in [21], but since Ω is not necessarily simply-connected, we need to be careful when defining a suitable function on the nodal domains. For completeness, we give the proof here.

Lemma 2.1. (Payne-Philippin [21]). *Let Ω be a compact connected surface with C^2 boundary. Then the first Steklov eigenfunction of (1.1) and (1.2) has no critical point in Ω.*

Proof. Let u be the first Steklov eigenfunction satisfying (1.1) and (1.2) with the corresponding first eigenvalue σ_1. Suppose that $\nabla u = 0$ at some point $p \in \Omega$. Define $v(x) = u(x) - u(p)$. Then we have

$$\begin{cases} \Delta v = 0 & \text{in } \Omega, \\ \frac{\partial v}{\partial \nu} - \sigma_1 v = \sigma_1 u(p) & \text{on } \partial\Omega. \end{cases}$$

Since $|\nabla v| = 0$ at p, it is known that the function v has at least two nodal curves intersecting at p (see [3] for example). Denote by m the number of nodal



(a) $m = 4$ (b) $m = 3$

FIGURE 1. (**a**) Four nodal domains for a simply-connected domain ($m = 4$) (**b**) Three nodal domains for a doubly-connected domain ($m = 3$)

domains. For $1 \leq k \leq m$, let Ω_k be the k-th nodal domain and let $\partial\Omega_k$ be the restriction of $\partial\Omega$ on Ω_k. Note that $m \geq 3$ since Ω is not necessarily simply-connected (see Fig. 1). We can define a continuous function w as follows:

$$w = \sum_{k=1}^{m} \alpha_k v_k$$

satisfying that

$$v_k = \begin{cases} v & \text{in } \Omega_k, \\ 0 & \text{in } \Omega \setminus \Omega_k, \end{cases}$$

where each α_k is chosen to be not all zero and satisfies the equations

$$\int_{\partial\Omega} w = 0 \quad \text{and} \quad \sum_{k=1}^{m} \alpha_k \int_{\partial\Omega_k} w = 0.$$

Observe that

$$\begin{aligned} \int_{\Omega_k} |\nabla w|^2 &= \int_{\partial\Omega_k} w \frac{\partial w}{\partial \nu} = \alpha_k^2 \int_{\partial\Omega_k} v_k \frac{\partial v_k}{\partial \nu} \\ &= \alpha_k^2 \int_{\partial\Omega_k} [u - u(p)] \frac{\partial v_k}{\partial \nu} \\ &= \sigma_1 \alpha_k^2 \int_{\partial\Omega_k} [u - u(p)] u \\ &= \sigma_1 \alpha_k^2 \int_{\partial\Omega_k} [u - u(p)]^2 + \sigma_1 u(p) \alpha_k^2 \int_{\partial\Omega_k} u - u(p). \end{aligned}$$

Adding up and using the above relations, we have

$$\begin{aligned}\int_\Omega |\nabla w|^2 &= \sigma_1 \sum_{k=1}^m \alpha_k^2 \int_{\partial\Omega_k} [u-u(p)]^2 + \sigma_1 u(p) \sum_{k=1}^m \alpha_k^2 \int_{\partial\Omega_k} u - u(p) \\ &= \sigma_1 \int_{\partial\Omega} w^2.\end{aligned}$$

However, since w has at least three nodal domains, w cannot be the first eigenfunction, which is a contradiction. □

Lemma 2.2. *Let Ω be a compact connected surface with Gaussian curvature $K_\Omega \geq 0$ and with C^2 boundary. Assume that u is a harmonic function satisfying that $\nabla u \neq 0$ in Ω. Then*

(i) $\frac{\partial}{\partial \nu} \log |\nabla u|^2 = \dfrac{2 \operatorname{Hess} u(\nabla u, \nu)}{|\nabla u|^2}$ *on* $\partial\Omega$,

(ii) $\Delta \log |\nabla u|^2 = 2K_\Omega \geq 0$ *in* Ω.

Proof. (i) can be seen by a direct calculation. For (ii), we observe that

$$\begin{aligned}\Delta \log |\nabla u|^2 &= \operatorname{div}\left(\nabla \log |\nabla u|^2\right) \\ &= 2 \operatorname{div}\left(\frac{\operatorname{Hess} u\,(\nabla u)}{\langle \nabla u, \nabla u\rangle}\right) \\ &= 2\,\frac{\langle \operatorname{div} \operatorname{Hess} u, \nabla u\rangle + |\operatorname{Hess} u|^2}{\langle \nabla u, \nabla u\rangle} - 4\,\frac{|\operatorname{Hess} u\,(\nabla u)|^2}{\langle \nabla u, \nabla u\rangle^2} \\ &= 2\frac{\operatorname{Ric}(\nabla u, \nabla u)}{|\nabla u|^2} + 2\frac{|\operatorname{Hess} u|^2}{|\nabla u|^2} - 4\,\frac{|\operatorname{Hess} u\,(\nabla u)|^2}{|\nabla u|^4} \\ &= 2K_\Omega + 2\frac{|\operatorname{Hess} u|^2}{|\nabla u|^2} - 4\frac{|\operatorname{Hess} u\,(\nabla u)|^2}{|\nabla u|^4},\end{aligned}$$

where we have used the identity

$$\operatorname{div} \operatorname{Hess} u = \nabla \Delta u + \operatorname{Ric}(\nabla u)$$

in the above. Since the dimension of Ω is 2, we have

$$|\operatorname{Hess} u|^2 |\nabla u|^2 = 2\,|\operatorname{Hess} u(\nabla u)|^2, \tag{3}$$

which leads to the conclusion that

$$\Delta \log |\nabla u|^2 = 2K_\Omega \geq 0.$$

□

We remark that (3) holds only for 2-dimensional cases as mentioned in [21]. Using these observations, we are able to prove the following.

Theorem 2.3. *Let Ω be a compact connected surface with Gaussian curvature $K_\Omega \geq 0$ and with C^2 boundary. If u is a solution to (1.1)-(1.3) with $\sigma = \sigma_1$ and with a nonpositive function $f \in C^\infty(\partial\Omega)$, then Ω is flat and $\partial\Omega$ is the disjoint union of geodesics and geodesic circles of radius $\frac{1}{\sigma_1}$.*

Proof. Note that the function u has no critical point in Ω by Lemma 2.1. By Lemma 2.2 (i) and (1.3), we see that

$$\begin{aligned} 0 \leq 2\int_\Omega K_\Omega &= \int_\Omega \Delta \log|\nabla u|^2 \\ &= \int_{\partial\Omega} \frac{\partial}{\partial \nu} \log|\nabla u|^2 \\ &= \int_{\partial\Omega} \frac{2f}{|\nabla u|^2} \\ &\leq 0. \end{aligned}$$

Combining with Lemma 2.2 (ii), we have

$$\begin{cases} \Delta \log|\nabla u|^2 = 0 & \text{in } \Omega, \\ \frac{\partial}{\partial \nu} \log|\nabla u|^2 = 0 & \text{on } \partial\Omega, \end{cases}$$

which shows that $|\nabla u|$ is constant in Ω by the maximum principle. Then applying the Bochner formula gives

$$0 = \frac{1}{2}\Delta|\nabla u|^2 = |\mathrm{Hess}\, u|^2 + K_\Omega |\nabla u|^2,$$

which implies that

$$\mathrm{Hess}\, u \equiv 0 \text{ and } K_\Omega \equiv 0$$

in Ω. Therefore, it follows that Ω is flat and $f = 0$. Since

$$\langle \nabla u, \nu \rangle = \sigma_1 u$$

along the boundary $\partial\Omega$ by (1.2), it follows that

$$\mathrm{Hess}\, u(e_1, \nu) + \langle \nabla u, \nabla_{e_1}\nu \rangle = \sigma_1 \nabla_{e_1} u,$$

which gives

$$(\sigma_1 + \langle \nabla_{e_1} e_1, \nu \rangle) \langle \nabla u, e_1 \rangle = 0$$

on $\partial\Omega$. Define two sets A and B as follows:

$$\begin{aligned} A &:= \{p \in \partial\Omega \mid \sigma_1 + \langle \nabla_{e_1} e_1, \nu \rangle = 0\} \quad \text{and} \\ B &:= \{p \in \partial\Omega \mid \langle \nabla u, e_1 \rangle = 0\}. \end{aligned}$$

Then it is clear that each connected component of A is contained in a geodesic circle of radius $\frac{1}{\sigma_1}$.

We claim that the set B consists of geodesics. To see this, we note that $\nabla u = g\nu$ on each connected component of B for some real-valued function $g \in C^\infty(\partial\Omega)$. Since

$$\nabla_{e_1} \nabla u = (e_1 g)\nu + g \nabla_{e_1} \nu,$$

we have

$$g \langle \nabla_{e_1} e_1, \nu \rangle = -\mathrm{Hess}\, u(e_1, e_1) = 0,$$

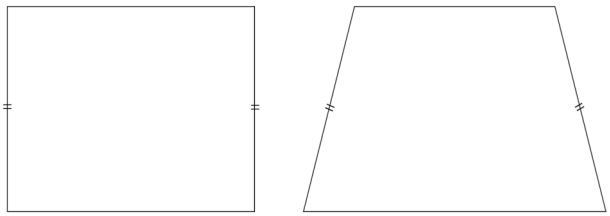


FIGURE 2. Possible doubly-connected flat surfaces

which implies that $g\langle\nabla_{e_1}e_1,\nu\rangle = 0$. Since $g \neq 0$ because otherwise ∇u vanishes, $\langle\nabla_{e_1}e_1,\nu\rangle = 0$. This means that $\nabla_{e_1}e_1 = 0$. Therefore, each connected component of B is contained in some geodesic. In conclusion, we see that $\partial\Omega$ is the disjoint union of geodesics and geodesic circles of radius $\frac{1}{\sigma_1}$ since $\partial\Omega$ is C^2. □

In particular, if the domain Ω in Theorem 2.3 is simply connected, we immediately get the following.

Corollary 2.4. *Let Ω be a compact simply connected surface with nonnegative Gaussian curvature with C^2 boundary. If u is a solution to* (1.1)-(1.3) *with $\sigma = \sigma_1$ and with a nonpositive function $f \in C^\infty(\partial\Omega)$, then Ω is a flat disk of radius $\frac{1}{\sigma_1}$.*

Remark 2.5. If Ω is doubly connected, then possible shapes of Ω are relatively complicated. To see this, let Ω be a doubly connected surface with Gaussian curvature $K_\Omega \geq 0$ and with C^2 boundary. If u is a solution to (1.1)-(1.3) with $\sigma = \sigma_1$ and with a nonpositive function $f \in C^\infty(\partial\Omega)$, then Theorem 2.3 implies that Ω is a flat surface whose topological type is an annulus, for example, a cylinder, a truncated cone, etc. See Fig. 2.

In [21], Payne-Philippin obtained a Serrin-type symmetry result for a simply connected bounded domain $\Omega \subset \mathbb{R}^2$. However, as seen in the proof of Theorem 2.3, the simple connectedness assumption can be dropped. This enables us to get the following result, which can be regarded as a generalization of the result in [21].

Theorem 2.6. *Let Ω be a connected bounded domain with C^2 boundary in $\mathbb{R}^2$. If u is a solution to the following overdetermined Steklov eigenvalue problem:*

$$\begin{cases} \Delta u = 0 & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = \sigma_1 u & \text{on } \partial\Omega, \\ |\nabla u| = const. = c & \text{on } \partial\Omega, \end{cases}$$

where $\nu = \nu(x)$ denotes the outward unit normal to $\partial\Omega$. Then Ω is a disk of radius $\frac{1}{\sigma_1}$.

Proof. As before, we have

$$\begin{cases} \Delta \log |\nabla u|^2 = 0 & \text{in } \Omega, \\ |\nabla u| = c & \text{on } \partial\Omega. \end{cases}$$

Thus we see that $|\nabla u| = c$ in Ω. The Bochner formula implies that

$$0 = \frac{1}{2}\Delta |\nabla u|^2 = |\mathrm{Hess}\, u|^2,$$

which shows that

$$\mathrm{Hess}\, u = 0.$$

Applying the same argument as in the proof of Theorem 2.3 shows that $\partial\Omega$ is the disjoint union of line segments or circles of radius $\frac{1}{\sigma_1}$. Since the boundary of Ω is C^2 and bounded in $\mathbb{R}^2$ by our assumption, $\partial\Omega$ is isometric to a circle of radius $\frac{1}{\sigma_1}$. This completes the proof. □

3. Compact Riemannian manifolds with nonnegative Ricci curvature

In this section, we study an overdetermined Steklov eigenvalue problem on higher-dimensional Riemannian manifolds and obtain a Serrin-type theorem. More precisely, we prove the following:

Theorem 3.1. *Let Ω be an n-dimensional compact connected Riemannian manifold with Ricci curvature $\mathrm{Ric}_\Omega \geq 0$ and with C^2 boundary. Let u be a solution to the overdetermined Steklov eigenvalue problem (1.1)-(1.3) with $\sigma = \sigma_1$ and with a nonpositive function $f \in C^\infty(\partial\Omega)$. Assume that u has no critical point in Ω. Then Ω is Ricci flat and the second fundamental form (h_{ij}) on $\partial\Omega$ is of the form*

$$(h_{ij}) = \left(\begin{array}{ccc|c} -\sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \cdots & -\sigma_1 & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right)$$

with respect to a suitable orthonormal basis on $\partial\Omega$.

Proof. We first observe that

$$\begin{aligned} \Delta|\nabla u| &= \mathrm{div}\nabla\left(\langle \nabla u, \nabla u\rangle^{\frac{1}{2}}\right) \\ &= \mathrm{div}\left(\langle \nabla u, \nabla u\rangle^{-\frac{1}{2}}\, \mathrm{Hess}\, u(\nabla u)\right) \\ &= -\frac{1}{2}\langle \nabla u, \nabla u\rangle^{-\frac{3}{2}} \cdot 2|\mathrm{Hess}\, u(\nabla u)|^2 \\ &\quad + \langle \nabla u, \nabla u\rangle^{-\frac{1}{2}}\left(\langle \mathrm{div}\,\mathrm{Hess}\, u, \nabla u\rangle + |\mathrm{Hess}\, u|^2\right) \end{aligned}$$

$$\begin{aligned}
&\geq -\langle \nabla u, \nabla u\rangle^{-\frac{3}{2}} |\mathrm{Hess}\, u(\nabla u)|^2 + \langle \nabla u, \nabla u\rangle^{-\frac{1}{2}} |\mathrm{Hess}\, u|^2 \\
&\geq -\langle \nabla u, \nabla u\rangle^{-\frac{3}{2}} |\mathrm{Hess}\, u|^2 |\nabla u|^2 + \langle \nabla u, \nabla u\rangle^{-\frac{1}{2}} |\mathrm{Hess}\, u|^2 \\
&= -|\nabla u|^{-1} |\mathrm{Hess}\, u|^2 + |\nabla u|^{-1} |\mathrm{Hess}\, u|^2 \\
&= 0. \qquad (4)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\frac{\partial}{\partial \nu} \langle \nabla u, \nabla u\rangle^{\frac{1}{2}} &= \frac{1}{2} \langle \nabla u, \nabla u\rangle^{-\frac{1}{2}} \langle \nabla_\nu \nabla u, \nabla u\rangle \\
&= \frac{\mathrm{Hess}\, u(\nabla u, \nu)}{\langle \nabla u, \nabla u\rangle^{\frac{1}{2}}} \\
&= \frac{f}{|\nabla u|} \\
&\leq 0. \qquad (5)
\end{aligned}$$

Integrating (4) over Ω and using (5), we get

$$0 \leq \int_\Omega \Delta |\nabla u| = \int_{\partial\Omega} \frac{\partial}{\partial \nu} |\nabla u| \leq 0.$$

Thus we obtain the following:

$$\begin{cases} \Delta |\nabla u| = 0 & \text{in } \Omega, \\ \frac{\partial}{\partial \nu} |\nabla u| = 0 & \text{on } \partial\Omega. \end{cases}$$

The maximum principle implies that $|\nabla u|$ is constant in Ω. It follows that

$$\mathrm{Hess}\, u \equiv 0 \quad \text{and} \quad \mathrm{Ric} \equiv 0$$

since

$$0 \leq |\mathrm{Hess}\, u|^2 + \mathrm{Ric}(\nabla u, \nabla u) = \frac{1}{2} \Delta |\nabla u|^2 = 0.$$

by the Bochner formula. Now let $\{e_i\}_{i=1}^{n-1}$ be an orthonormal frame in the neighborhood of a point on $\partial\Omega$ such that, at the point,

$$(\nabla_{e_i} e_j)^\top = 0 \quad \text{for } 1 \leq i, j \leq n-1 \text{ and}$$

the second fundamental form h_{ij} is diagonalized, i.e.,

$$(h_{ij}) = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{n-1} \end{pmatrix},$$

where $\top$ denotes the orthogonal projection onto the tangent space $T_p(\partial\Omega)$ and $h_{ij} = \langle \nu, \nabla_{e_i} e_j \rangle$. Then it follows from (1.2) that

$$e_j \langle \nabla u, \nu \rangle = \sigma_1 e_j u, \qquad \forall j = 1, \cdots, n-1.$$

Hence we have

$$\begin{aligned}\sigma_1 \langle \nabla u, e_j \rangle = e_j \langle \nabla u, \nu \rangle &= \langle \nabla_{e_j} \nabla u, \nu \rangle + \langle \nabla u, \nabla_{e_j} \nu \rangle \\ &= \langle \nabla u, \nabla_{e_j} \nu \rangle \\ &= \langle \nabla u, -\lambda_j e_j \rangle\end{aligned}$$

or, equivalently,

$$(\sigma_1 + \lambda_j) \langle \nabla u, e_j \rangle = 0$$

for each $j = 1, \cdots, n-1$. As in the proof of Theorem 2.3, we divide into three cases as follows.

Case I: Suppose that $\sigma_1 + \lambda_j = 0$ for all $j = 1, \cdots, n-1$. Then $\partial\Omega$ is totally umbilic.

Case II: Suppose that $\langle \nabla u, e_j \rangle = 0$ for all $j = 1, \cdots, n-1$. Then $\nabla u = f\nu$ for some function f. This gives

$$\begin{aligned}e_i f &= \langle \nabla_{e_i} \nabla u, \nu \rangle \\ &= \langle (e_i f)\nu + f \nabla_{e_i} \nu, \nu \rangle \\ &= \operatorname{Hess} u(e_i, \nu) \\ &= 0\end{aligned}$$

for all $i = 1, \cdots, n-1$. It follows that the function f should be a nonzero constant function on $\partial\Omega$. In other words, $\nabla u = c\nu$ for some constant $c \neq 0$. Thus

$$\begin{aligned}0 &= \operatorname{Hess} u(e_i, e_k) \\ &= \langle \nabla_{e_i} \nabla u, e_k \rangle \\ &= c \langle \nabla_{e_i} \nu, e_k \rangle \\ &= c \langle -h_{ij} e_j, e_k \rangle \\ &= -c h_{ik}\end{aligned}$$

for each $k = 1, \cdots, n-1$. Therefore, we see that $h_{ik} \equiv 0$, which implies that $\partial\Omega$ is totally geodesic.

Case III: Suppose that

$$\begin{cases} \sigma_1 + \lambda_j = 0 & \text{for } 1 \leq j \leq k, \\ \langle \nabla u, e_j \rangle = 0 & \text{for } k+1 \leq j \leq n-1 \end{cases}$$

by rearranging the entries if necessary. In other words, the second fundamental form h_{ij} is given by

$$(h_{ij}) = \left(\begin{array}{ccc|c} -\sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \cdots & -\sigma_1 & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right).$$

Since

$$\nabla u = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k + \beta\nu,$$

we have

$$\begin{aligned} \nabla_{e_j}\nabla u &= \sum_{i=1}^{k} \nabla_{e_j}(\alpha_i e_i) + \nabla_{e_j}(\beta\nu) \\ &= \sum_{i=1}^{k} e_j(\alpha_i)e_i + \sum_{i=1}^{k} \alpha_i(\nabla_{e_j} e_i) + e_j(\beta)\nu + \beta\nabla_{e_j}\nu \\ &= \sum_{i=1}^{k} e_j(\alpha_i)e_i + \alpha_j\lambda_j\nu + e_j(\beta)\nu - \beta\lambda_j e_j \\ &= \sum_{i=1}^{k} e_j(\alpha_i)e_i - \beta\lambda_j e_j + (e_j(\beta) + \alpha_j\lambda_j)\nu. \end{aligned}$$

Since

$$\operatorname{Hess} u = 0 \quad \text{for } k+1 \le j \le n-1,$$

we get

$$\beta\lambda_j = 0.$$

Suppose that $\beta = 0$. Then we have $\nabla u = \alpha_1 e_1 + \cdots + \alpha_k e_k$, which implies that

$$\sigma_1 u = \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega.$$

Thus, we have $u = 0$ on $\partial\Omega$. Since $\Delta u = 0$ in Ω, it follows that $u \equiv 0$ in Ω. This contradicts the assumption that u is the first eigenfunction. Therefore, we have $\lambda_j = 0$ for $k+1 \le j \le n-1$, which shows that

$$(h_{ij}) = \left(\begin{array}{ccc|c} -\sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \cdots & -\sigma_1 & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right)$$

with respect to the chosen orthonormal frame. □

Remark 3.2. In the 2-dimensional case, the first Steklov eigenfunction has no critical point in Ω by Lemma 2.1. However, we do not know whether this property is still true for higher-dimensional Riemannian manifolds or not. It would be interesting to get rid of the condition that the first Steklov eigenfunction has no critical point in the domain in Theorem 3.1.

In particular, if Ω is a compact connected domain with C^2 boundary in a Euclidean space, then we immediately get the following consequences.

Corollary 3.3. *Let $\Omega \subset \mathbb{R}^n$ be a compact connected domain. Let u be a solution to the overdetermined Steklov eigenvalue problem* (1.1)-(1.3) *with $\sigma = \sigma_1$ and with a nonpositive function $f \in C^\infty(\partial\Omega)$. Assume that u has no critical point in Ω. Then $\partial\Omega$ is isometric to $\mathbb{S}^{n-1}\left(\frac{1}{\sigma_1}\right)$, where $\mathbb{S}^m(r)$ denotes an m-dimensional Euclidean sphere of radius r.*

Proof. From Theorem 3.1 it follows that if the number of the eigenvalue σ_1 is k, then $\partial\Omega$ is isometric to the product of a k-dimensional geodesic sphere and an $(n-k-1)$-dimensional Euclidean space. In fact, this comes from a kind of a generalized Bonnet theorem (see [29] for instance). However, since Ω is compact, we obtain that $\partial\Omega$ is isometric to $\mathbb{S}^{n-1}\left(\frac{1}{\sigma_1}\right)$. □

Acknowledgements

The authors would like to thank the reviewers deeply for their careful reading of our manuscript and their helpful comments and suggestions. The second author was supported by the National Research Foundation of Korea (NRF-2021R1A2C1003365).

Data availability Data sharing is not applicable to this article as no data sets were generated or analysed during the current study.

Declarations

Competing interests The corresponding author states that there is no conflict of interest.

Open Access. This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article

are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

[1] Alessandrini, G., Magnanini, R.: Symmetry and nonsymmetry for the overdetermined Stekloff eigenvalue problem. Z. Angew. Math. Phys. **45**(1), 44–52 (1994)

[2] Brandolini, B., Nitsch, C., Salani, P., Trombetti, C.: Serrin-type overdetermined problems: an alternative proof. Arch. Ration. Mech. Anal. **190**(2), 267–280 (2008)

[3] Cheng, S.Y.: Eigenfunctions and nodal sets. Comment. Math. Helv. **51**(1), 43–55 (1976)

[4] Ciraolo, G., Vezzoni, L.: A rigidity problem on the round sphere. Commun. Contemp. Math. **19**(5), 1750001, 11 pp (2017)

[5] Ciraolo, G., Vezzoni, L.: On Serrin's overdetermined problem in space forms. Manuscripta Math. **159**(3–4), 445–452 (2019)

[6] Colbois, B., Girouard, A., Gordon, C., Sher, D.: Some recent developments on the Steklov eigenvalue problem. Rev. Mat. Complut. (2023). https://doi.org/10.1007/s13163-023-00480-3

[7] Escobar, J.F.: The geometry of the first non-zero Stekloff eigenvalue. J. Funct. Anal. **150**(2), 544–556 (1997)

[8] Escobar, J.F.: An isoperimetric inequality and the first Steklov eigenvalue. J. Funct. Anal. **165**(1), 101–116 (1999)

[9] Escobar, J.F.: A comparison theorem for the first non-zero Steklov eigenvalue. J. Funct. Anal. **178**(1), 143–155 (2000)

[10] Fall, M.M., Minlend, I.A., Weth, T.: Serrin's overdetermined problem on the sphere. Calc. Var. Partial Differential Equations **57**(3), 24 (2018)

[11] Fraser, A., Schoen, R.: The first Steklov eigenvalue, conformal geometry, and minimal surfaces. Adv. Math. **226**(5), 4011–4030 (2011)

[12] Fraser, A., Schoen, R.: Sharp eigenvalue bounds and minimal surfaces in the ball. Invent. Math. **203**(3), 823–890 (2016)

[13] Fraser, A., Schoen, R.: Some results on higher eigenvalue optimization, Calc. Var. Partial Differential Equations **59**, no. 5, Paper No. 151, 22 pp (2020)

[14] Girouard, A., Polterovich, I.: Spectral geometry of the Steklov problem. J. Spectr. Theory **7**(2), 321–359 (2017)

[15] Kumaresan, S., Prajapat, J.: Serrin's result for hyperbolic space and sphere. Duke Math. J. **91**(1), 17–28 (1998)

[16] Lee, J., Seo, K.: Radial symmetry and partially overdetermined problems in a convex cone. Math. Nachr. **296**(3), 1204–1224 (2023)

[17] Lee, J., Seo, K.: Overdetermined problems in annular domains with a spherical-boundary component in space forms, J. Inequal. Appl., Paper No. 45, 22 pp (2023)

[18] Molzon, R.: Symmetry and overdetermined boundary value problems. Forum Math. **3**(2), 143–156 (1991)

[19] Morabito, F.: Serrin's overdetermined problem on $\mathbb{S}^n \times \mathbb{R}$, J. Geom. Anal. **33**, no. 10, Paper No. 327, 17 pp (2023)

[20] Pacella, F., Tralli, G.: Overdetermined problems and constant mean curvature surfaces in cones. Rev. Mat. Iberoam. **36**(3), 841–867 (2020)

[21] Payne, L.E., Philippin, G.A.: Some overdetermined boundary value problems for harmonic functions. Z. Angew. Math. Phys. **42**(6), 864–873 (1991)

[22] Payne, L.E., Philippin, G.A.: On two free boundary problems in potential theory. J. Math. Anal. Appl. **161**(2), 332–342 (1991)

[23] Philippin, G.A.: On a free boundary problem in electrostatics. Math. Methods Appl. Sci. **12**(5), 387–392 (1990)

[24] Qiu, G., Xia, C.: Overdetermined boundary value problems in $\mathbb{S}^n$. J. Math. Study **50**(2), 165–173 (2017)

[25] Roncoroni, A.: A Serrin-type symmetry result on model manifolds: An extension of the Weinberger argument. C. R. Math. Acad. Sci. Paris **356**(6), 648–656 (2018)

[26] Ros, A., Ruiz, D., Sicbaldi, P.: A rigidity result for overdetermined elliptic problems in the plane. Comm. Pure Appl. Math. **70**(7), 1223–1252 (2017)

[27] Ros, A., Sicbaldi, P.: Geometry and topology of some overdetermined elliptic problems. J. Differential Equations **255**(5), 951–977 (2013)

[28] Serrin, J.: A symmetry problem in potential theory. Arch. Ration. Mech. Anal. **43**(4), 304–318 (1971)

[29] Spivak, M.: A comprehensive introduction to differential geometry Vol. IV, Second edition, Publish or Perish, Inc., Wilmington, DE, viii+561 pp (1979)

[30] Weinberger, H.F.: Remark on the preceding paper of Serrin. Arch. Ration. Mech. Anal. **43**(4), 319–320 (1971)

Eunjoo Lee
Department of Mathematics
Soongsil University
Sangdo-ro 369
Dongjak-gu, Seoul 06978
Korea
e-mail: eunjoolee@ssu.ac.kr

Keomkyo Seo
Department of Mathematics and Research Institute of Natural Sciences
Sookmyung Women's University
Cheongpa-ro 47-gil 100
Yongsan-ku, Seoul 04310
Korea
e-mail: kseo@sookmyung.ac.kr
URL: http://sites.google.com/site/keomkyo/

Received: May 16, 2024.
Accepted: April 11, 2025.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The Journal of Geometric Analysis (2026) 36:59
https://doi.org/10.1007/s12220-025-02309-5

Harmonic minimal graphs in product spaces

Donghoon Park[1] · **Keomkyo Seo**[2]

Received: 25 October 2024 / Accepted: 17 December 2025
© The Author(s) 2025

Abstract
It is known that a harmonic minimal graph in 3-dimensional Euclidean space is locally either a plane or a helicoid. We prove that any harmonic minimal graph in $\mathbb{R}^{n+1}$ is locally foliated by $(n-1)$-dimensional minimal graphs. Moreover, it is proved that a harmonic minimal graph in $M^2 \times \mathbb{R}$ is locally either a horizontal level surface or a helicoid, where M^2 denotes either 2-dimensional hyperbolic space $\mathbb{H}^2$ or the 2-dimensional unit sphere $\mathbb{S}^2$.

Keywords Harmonic graph · Minimal surface equation · Foliation · Helicoid

Mathematics Subject Classification 53C42 · 53A10 · 53A35

1 Introduction

During recent decades, minimal surface theory in product spaces has been actively researched and developed by many geometers. For instance, Rosenberg [17] studied the geometry and topology of minimal surfaces in $M^2 \times \mathbb{R}$, where M is a complete surface of nonnegative curvature. Nelli-Rosenberg [15] obtained a Jenkins-Serrin theorem in $\mathbb{H}^2 \times \mathbb{R}$. Moreover, Abresch-Rosenberg [1] used the generalized notion of a holomorphic Hopf differential to constant mean curvature surfaces in $\mathbb{H}^2 \times \mathbb{R}$ to show that any topological sphere in $\mathbb{H}^2 \times \mathbb{R}$ with constant mean curvature is embedded and rotationally symmetric. We refer the reader to [2, 5–8, 10, 11, 14–19] and references therein. On the other hand, O. Kobayashi [13] observed the following interesting fact.

✉ Keomkyo Seo
kseo@sookmyung.ac.kr
http://sites.google.com/site/keomkyo/

Donghoon Park
dhpark12@gmail.com

[1] Department of Mathematics, Yonsei University, 50 Yonsei-Ro, Seodaemun-Gu, 03722 Seoul, Korea

[2] Department of Mathematics and Research Institute of Natural Sciences, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-ku, 04310 Seoul, Korea

Published online: 27 December 2025

Theorem 1.1 (Kobayashi, [13]) *Let Ω be a domain in $\mathbb{R}^2$ and let $f : \Omega \to \mathbb{R}$ be a smooth function. Then any two of the following conditions for $z = f(x, y)$ imply that the graph of $z = f(x, y)$ is locally either a plane or a helicoid which is given by $y = x \tan z$:*

- *The function f is harmonic.*
- *The graph of f in a Euclidean space $\mathbb{R}^3$ is a minimal surface.*
- *The graph of f in the Lorentz-Minkowski spacetime $\mathbb{L}^3 = (\mathbb{R}^3, dx^2 + dy^2 - dz^2)$ is a maximal surface, i.e., it has mean curvature zero.*

Indeed, it is known that there are six different kinds of helicoids in $\mathbb{L}^3$: the causal character of the normal vectors determines spacelike or timelike helicoids and the causal character of the axes of the screw motion determines elliptic, parabolic or hyperbolic ones. The helicoid considered by Kobayashi is the elliptic one. Kim-Lee-Yang [12] showed that the same result holds for other kinds of helicoids in $\mathbb{L}^3$.

Motivated by the above results, we study the geometry of harmonic minimal graphs in product spaces. In Section 2, it turns out that any n-dimensional harmonic minimal graph is locally foliated by $(n-1)$-dimensional minimal graphs (see Theorem 2.1). As a consequence, our result recovers Theorem 1.1 by Kobayashi. In Section 3, it is shown that any harmonic minimal graph in $M^2 \times \mathbb{R}$ is foliated by geodesics, where M^2 denotes either 2-dimensional hyperbolic space $\mathbb{H}^2$ or the 2-dimensional unit sphere $\mathbb{S}^2$ (see Theorem 3.1). This result can be regarded as an extension of Theorem 1.1 by Kobayashi.

2 Harmonic minimal graphs in a higher-dimensional Euclidean space

Let $\Omega \subset \mathbb{R}^n$ be a C^2 bounded domain. In [9], Jenkins-Serrin proved that the Dirichlet problem for the following minimal surface equation in Ω with $f = f(x_1, x_2, \cdots, x_n)$ is solvable for C^2 boundary values on $\partial\Omega$ if and only if $\partial\Omega$ is convex:

$$(1 + |\nabla f|^2)\Delta f - \sum_{i,j=1}^{n} f_i f_j f_{ij} = 0, \tag{1}$$

where ∇ and Δ denote the gradient and the Laplace operator on Ω, respectively. In this section, we show that any n-dimensional harmonic minimal graph is locally foliated by $(n-1)$-dimensional minimal graphs. More precisely, we prove the following.

Theorem 2.1 *Let $\Omega \subset \mathbb{R}^n$ be a domain and let $f : \Omega \to \mathbb{R}$ be a harmonic function. Assume that the function f satisfies the minimal surface equation (1). Then the harmonic minimal graph $f(\Omega)$ is locally foliated by $(n-1)$-dimensional minimal graphs.*

Proof The graph of the function $t = f(x_1, \cdots, x_n)$ can be locally regarded as a graph of the function h defined by

$$\begin{aligned} x_n &= h(x_1, \cdots, x_{n-1}, t) \\ &= h(x_1, \cdots, x_{n-1}, f(x_1, \cdots, x_n)) \end{aligned}$$

provided that $\frac{\partial f}{\partial x_n} \neq 0$. Note that we may assume that $\frac{\partial f}{\partial x_n} \neq 0$ almost everywhere in Ω by Sard's theorem unless the function f is constant. For $1 \leq j, k \leq n-1$, a simple computation shows that

$$\begin{aligned}
0 &= \frac{\partial x_n}{\partial x_j} = h_{x_j} + h_t f_{x_j} \\
1 &= \frac{\partial x_n}{\partial x_n} = h_t f_{x_n} \\
0 &= \frac{\partial^2 x_n}{\partial x_j^2} = h_{x_j x_j} + h_{x_j t} f_{x_j} + h_{t x_j} f_{x_j} + h_{tt} f_{x_j}^2 + h_t f_{x_j x_j} \\
0 &= \frac{\partial^2 x_n}{\partial x_k \partial x_j} = h_{x_j x_k} + h_{x_j t} f_{x_k} + h_{t x_k} f_{x_j} + h_{tt} f_{x_j} f_{x_k} + h_t f_{x_j x_k} \\
0 &= \frac{\partial^2 x_n}{\partial x_j \partial x_n} = h_{t x_j} f_{x_n} + h_{tt} f_{x_j} f_{x_n} + h_t f_{x_j x_n} \\
0 &= \frac{\partial^2 x_n}{\partial x_n^2} = h_{tt} f_{x_n}^2 + h_t f_{x_n x_n}.
\end{aligned}$$

In particular, we see that $h_t \neq 0$ from the second identity. Thus we have

$$\begin{aligned}
f_{x_j} &= -\frac{h_{x_j}}{h_t} \\
f_{x_n} &= \frac{1}{h_t} \\
f_{x_j x_j} &= \frac{-h_{x_j x_j} h_t^2 + 2 h_{x_j t} h_{x_j} h_t - h_{tt} h_{x_j}^2}{h_t^3} \\
f_{x_j x_k} &= \frac{-h_{x_j x_j} h_t^2 + h_{x_j t} h_{x_k} h_t + h_{x_k t} h_{x_j} h_t - h_{tt} h_{x_j} h_{x_k}}{h_t^3} \\
f_{x_j x_n} &= \frac{-h_{x_j t} h_t + h_{tt} h_{x_j}}{h_t^3} \\
f_{x_n x_n} &= -\frac{h_{tt}}{h_t^3}.
\end{aligned} \tag{2}$$

Since the harmonic function f satisfies the minimal surface equation (1), we have the following:

$$\sum_{i=1}^{n} f_{x_i x_i} = 0$$

and

$$\sum_{i=1}^{n} \left(1 + \sum_{j \neq i} f_{x_j}^2 \right) f_{x_i x_i} - 2 \sum_{i=1}^{n} \sum_{j<i} f_{x_i} f_{x_j} f_{x_i x_j} = 0.$$

From these two equations, it follows that

$$\sum_{i=1}^{n}\sum_{j\neq i}\left(f_{x_j}^2 f_{x_ix_i} - f_{x_i} f_{x_j} f_{x_ix_j}\right) = 0. \tag{3}$$

Using (2), one can rewrite (3) in terms of the function h as follows:

$$\begin{aligned}
0 &= \sum_{i=1}^{n}\sum_{j\neq i}\left(f_{x_j}^2 f_{x_ix_i} - f_{x_i} f_{x_j} f_{x_ix_j}\right)\\
&= \sum_{i=1}^{n-1}\sum_{j\neq i,\ j<n}\left(f_{x_j}^2 f_{x_ix_i} - f_{x_i} f_{x_j} f_{x_ix_j}\right)\\
&\quad + \sum_{i=1}^{n-1}\left(f_{x_n}^2 f_{x_ix_i} - f_{x_i} f_{x_n} f_{x_ix_n}\right) + \sum_{j=1}^{n-1}\left(f_{x_j}^2 f_{x_nx_n} - f_{x_n} f_{x_j} f_{x_nx_j}\right)\\
&= -\frac{1}{h_t^3}\sum_{i=1}^{n-1}\left(h_{x_ix_i} + \sum_{j\neq i,\ j<n}(h_{x_i}^2 h_{x_ix_i} - h_{x_i} h_{x_j} h_{x_ix_j})\right).
\end{aligned}$$

Thus, for any fixed $t = t_0$, this implies that $x_n = h(x_1, \cdots, x_{n-1}, t_0)$ satisfies the following $(n-1)$-dimensional minimal surface equation

$$0 = \sum_{i=1}^{n-1}\left(1 + \sum_{j\neq i,\ j<n} h_{x_i}^2\right) h_{x_ix_i} - 2\sum_{i=1}^{n-1}\sum_{j<i} h_{x_i} h_{x_j} h_{x_ix_j},$$

which shows that the harmonic minimal graph of $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is locally foliated by $(n-1)$-dimensional minimal graphs of the function $h(x_1, \cdots, x_{n-1}, t_0)$. This completes the proof. □

In particular, when $n = 2$, we immediately get the following consequence. Since the only 1-dimensional minimal graph in $\mathbb{R}^2$ is a straight line, the harmonic minimal graph of a planar domain is locally foliated by straight lines by Theorem 2.1. From the well-known fact that the only ruled minimal surface in $\mathbb{R}^3$ is either a plane or a helicoid (see [3] for example), we are able to recover the result by Kobayashi as follows:

Corollary 2.2 (Kobayashi, [13]) *Let $\Omega \subset \mathbb{R}^2$ be a domain and let $f : \Omega \to \mathbb{R}$ be a harmonic function. Assume that the function f satisfies the minimal surface equation (1). Then the harmonic minimal graph $f(\Omega)$ is locally either a plane or a helicoid.*

It would be interesting if Corollary 2.2 were still true in arbitrary dimensions. Motivated by Theorem 2.1 and Corollary 2.2, it is natural to ask the following.

Question Is every harmonic minimal graph in $\mathbb{R}^{n+1}$ locally congruent to either a hyperplane or a generalized helicoid?

3 Harmonic minimal graphs in $\mathbb{H}^2 \times \mathbb{R}$ and $\mathbb{S}^2 \times \mathbb{R}$

In this section, we investigate harmonic minimal graphs in $M^2 \times \mathbb{R}$, where M is either 2-dimensional hyperbolic space $\mathbb{H}^2$ or the unit sphere $\mathbb{S}^2$. We show that any harmonic minimal graph in $M^2 \times \mathbb{R}$ is foliated by geodesics, which can be regarded as an extension of Theorem 1.1 by Kobayashi.

Theorem 3.1 *Let Ω be a domain in M^2, where M^2 is either 2-dimensional hyperbolic space $\mathbb{H}^2$ or the unit sphere $\mathbb{S}^2$. Let $f : \Omega \to \mathbb{R}$ be a harmonic function. Assume that the graph of the function f is minimal. Then the harmonic minimal graph $f(\Omega)$ in $M^2 \times \mathbb{R}$ is locally either a horizontal level surface $M^2 \times \{t_0\}$ or a helicoid.*

Proof We first consider the case where M^2 is a 2-dimensional hyperbolic space. Consider the upper half-plane model of $\mathbb{H}^2$ as follows:

$$\mathbb{H}^2 = \{(x, y) \in \mathbb{R}^2 : y > 0\}$$

with the metric $ds^2 = \frac{dx^2+dy^2}{y^2}$. Then the product space $\mathbb{H}^2 \times \mathbb{R}$ is equipped with the metric $d\sigma^2 = ds^2 + dt^2$. Let $\Omega \subset \mathbb{H}^2$ be a domain and let $f : \Omega \to \mathbb{R}$ be a C^2 function on Ω. Then the vertical graph in $\mathbb{H}^2 \times \mathbb{R}$ is given by $G = \{(x, f(x)) : x \in \Omega\}$ and the mean curvature H of G satisfies

$$H = \operatorname{div}_{\mathbb{H}^2}\left(\frac{\nabla^{\mathbb{H}^2} f}{\sqrt{1 + |\nabla^{\mathbb{H}^2} f|^2}}\right),$$

where $\operatorname{div}_{\mathbb{H}^2}$ and $\nabla^{\mathbb{H}^2}$ denote the divergence and the gradient in $\mathbb{H}^2$, respectively. Thus, if the (vertical) graph $G \subset \mathbb{H}^2 \times \mathbb{R}$ is minimal, then it satisfies the following minimal surface equation (see [18] for example)

$$(1 + y^2 f_y^2) f_{xx} + (1 + y^2 f_x^2) f_{yy} - 2y^2 f_x f_y f_{xy} - y f_y (f_x^2 + f_y^2) = 0. \tag{4}$$

Furthermore, in case where the graph is harmonic, the function f satisfies

$$\Delta_{\mathbb{H}^2} f = 0,$$

where $\Delta_{\mathbb{H}^2}$ denotes the Laplace operator on $\mathbb{H}^2$. In other words, we have

$$f_{xx} + f_{yy} = 0. \tag{5}$$

As in the previous section, since we may assume that $\frac{\partial f}{\partial y} \neq 0$, we compute

$$\begin{aligned} t &= f(x, y) \\ y &= h(x, t) = h\,(x,\, f(x, y)) \\ f_x &= -\frac{h_x}{h_t} \\ f_y &= \frac{1}{h_t} \\ f_{xx} &= \frac{-h_{xx}h_t^2 + 2h_{xt}h_x h_t - h_{tt}h_x^2}{h_t^3} \\ f_{xy} &= \frac{-h_{xt}h_t + h_{tt}h_x}{h_t^3} \\ f_{yy} &= -\frac{h_{tt}}{h_t^3}. \end{aligned} \tag{6}$$

Note that $h_t \neq 0$. Using (6), we can rewrite the minimal surface equation (4) as

$$\begin{aligned} (h_t^2 + h^2)(h_{xx}h_t^2 - 2h_x h_t h_{xt} + h_x^2 h_{tt}) + (h_t^2 + h^2 h_x^2)h_{tt} \\ + 2h^2 h_x (h_t h_{xt} - h_x h_{tt}) + h h_t^2 (1 + h_x^2) = 0. \end{aligned} \tag{7}$$

Moreover, (5) becomes

$$h_t^2 h_{xx} - 2h_x h_t h_{xt} + (1 + h_x^2)h_{tt} = 0. \tag{8}$$

Applying (8) to the minimal surface equation (7), we get

$$(-h^2 - h^2 h_x^2)h_{tt} + 2h^2 h_x h_t h_{xt} + h h_t^2 (1 + h_x^2) = 0.$$

Using (8) again, we obtain

$$h h_{xx} + h_x^2 + 1 = 0,$$

which gives

$$(h h_x)_x = -1.$$

This shows that the function h satisfies that

$$h^2 = -x^2 + x g_1(t) + g_2(t),$$

where g_1 and g_2 are the functions of t. This implies that

$$\left(x - \frac{g_1(t)}{2}\right)^2 + h^2 = \frac{g_1(t)^2}{4} + g_2(t),$$

which is a geodesic on the upper half-plane $\mathbb{H}^2$ for each level t whenever $\frac{g_1(t)^2}{4} + g_2(t) > 0$. Therefore we see that the harmonic minimal graph in $\mathbb{H}^2 \times \mathbb{R}$ is locally

foliated by horizontal geodesics. It is known that any ruled minimal surface in $\mathbb{H}^2 \times \mathbb{R}$ is either a part of a cylinder $\gamma \times \mathbb{R}$ over a horizontal geodesic γ, a horizontal level surface $\mathbb{H}^2 \times \{t_0\}$ for some $t_0 \in \mathbb{R}$, or a helicoid, which was proved by Kim-Koh-Shin-Yang [10]. Since the graph over the domain $\Omega \subset \mathbb{H}^2$ cannot be a part of a cylinder $\gamma \times \mathbb{R}$ over a horizontal geodesic γ, we get the conclusion.

Now consider the case where M^2 is the unit sphere $\mathbb{S}^2$. We use the following conformal model of $\mathbb{S}^2$:

$$\left(\mathbb{R}^2, ds^2 = \lambda^2(dx^2 + dy^2)\right),$$

where the conformal factor λ is given by

$$\lambda = \frac{2}{1 + x^2 + y^2}.$$

Moreover, the product space $\mathbb{S}^2 \times \mathbb{R}$ is equipped with the metric $d\sigma^2 = ds^2 + dt^2$. Let $\Omega \subset \mathbb{S}^2$ be a domain and let $f : \Omega \to \mathbb{R}$ be a C^2 function defined on Ω. The graph in $\mathbb{S}^2 \times \mathbb{R}$ is denoted by $G = \{(x, f(x)) : x \in \Omega\}$. The harmonicity of f implies that

$$\Delta_{\mathbb{S}^2} f = 0,$$

where $\Delta_{\mathbb{S}^2}$ denotes the Laplace operator on $\mathbb{S}^2$. This is equivalent to

$$f_{xx} + f_{yy} = 0. \tag{9}$$

Since the graph G is minimal by assumption, the function f satisfies the following minimal surface equation (see [18] for example)

$$0 = \frac{2(\lambda_x f_x + \lambda_y f_y)}{\lambda^3 \sqrt{1 + \frac{f_x^2 + f_y^2}{\lambda^2}}} + \frac{\partial}{\partial x}\left(\frac{\frac{f_x}{\lambda^2}}{\sqrt{1 + \frac{f_x^2 + f_y^2}{\lambda^2}}}\right) + \frac{\partial}{\partial y}\left(\frac{\frac{f_y}{\lambda^2}}{\sqrt{1 + \frac{f_x^2 + f_y^2}{\lambda^2}}}\right). \tag{10}$$

As before, we may think the graph of the function $t = f(x, y)$ locally as that of the function

$$y = h(x, t) = h(x, f(x, y))$$

provided that $\frac{\partial f}{\partial y} \neq 0$. In terms of the function h, (9) is given by

$$h_t h_{xx} - 2h_x h_t h_{xt} + (1 + h_x^2) h_{tt} = 0 \tag{11}$$

Using the fact that

$$\lambda_x = -x\lambda^2 \quad \text{and} \quad \lambda_y = -y\lambda^2,$$

the minimal surface equation (10) can be rewritten as follows:

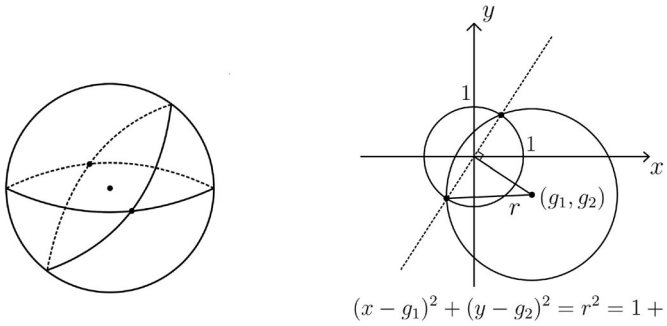


Fig. 1 Two geodesics on $\mathbb{S}^2$. The equator on $\mathbb{S}^2$ corresponds to the unit circle centered at the origin in the xy-plane under the stereographic projection from the north pole. The other corresponds to the circle $(x-g_1)^2+(y-g_2)^2=r^2$.

$$0=(\lambda^2+f_x^2+f_y^2)(f_{xx}+f_{yy})$$
$$-f_x\left(f_xf_{xx}+f_yf_{xy}+x\lambda(f_x^2+f_y^2)\right)-f_y\left(f_xf_{xy}+f_yf_{yy}+y\lambda(f_x^2+f_y^2)\right).$$

From (11), it follows that

$$0=f_x^2f_{xx}+f_y^2f_{yy}+2f_xf_yf_{xy}+(f_x^2+f_y^2)(x\lambda f_x+y\lambda f_y). \tag{12}$$

In terms of h, combining (12) with (11), we get

$$h_{xx}-\lambda(1+h_x^2)(xh_x-h)=0. \tag{13}$$

One can check that the following geodesic on $\left(\mathbb{R}^2, ds^2=\lambda^2(dx^2+dy^2)\right)$

$$(x-g_1(t))^2+(y-g_2(t))^2=1+g_1^2(t)+g_2^2(t)$$

satisfies (13) for fixed t (see Figure 1). Picard's iteration (see [4, Theorem I.3.1] for example) guarantees that these geodesics are the unique solutions to the equation (13) under an initial condition. Indeed, by using integrating factors, we are able to solve (13) directly. To see this, multiplying both sides of (13) by $\frac{-x(1+x^2+h^2)}{(xh_x-h)^2}$ gives

$$\begin{aligned}0&=-\frac{xh_{xx}(1+x^2+h^2)}{(xh_x-h)^2}+\frac{2x(1+h_x^2)}{xh_x-h}\\&=\frac{2xh_x^2+2xhh_{xx}+2x}{xh_x-h}-\frac{xh_{xx}(2xhh_x+x^2-h^2+1)}{(xh_x-h)^2}\\&=\frac{\partial}{\partial x}\left(\frac{2xhh_x+x^2-h^2+1}{xh_x-h}\right),\end{aligned}$$

which implies that

$$\frac{2xhh_x+x^2-h^2+1}{xh_x-h}=2g_2(t)$$

for some function $g_2 : \mathbb{R} \to \mathbb{R}$. Multiplying both sides of the above equation by $\frac{h - xh_x}{x^2}$ shows that

$$\begin{aligned} 0 &= \frac{2g_2 h}{x^2} - \frac{2g_2 h_x}{x} + \frac{2hh_x}{x} - \frac{h^2}{x^2} + 1 + \frac{1}{x^2} \\ &= \frac{\partial}{\partial x}\left(-\frac{2g_2 h}{x} + \frac{h^2}{x} + x - \frac{1}{x}\right) \end{aligned}$$

and thus

$$\frac{-2g_2 h + x^2 + h^2 - 1}{x} = 2g_1$$

for some function $g_1(t)$. Hence we obtain

$$(x - g_1(t))^2 + (h - g_2(t))^2 = 1 + g_1^2(t) + g_2^2(t),$$

which is a horizontal geodesic in $\mathbb{S}^2 \times \mathbb{R}$ for each t. Therefore it follows that the harmonic minimal graph in $\mathbb{S}^2 \times \mathbb{R}$ is locally foliated by horizontal geodesics. Using the classification of ruled minimal surfaces in $\mathbb{S}^2 \times \mathbb{R}$ by [10] again, we obtain that the harmonic minimal graph over the domain is locally either a horizontal level surface $\mathbb{S}^2 \times \{t_0\}$ or a helicoid for some $t_0 \in \mathbb{R}$. This completes the proof. □

Remark The coordinate computations carried out in Sections 2 and 3 can be interpreted in a more geometric way. To see this, consider a harmonic minimal graph $t = f(x)$ in $M^n \times \mathbb{R}$. Then the minimal graph equation is given by

$$0 = \mathrm{div}_M\left(\frac{\nabla^M f}{\sqrt{1 + |\nabla^M f|^2}}\right),$$

where div_M and ∇^M denote the divergence and the gradient in M^n, respectively. This gives

$$(1 + |\nabla^M f|^2)\,\Delta_M f - \mathrm{Hess}_M f(\nabla^M f, \nabla^M f) = 0,$$

where Δ_M and Hess_M denote the Laplacian and the Hessian in M^n, respectively. Since $\Delta_M f = 0$ in M, we see that

$$\mathrm{Hess}_M f(\nabla^M f, \nabla^M f) = 0. \tag{14}$$

For a regular value t, denote by Σ_t the level hypersurface $\{f = t\} \subset M^n$. Choose an orthonormal frame $\{e_1, \ldots, e_{n-1}, e_n\}$ on M^n such that $e_1, \ldots, e_{n-1}$ are tangent to Σ_t and $e_n = \frac{\nabla^M f}{|\nabla^M f|}$. Then (14) implies $\mathrm{Hess}_M f(e_n, e_n) = 0$. From this, it follows that

$$0 = \Delta_M f = \sum_{i=1}^{n-1} \mathrm{Hess}_M f(e_i, e_i) + \mathrm{Hess}_M f(e_n, e_n)$$

$$= \sum_{i=1}^{n-1} \mathrm{Hess}_M f(e_i, e_i). \tag{15}$$

On the other hand, the mean curvature H_{Σ_t} of the level hypersurface $\Sigma_t = \{f = t\} \subset M^n$ is given by

$$H_{\Sigma_t} = \sum_{i=1}^{n-1} \mathrm{II}(e_i, e_i),$$

where the second fundamental form II on Σ_t in M^n is defined by

$$\mathrm{II}(X, Y) = \langle \nabla^M_X e_n, Y \rangle, \quad X, Y \in T\Sigma_t.$$

Moreover, for $X \in T\Sigma_t$, we see that

$$\begin{aligned} \nabla^M_X e_n &= \nabla^M_X \left(\frac{\nabla^M f}{|\nabla^M f|} \right) \\ &= \frac{1}{|\nabla^M f|} \nabla^M_X \nabla^M f + X\left(\frac{1}{|\nabla^M f|} \right) \nabla^M f \\ &= \frac{1}{|\nabla^M f|} \nabla^M_X \nabla^M f, \end{aligned}$$

which implies

$$\mathrm{II}(X, Y) = \frac{1}{|\nabla^M f|} \langle \nabla^M_X \nabla^M f, Y \rangle = \frac{1}{|\nabla^M f|} \mathrm{Hess}_M f(X, Y).$$

By (15), we obtain

$$H_{\Sigma_t} = \frac{1}{|\nabla^M f|} \sum_{i=1}^{n-1} \mathrm{Hess}_M f(e_i, e_i) = 0.$$

This shows that each regular level set Σ_t is a minimal hypersurface in M^n.

Acknowledgements The authors would like to express their sincere gratitude to the reviewer for his/her careful reading of our manuscript and helpful comments and valuable suggestions. The second author was supported by the National Research Foundation of Korea (NRF-2021R1A2C1003365).

Data Availability Data sharing not applicable to this article as no data sets were generated or analysed during the current study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Abresch, U., Rosenberg, H.: A Hopf differential for constant mean curvature surfaces in $\mathbb{S}^2 \times \mathbb{R}$ and $\mathbb{H}^2 \times \mathbb{R}$. Acta Math. **193**(2), 141–174 (2004)
2. Castro-Infantes, J., Manzano, J. M., Rodríguez, M.: *A construction of constant mean curvature surfaces in* $\mathbb{H}^2 \times \mathbb{R}$ *and the Krust property*, Int. Math. Res. Not. IMRN (2022), no. **19**, 14605-14638
3. Catalan, E.: Sur les surfaces réglées dont l'aire est un minimum. J. Math. Pure Appl. **7**, 203–211 (1842)
4. Coddington, E. A., Levinson, N.: *Theory of ordinary differential equations*, McGraw-Hill Book Co., Inc., New York-Toronto-London, 1955. xii+429 pp
5. Daniel, B.: Isometric immersions into $\mathbb{S}^n \times \mathbb{R}$ and $\mathbb{H}^n \times \mathbb{R}$ and applications to minimal surfaces. Trans. Amer. Math. Soc. **361**(12), 6255–6282 (2009)
6. Elbert, M.F., Rosenberg, H.: Minimal graphs in $M \times \mathbb{R}$. Ann. Global Anal. Geom. **34**(1), 39–53 (2008)
7. Hauswirth, L., Menezes, A., Rodríguez, M.: Slab theorem and halfspace theorem for constant mean curvature surfaces in $\mathbb{H}^2 \times \mathbb{R}$. Rev. Mat. Iberoam. **39**(1), 307–320 (2023)
8. Hauswirth, L., Rosenberg, H.: Minimal surfaces of finite total curvature in $\mathbb{H} \times \mathbb{R}$. Mat. Contemp. **31**, 65–80 (2006)
9. Jenkins, H., Serrin, J.: The Dirichlet problem for the minimal surface equation in higher dimensions. J. Reine Angew. Math. **229**, 170–187 (1968)
10. Kim, Y.W., Koh, S.-E., Shin, H., Yang, S.-D.: Helicoids in $\mathbb{S}^2 \times \mathbb{R}$ and $\mathbb{H}^2 \times \mathbb{R}$. Pacific J. Math. **242**(2), 281–297 (2009)
11. Kim, Y.W., Koh, S.-E., Shin, H., Yang, S.-D.: Helicoidal minimal surfaces in $\mathbb{H}^2 \times \mathbb{R}$. Bull. Aust. Math. Soc. **86**(1), 135–149 (2012)
12. Kim, Y.W., Lee, H.Y., Yang, S.-D.: Minimal harmonic graphs and their Lorentzian cousins. J. Math. Anal. Appl. **353**(2), 666–670 (2009)
13. Kobayashi, O.: Maximal surfaces in the 3- dimensional Minkowski space of $\mathbb{L}^3$. Tokyo J. Math. **6**(2), 297–309 (1983)
14. Meeks, W.H., Rosenberg, H.: The theory of minimal surfaces in $M \times \mathbb{R}$. Comment. Math. Helv. **80**(4), 811–858 (2005)
15. Nelli, B., Rosenberg, H.: *Minimal surfaces in* $\mathbb{H}^2 \times \mathbb{R}$, Bull. Braz. Math. Soc. (N.S.) **33** (2002), no. 2, 263-292
16. Pyo, J., Rodríguez, M.: *Simply connected minimal surfaces with finite total curvature in* $\mathbb{H}^2 \times \mathbb{R}$, Int. Math. Res. Not. IMRN(2014), no. 11, 2944-2954
17. Rosenberg, H.: Minimal surfaces in $M^2 \times \mathbb{R}$. Illinois J. Math. **46**(4), 1177–1195 (2002)
18. Sà Earp, R., Toubiana, E.: *Minimal graphs in* $\mathbb{H}^n \times \mathbb{R}$ *and* $\mathbb{R}^{n+1}$, Ann. Inst. Fourier (Grenoble) 60 (2010), no. 7, 2373-2402
19. Seo, K.: Sphere-foliated minimal and constant mean curvature hypersurfaces in product spaces. Bull. Korean Math. Soc. **48**(2), 335–342 (2011)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

신용현 교수

논문

- K. Lee, H. Liu, and Y.H. Shin (2025). An Optimal Portfolio Choice Problem with Delays. Mathematical Control and Related Fields. 15 (1), 115-142.

- Q. Li, Y.H. Shin, and J.-H. Yoon (2025). An Optimal Retirement Problem with Job Switching and Unemployment Risks under Subsistence Consumption Constraints. Quantitative Finance. 25 (5), 795--815.

- Q. Li, Y.H. Shin, and J.-H. Yoon (2025). Job Switching and Bequest Motives in an Optimal Consumption-Investment Model under Inflation and Mortality Risks. Economic Modelling. 153, 107307.

- Y.H. Shin and H.-S. Lee (2025). Work Hour Flexibility and Job Mobility. Finance Research Letters. 86 (D), 108573.

- T.U. Gang and Y.H. Shin (2026). Impact of Income and Leisure on Optimal Portfolio, Consumption, and Retirement Decisions under Exponential Utility. Journal of Computational and Applied Mathematics. 472, 116810.

- H.-T. Lee, B.-G. Jang, K.-H. Roh, and Y.H. Shin (2026). Household Consumption, Financial Constraints, and Size Structure of Rent. Finance Research Letters. 89, 109176.

Mathematical Control and Related Fields
Vol. 15, No. 1, March 2025, pp. 115-142
doi:10.3934/mcrf.2024003

AN OPTIMAL PORTFOLIO CHOICE PROBLEM WITH DELAYS

Kiseop Lee[1], Haibo Liu[2] and Yong Hyun Shin[*3]

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

[2]Department of Statistics and Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA

[3]Department of Mathematics & Research Institute of Natural Sciences, Sookmyung Women's University, Seoul 04310, Republic of Korea

(Communicated by Qing Zhang)

Abstract. We investigate the impact of the historical performance (delay) of a risky asset on the investment decisions of an investor in the context of terminal utility maximization. By assuming either a power-type or an exponential-type utility function, we derive a closed-form solution for the value function under mild assumptions. Our findings show that, in terms of the optimal portfolio ratio, an investor with power-type utility and delays tends to invest more in the risky asset compared to a CRRA investor without delays (as per Merton's model). Moreover, in terms of the optimal investment amount, an investor with exponential-type utility and delays tends to invest less in the risky asset compared to a CARA investor without delays (as per Merton's model). We also provide numerical simulation results to examine the influence of various key parameters on the wealth process and the optimal investment strategy.

1. **Introduction.** The majority of models used to analyze financial assets assume that the Markov property holds, meaning that future movements of stock prices are only dependent on their current value and hence do not depend on past history. However, empirical studies have shown that this assumption is often not accurate [4]. Despite this, the Markov assumption remains popular due to its mathematical tractability, which is difficult to abandon. To overcome the limitations of Markovian models, researchers have proposed alternative models such as fractional Brownian motion and stochastic delay models to capture short-term dependencies in asset price processes.

Models that use fractional Brownian motion are able to capture short-term dependencies by design [2], however, this approach also introduces a significant difficulty as it fails to be a semimartingale. In this study, we instead focus on a stochastic delay model, which is often more intuitive to understand and work with. Specifically, the performance of a risky asset is affected by its historical performance

2020 *Mathematics Subject Classification.* Primary: 91G10, 91G60; Secondary: 91B16.

Key words and phrases. Portfolio selection, delay for stock, dynamic programming method, power-type utility, exponential-type utility.

Haibo Liu acknowledges the financial support from a start-up research fund provided by Purdue University. Yong Hyun Shin acknowledges the financial support from the NRF grant funded by the Korea government (MSIT) [RS-2023-00240574].

*Corresponding author: Yong Hyun Shin.

through two stochastic delay variables defined in Remark 2.1; see Section 2 of [3] for related discussions on these variables.

Since the pioneering works of [7, 8], optimal portfolio selection has been a central topic of research in mathematical finance and financial economics. Recently, [3] examined the optimal consumption and investment problem for a risky asset with delays, and provided a closed-form solution for the Merton problem with a power-type utility function under certain assumptions. Since then, there have been numerous extensions to their work, including multi-dimensional cases [11], the incorporation of stochastic factors [1, 6], infinite time horizons [9], and complete memory [10], among others.

In this paper, we investigate a portfolio optimization problem under a stochastic delay model, with a focus on terminal utility maximization using power-type and exponential-type utility functions, based on the financial model of [3]. We derive closed-form solutions for the value function under both utility functions. While [10] previously obtained a time-separable value function for the optimal consumption and investment problem with complete memory and exponential-type utility (see Section 4 of [10]), it is difficult to conjecture the form of the solution for exponential-type utility. In this work, we demonstrate how to derive a non-time-separable solution for the terminal utility maximization (see Subsection 3.4). Additionally, we analyze the optimal portfolios in comparison to Merton's case, which refers to the terminal utility maximization without considering the effects of delays. Our results show that, in terms of the optimal portfolio ratio, an investor with power-type utility and delays tends to invest more in the risky asset compared to a constant relative risk aversion (CRRA) investor without delays (as per Merton's model), but, in terms of the optimal investment amount, an investor with exponential-type utility and delays tends to invest less in the risky asset compared to a constant absolute risk aversion (CARA) investor without delays (as per Merton's model).

Moreover, we simulate sample paths of the total wealth of the investor $X(t)$ in (6), the exponential moving average of the total wealth $Y(t)$ in (1) and the parallel movement of historical wealth $Z(t)$ in (2) for different values of various parameters h, λ and μ_3 for each utility function. For both a power-type and an exponential-type utility function, we find that μ_3 has a greater impact on the wealth $X(t)$ than h and λ. We also simulate sample paths of the optimal portfolio π_t^* for different values of h, λ and μ_3 for each utility function. For a power-type utility function, we see that h and μ_3 have larger effects on the optimal portfolio π_t^* compared to λ. For an exponential-type utility function, we see that μ_3 has a greater impact on the optimal portfolio π_t^* compared to h and λ. We also find that μ_3 plays different roles in power-type and exponential-type utilities. With power-type utility, we see that π_t^* increases as μ_3 increases, but with exponential-type utility, we see that π_t^* decreases as μ_3 increases.

The rest of the paper is organized as follows. We introduce the stochastic delay model in Section 2. We solve two optimization problems under different utility functions in Section 3 and provide some related simulation results. We discuss further in Section 4.

2. **The model and the economy.**

We consider a financial market within a finite-time horizon $T > 0$. In the market, there are one risky asset and one riskless asset that grows at a constant interest rate $r > 0$. The investor can freely invest in these assets. There is no short sale constraint.

Remark 2.1 (Delay Variables $Y(t)$ and $Z(t)$ [3])**.**
We assume that the investor decides the amount of capital invested in the risky asset based on the past performance of his/her portfolio, and that the past performance of the portfolio affects the dynamics through the delay variables $Y(t)$ and $Z(t)$ defined as follows:

$$Y(t) \equiv \int_{-h}^{0} e^{\lambda u} X(t+u)du, \text{ for } \lambda > 0, \ h > 0, \tag{1}$$

$$Z(t) \equiv X(t-h), \ \forall \ t \in [s, T], \tag{2}$$

where $X(t)$ is the portfolio value (or the total wealth of the investor) at time t. In this setting, the performance of the risky asset has delay; the reader is referred to pages 605 and 606 of [3] for related discussions.

Notice that $Y(t)$ is the exponential moving average of the total wealth, h represents the duration of the delay effect, λ determines how fast the past dependency diminishes, and $Z(t)$ is simply the historical wealth at time $t-h$.

Denote by $\Pi_t^{(0)}$ and $\Pi_t^{(1)}$ the amount of capital invested in the riskless asset and that invested in the risky asset at time t, respectively. We assume that

$$d\Pi_t^{(0)} = \left[r\Pi_t^{(0)} - I_t \right] dt, \tag{3}$$

$$d\Pi_t^{(1)} = \left[\mu_1 \Pi_t^{(1)} + \mu_2 Y(t) + \mu_3 Z(t) + I_t \right] dt + \sigma \Pi_t^{(1)} dB_t, \tag{4}$$

where μ_1, μ_2, μ_3 and σ are positive constants, I_t is the investment rate on the risky asset at time t, and $\{B_t\}_{0 \le t \le T}$ is a standard Brownian motion defined on the complete filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{0 \le t \le T}, \mathbb{P})$ in which $\{\mathcal{F}_t\}_{0 \le t \le T}$ is the $\mathbb{P}$-augmented natural filtration generated by $\{B_t\}_{0 \le t \le T}$. Moreover, we make the following assumption throughout the paper:

Assumption 2.2. *The excess return of the risky asset is positive, that is,*

$$\mu_1 - r > 0.$$

Note from (4) that the dynamics of the wealth invested in the risky asset $\Pi^{(1)}$ depends on the past performance of the portfolio through delay variables $Y(t)$ and $Z(t)$. The role of $Y(t)$ is the time adjusted cumulative effect over a recent period of length h and the role of $Z(t)$ is the impact of the historical wealth h time units ago. Therefore, $\Pi^{(1)}$ is impacted both by the historical wealth at time $t-h$ and the cumulative one between time $t-h$ and t.

The total wealth of the investor at time t is given by

$$X(t) = \Pi_t^{(0)} + \Pi_t^{(1)},$$

and it follows from (3) and (4) that the wealth process $\{X(t)\}_{0 \le t \le T}$ evolves according to the stochastic differential equation (SDE):

$$dX(t) = \left[rX(t) + (\mu_1 - r)\, \Pi_t^{(1)} + \mu_2 Y(t) + \mu_3 Z(t) \right] dt + \sigma \Pi_t^{(1)} dB_t. \tag{5}$$

Denote the proportion of wealth invested in the risky asset at time t by

$$\pi_t \equiv \frac{\Pi_t^{(1)}}{X(t)}.$$

Then the dynamics of the wealth process (5) can be rewritten as

$$dX(t) = \Big[rX(t) + (\mu_1 - r)\pi_t X(t) + \mu_2 Y(t) + \mu_3 Z(t)\Big]dt + \sigma\pi_t X(t)dB_t, \ \forall\, t \in [s,T] \tag{6}$$

with the initial condition

$$X(t) = \varphi(t-s), \ \forall\, t \in [s-h, s]. \tag{7}$$

In the above, $\varphi \in C[-h,0]$ and $\varphi(u) > 0$ for $u \in [-h,0]$.

Remark 2.3. Since $\varphi \in C[-h,0]$, there exists $M > 0$ such that

$$||\varphi|| \equiv \max_{u\in[-h,0]} |\varphi(u)| = M < \infty.$$

Apart from Assumption 2.2, we also make the following assumption:

Assumption 2.4. *The following equality holds:*

$$\mu_2 = \Big(r + \lambda + \mu_3 e^{\lambda h}\Big)\mu_3 e^{\lambda h}. \tag{8}$$

We refer the reader to Section 5 of [3] for a detailed discussion on Assumption 2.4.

The objective of the investor is to maximize his/her expected terminal utility function with $T > 0$ subject to (6) with (7). We assume that the terminal utility function $U(\cdot,\cdot)$ is a function of $X(T)$ and $Y(T)$, so the objective function is given by

$$V(s,\varphi) = \sup_{\pi\in\mathcal{A}} \mathbb{E}^{s,\varphi,\pi}\Big[U(X(T), Y(T))\Big],$$

where $\mathcal{A}$ is the set of admissible investment strategies (see Definitions 3.2 and 3.12 for details).

Remark 2.5. Following [3], we can write the value function $V(s,\varphi)$ as $V(s,x,y,z)$, i.e., $V(s,\varphi) = V(s,x,y,z)$. Moreover, under Assumption 2.4 the value function depends only on s,x,y, but not on z, i.e., $V(s,x,y,z) = V(s,x,y)$. Therefore, we have

$$V(s,x,y) = V(s,\varphi) = \sup_{\pi\in\mathcal{A}} \mathbb{E}^{s,\varphi,\pi}\Big[U(X(T), Y(T))\Big]. \tag{9}$$

In the sequel, we study the optimization problem (9) with two types of utility functions: power-type and exponential-type.

3. Optimization problems.

In this section, we formulate and solve optimization problems under two different utility functions and present some numerical simulation results.

3.1. The optimization problem with power-type utility.

We first consider a power-type utility function defined below.

Definition 3.1 (Power-Type Utility Function)**.**
The utility function is given by

$$U(x,y) \equiv \frac{\big(x + \mu_3 e^{\lambda h} y\big)^{1-\gamma}}{1-\gamma}, \tag{10}$$

where $\gamma > 0\,(\gamma \neq 1)$ is the risk aversion coefficient.

With the power-type utility function (10), the objective function (9) can be written as

$$V(s,x,y)=\sup_{\pi\in\mathcal{A}}\mathbb{E}^{s,\varphi,\pi}\left[\frac{\left(X(T)+\mu_3e^{\lambda h}Y(T)\right)^{1-\gamma}}{1-\gamma}\right]. \tag{11}$$

We next define the admissible set of investment strategies for the optimization problem (11).

Definition 3.2 (Admissible Set $\mathcal{A}$)**.**
An investment strategy $\pi\equiv\{\pi_t\}_{0\le t\le T}$ is admissible if

(1) π_t is $\mathcal{F}_t$-measurable for all $t\in[0,T]$;
(2) there exists a constant $\Lambda>0$ such that

$$|\pi_tX(t)|\le\Lambda\left|X(t)+\mu_3e^{\lambda h}Y(t)\right|,\qquad\forall t\in[0,T].$$

The optimization problem (11) can be solved similarly as the one in [3]. For the sake of completeness, we provide the detailed proof as follows. Analogous to Lemmas 2.1 and 2.2 of [3], we obtain the following lemma:

Lemma 3.3. *For any admissible control policy π, the SDE (6) with the initial condition (7) admits a unique strong solution $X:[s-h,T]\times\Omega\to\mathbb{R}$. For any positive integer $k>0$, we have*

$$\mathbb{E}||X_t||^{2k}\le C_k\left(1+||\varphi||^{2k}\right),\qquad\forall t\in[s,T],$$

where $X_t(u)=X(t+u)$, for $u\in[-h,0]$, $C_k>0$ is a constant, and $||\cdot||$ is the sup norm. Moreover, $X(t)>0$ a.s. for all $t\in[s,T]$.

The following lemma is a version of Itô's Formula, which is also Lemma 3.1 of [3].

Lemma 3.4 (Itô's Formula)**.**
Let $f(t,x,y)\in C^{1,2,1}\left([0,T]\times\mathbb{R}^2\right)$ and $G(t)=f(t,X(t),y(X_t))$. Under the SDE (6) with (7), we have

$$\begin{aligned}dG(t)=&\left[f_t+f_x\cdot\left\{rx+(\mu_1-r)\pi x+\mu_2y+\mu_3z\right\}+\frac{1}{2}\sigma^2\pi^2x^2\cdot f_{xx}\right]dt\\&+f_y\cdot\left[x-\lambda y-e^{-\lambda h}z\right]dt+f_x\cdot\sigma\pi x\,dB_t,\end{aligned}$$

where the functions on the right-hand side are evaluated at

$$x=X(t),\ X_t(u)=X(t+u),\ y=y(X_t)=\int_{-h}^{0}e^{\lambda u}X(t+u)du,$$

$$z=z(X_t)=X(t-h),\ \pi=\pi_t.$$

Lemma 3.4 (Itô's Formula) gives us the Hamilton-Jacobi-Bellman (HJB) equation of the optimization problem (9) and we describe it in the next Proposition.

Proposition 3.5 (Hamilton-Jacobi-Bellman Equation)**.**
Assume that $V(s,x,y)\in C^{1,2,1}\left([0,T]\times\mathbb{R}^2\right)$. Then we have the following HJB equation:

$$\begin{aligned}\sup_{\pi}\Big[&V_s+\left\{rx+(\mu_1-r)\pi x+\mu_2y+\mu_3z\right\}V_x\\&+\left\{x-\lambda y-e^{-\lambda h}z\right\}V_y+\frac{1}{2}\sigma^2\pi^2x^2V_{xx}\Big]=0,\end{aligned} \tag{12}$$

with the terminal condition

$$V(T,x,y) = U(x,y). \tag{13}$$

We proceed to solve the HJB equation (12) with the power-type utility function (10).

Theorem 3.6. *The solution to the HJB equation (12) with the power-type terminal utility function*

$$U(x,y) = \frac{\left(x+\mu_3 e^{\lambda h} y\right)^{1-\gamma}}{1-\gamma}$$

is given by

$$V(s,x,y) = e^{-\gamma K(T-s)} \cdot \frac{\left(x+\mu_3 e^{\lambda h} y\right)^{1-\gamma}}{1-\gamma}, \tag{14}$$

where

$$K \equiv \frac{\gamma-1}{\gamma}\left(r+\mu_3 e^{\lambda h}\right) + \frac{\gamma-1}{2\gamma^2}\theta^2 \tag{15}$$

and $\theta \equiv (\mu_1 - r)/\sigma$ is the market price of risk. Furthermore, the optimal portfolio ratio π_s^* at time s is given by

$$\pi_s^* = \frac{\theta(X(s) + \mu_3 e^{\lambda h} Y(s))}{\sigma\gamma X(s)}. \tag{16}$$

Proof. First we consider the first order condition (FOC) with respect to π. From (12), we have

$$\pi^* = -\frac{(\mu_1 - r)V_x}{\sigma^2 x V_{xx}} = -\frac{\theta V_x}{\sigma x V_{xx}}, \tag{17}$$

where $\theta \equiv (\mu_1 - r)/\sigma$ is the market price of risk. Plugging (17) into (12), we obtain

$$0 = V_s + (rx + \mu_2 y + \mu_3 z)\, V_x + \left(x - \lambda y - e^{-\lambda h} z\right) V_y - \frac{1}{2}\theta^2 \frac{(V_x)^2}{V_{xx}}. \tag{18}$$

We conjecture the solution form as

$$V(s,x,y) = A(s)U(x,y). \tag{19}$$

Then we have

$$V_s = A'(s)U(x,y), \quad V_x = A(s)U_x(x,y) = A(s)\frac{1-\gamma}{x+\mu_3 e^{\lambda h} y}U(x,y),$$

$$V_{xx} = A(s)U_{xx}(x,y) = -A(s)\frac{\gamma(1-\gamma)}{(x+\mu_3 e^{\lambda h} y)^2}U(x,y),$$

$$V_y = A(s)U_y(x,y) = A(s)\frac{\mu_3 e^{\lambda h}(1-\gamma)}{x+\mu_3 e^{\lambda h} y}U(x,y), \quad \frac{(V_x)^2}{V_{xx}} = -\frac{1-\gamma}{\gamma}A(s)U(x,y).$$

Plugging the above equations into (18) gives the first-order ordinary differential equation (ODE) as follows:

$$\begin{aligned} 0 &= A'(s) + \Bigg[(rx + \mu_2 y + \mu_3 z) + (x - \lambda y - e^{-\lambda h} z)\mu_3 e^{\lambda h} \\ &\quad + \frac{1}{2\gamma}\theta^2\left(x + \mu_3 e^{\lambda h} y\right)\Bigg]\frac{1-\gamma}{x+\mu_3 e^{\lambda h} y}A(s) \\ &= A'(s) + \Big[\left(r + \mu_3 e^{\lambda h}\right)x + \left(\mu_2 - \lambda\mu_3 e^{\lambda h}\right)y \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2\gamma}\theta^2\left(x+\mu_3 e^{\lambda h}y\right)\Bigg]\frac{1-\gamma}{x+\mu_3 e^{\lambda h}y}A(s)\\
&=A'(s)+\Bigg[\left(r+\mu_3 e^{\lambda h}\right)x+\left(r+\mu_3 e^{\lambda h}\right)\mu_3 e^{\lambda h}y\\
&\quad+\frac{1}{2\gamma}\theta^2\left(x+\mu_3 e^{\lambda h}y\right)\Bigg]\frac{1-\gamma}{x+\mu_3 e^{\lambda h}y}A(s)\\
&=A'(s)+\left[(1-\gamma)\left(r+\mu_3 e^{\lambda h}\right)+\frac{1-\gamma}{2\gamma}\theta^2\right]A(s)\\
&=A'(s)-\gamma K A(s), \qquad (20)
\end{aligned}
$$

where the third equality is obtained from (8) in Assumption 2.4 and the last equality is due to (15). It is trivial that the solution $A(s)$ to the ODE (20) is given by

$$A(s)=D_1 e^{\gamma K s},$$

for some constant D_1. From (13) and (19), we see that

$$V(T,x,y)=U(x,y)=A(T)U(x,y),$$

implying $A(T)=1$, and hence

$$A(s)=e^{-\gamma K(T-s)}.$$

Combining the above results,

$$V(s,x,y)=A(s)U(x,y)=e^{-\gamma K(T-s)}\cdot\frac{\left(x+\mu_3 e^{\lambda h}y\right)^{1-\gamma}}{1-\gamma}$$

is the solution to the HJB equation (12) and the optimal portfolio ratio π^* in (17) is given by

$$\pi^*=-\frac{\theta V_x}{\sigma x V_{xx}}=\frac{\theta(x+\mu_3 e^{\lambda h}y)}{\sigma\gamma x}.$$

This completes the proof. □

To solve the optimization problem (11), we also provide the following verification theorem:

Theorem 3.7 (Verification Theorem)**.**
Let $X(t)$ be a strong solution to (6) and (7) with $Y(t)$ in (1) and $Z(t)$ in (2). Assume that $g(s,x,y)\in C^{1,2,1}([0,T]\times\mathbb{R}^2)$ is a solution to the HJB equation (12) with the terminal condition (13) and the power-type utility function (10), and that

$$\mathbb{E}\left[\int_0^T\left(\pi_t X(t)g_x(t,X(t),Y(t))\right)^2 dt\right]<\infty, \qquad (21)$$

for any admissible control policy π. Then we have

$$g(s,x,y)\geq\sup_{\pi\in\mathcal{A}}\mathbb{E}^{s,\varphi,\pi}\left[\frac{\left(X(T)+\mu_3 e^{\lambda h}Y(T)\right)^{1-\gamma}}{1-\gamma}\right]. \qquad (22)$$

Let $\hat{\pi}=\{\hat{\pi}(t,X(t),Y(t))\}_{0\leq t\leq T}$, where

$$\hat{\pi}(t,x,y)=-\frac{\theta g_x(t,x,y)}{\sigma x g_{xx}(t,x,y)},\qquad (t,x,y)\in[0,T]\times\mathbb{R}^2. \qquad (23)$$

If $\hat{\pi}$ is admissible, then $\hat{\pi}$ is the optimal control policy. Moreover, when $\hat{\pi}$ is taken as the control policy, we have

$$g(s,x,y) = \mathbb{E}^{s,\varphi,\hat{\pi}}\left[\frac{\left(X(T)+\mu_3 e^{\lambda h}Y(T)\right)^{1-\gamma}}{1-\gamma}\right]. \tag{24}$$

Proof. Let $g(s,x,y)$ be a solution to the HJB equation (12) with (13). For any given admissible control policy π and $(s,x,y)\in[0,T]\times\mathbb{R}^2$, we have

$$g_s + \{rx + (\mu_1 - r)\pi x + \mu_2 y + \mu_3 z\}\, g_x + \left\{x - \lambda y - e^{-\lambda h} z\right\} g_y + \frac{1}{2}\sigma^2\pi^2 x^2 g_{xx} \le 0.$$

It follows from Itô's Formula in Lemma 3.4 that

$$\begin{aligned}
&dg(t,X(t),Y(t))\\
&= \Big[g_t + \{rX(t) + (\mu_1 - r)\pi_t X(t) + \mu_2 Y(t) + \mu_3 Z(t)\}\, g_x\\
&\quad + \left\{X(t) - \lambda Y(t) - e^{-\lambda h} Z(t)\right\} g_y\Big] dt\\
&\quad + \frac{1}{2}\sigma^2\pi_t^2 X^2(t) g_{xx} dt + \sigma\pi_t X(t) g_x dB_t.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&g(T,X(T),Y(T)) - g(s,x,y)\\
&= \int_s^T \Big[g_t + \{rX(t) + (\mu_1 - r)\pi_t X(t) + \mu_2 Y(t) + \mu_3 Z(t)\}\, g_x\\
&\quad + \left\{X(t) - \lambda Y(t) - e^{-\lambda h} Z(t)\right\} g_y\Big] dt\\
&\quad + \int_s^T \frac{1}{2}\sigma^2\pi_t^2 X^2(t) g_{xx} dt + \int_s^T \sigma\pi_t X(t) g_x dB_t\\
&\le \int_s^T \sigma\pi_t X(t) g_x dB_t.
\end{aligned} \tag{25}$$

From (13), we see that $g(T,X(T),Y(T)) = U(X(T),Y(T))$. Also, (21) implies that

$$\int_s^T \sigma\pi_t X(t) g_x dB_t$$

is an $\{\mathcal{F}_t\}_{0\le t\le T}$ martingale under $\mathbb{P}$. By taking expectations on both sides of the inequality (25), we get

$$\begin{aligned}
g(s,x,y) \ge& \mathbb{E}^{s,\varphi,\pi}\Big[U(X(T),Y(T))\Big]\\
=& \mathbb{E}^{s,\varphi,\pi}\left[\frac{\left(X(T)+\mu_3 e^{\lambda h}Y(T)\right)^{1-\gamma}}{1-\gamma}\right].
\end{aligned} \tag{26}$$

Since the inequality (26) holds for any admissible control policy π, (22) holds.

Suppose the utility function is given by (10), the control policy $\hat{\pi}$ is given by (23), and $\hat{\pi}$ is admissible. Replace π by $\hat{\pi}$ in the above proof, then all the above inequalities become equalities, which completes the proof of (24). □

Utilizing Theorems 3.6 and 3.7, we solve the optimization problem (11) below.

Theorem 3.8. *Assume the utility function is given by (10) and the condition (8) holds. Then the function $V(s,x,y)$ given by (14) is a solution to the HJB equation (12) with (13), and is equal to the value function defined in (11). Moreover, the optimal control policy is given by (16).*

Proof. By Theorem 3.6, $V(s,x,y)$ given by (14) is a solution to the HJB equation (12) with (13). Moreover, the associated optimal portfolio ratio π^* is given by (16).

For the remainder of the proof, we apply Theorem 3.7 to derive the solution of the optimization problem (11). We first show that (21) holds for any admissible control policy π when $g(s,x,y) \equiv V(s,x,y)$. From (14),

$$V_x(t,X(t),Y(t)) = A(t)\left(X(t)+\mu_3 e^{\lambda h}Y(t)\right)^{-\gamma}.$$

By Lemma 3.3, $X(t) > 0$ a.s. and $Y(t) > 0$ a.s. From Definition 3.2, we have

$$|\pi_t X(t)| \le \Lambda \left|X(t)+\mu_3 e^{\lambda h}Y(t)\right|.$$

Recall $Y(t)$ defined in (1). Then

$$|Y(t)| \le \int_{-h}^{0} |X(t+u)|du \le h \max_{u\in[-h,0]} |X(t+u)| = h\,||X_t||. \tag{27}$$

Clearly, $|X(t)| \le ||X_t||$. Recall $A(t) = e^{-\gamma K(T-t)}$, which is bounded on $[0,T]$. Therefore,

$$\begin{aligned}
|\pi_t X(t) V_x(t,X(t),Y(t))| &\le \Lambda \left|X(t)+\mu_3 e^{\lambda h}Y(t)\right| A(t) \left|X(t)+\mu_3 e^{\lambda h}Y(t)\right|^{-\gamma} \\
&= \Lambda\, A(t) \left|X(t)+\mu_3 e^{\lambda h}Y(t)\right|^{1-\gamma} \\
&\le \Lambda\, A(t) \Big(|X(t)| + \mu_3 e^{\lambda h} |Y(t)| \Big)^{1-\gamma} \\
&\le \Lambda_1 ||X_t||^{1-\gamma},
\end{aligned}$$

where Λ_1 is a constant that does not depend on t. By Lemma 3.3 and noting that $\gamma > 0$, we derive

$$\begin{aligned}
\mathbb{E}\left[\int_0^T (\pi_t X(t) V_x(t,X(t),Y(t)))^2 dt\right] &\le \Lambda_1^2\, \mathbb{E}\left[\int_0^T ||X_t||^{2(1-\gamma)} dt\right] \\
&\le \Lambda_1^2\, \mathbb{E}\left[\int_0^T (1+||X_t||)^{2(1-\gamma)} dt\right] \\
&\le \Lambda_1^2\, \mathbb{E}\left[\int_0^T (1+||X_t||)^{2} dt\right] \\
&\le \Lambda_1^2\, \mathbb{E}\left[\int_0^T \left(2+2||X_t||^2\right) dt\right] \\
&= 2\Lambda_1^2 \int_0^T \mathbb{E}\left[1+||X_t||^2\right] dt \\
&\le 2\Lambda_1^2 T \left[1 + C_1\left(1+||\varphi||^2\right)\right]
\end{aligned}$$

$$< \infty,$$

where $C_1 > 0$ is a positive constant. Therefore, (21) holds for any admissible control policy π.

Clearly, the control policy π^* given by (16) satisfies (23) with $g(s,x,y) \equiv V(s,x,y)$. We proceed to show that π^* is admissible. It is easy to see that π_t^* is $\mathcal{F}_t$-measurable for $t \in [0,T]$. Note that

$$\pi_t^* X(t) = \frac{\theta\left[X(t) + \mu_3 e^{\lambda h} Y(t)\right]}{\sigma\gamma}.$$

Since $X(t) > 0$ and $Y(t) > 0$ a.s., it is easy to see from Lemma 3.3 that the second condition in Definition 3.2 is also satisfied. Thus, π^* is admissible.

From the above results and Theorem 3.7, the function $V(s,x,y)$ given by (14) is equal to the value function defined in (11) and π^* is the optimal policy. □

3.2. **The value function and optimal portfolio ratio with power-type Utility.** First we consider the original Merton problem (without delays). Our model becomes the original Merton model if we do not consider the effects of delays of the risky asset, that is, if we let $\mu_2 \equiv 0$ and $\mu_3 \equiv 0$ in the value function (11) and the wealth process (6).

Remark 3.9 (Merton's Problem with CRRA)**.**
The investor seeks to maximize his/her expected terminal utility function. Specifically, the objective function is given by

$$V^M(s,x) = \sup_{\pi\in\mathcal{A}} \mathbb{E}^{s,X(s)=x}\left[\frac{(X(T))^{1-\gamma}}{1-\gamma}\right]$$

subject to

$$dX(t) = \Big[rX(t) + (\mu_1 - r)\pi_t X(t)\Big]dt + \sigma\pi_t X(t) dB_t,\ X(s) = x,\ \forall\, t \in [s,T].$$

Then the associated HJB equation is derived as

$$\sup_{\pi}\left[V_s^M + \{rx + (\mu_1 - r)\pi x\} V_x^M + \frac{1}{2}\sigma^2\pi^2 x^2 V_{xx}^M\right] = 0, \tag{28}$$

with the terminal condition

$$V^M(T,x) = \frac{x^{1-\gamma}}{1-\gamma}. \tag{29}$$

If we consider the FOC with respect to π, then, from (28), we obtain

$$\pi^{M*} = -\frac{(\mu_1 - r)V_x^M}{\sigma^2 x V_{xx}^M} = -\frac{\theta V_x^M}{\sigma x V_{xx}^M}. \tag{30}$$

From (28)–(30), we obtain

$$V^M(s,x) = e^{-\gamma K_M (T-s)} \cdot \frac{x^{1-\gamma}}{1-\gamma}, \tag{31}$$

where

$$K_M \equiv \frac{\gamma - 1}{\gamma} r + \frac{\gamma - 1}{2\gamma^2}\theta^2. \tag{32}$$

Also using the FOC in (30) and the value function (31), we derive the optimal portfolio ratio at time s as

$$\pi_s^{M*} = \frac{\theta}{\sigma\gamma}. \tag{33}$$

We next compare the value function (14) and optimal portfolio (16) with Merton's (31) and (33), respectively.

Proposition 3.10. *For $0 \le s \le T$ and $x, y > 0$, we have*

$$V(s,x,y) = e^{-\gamma K(T-s)} \cdot \frac{\left(x + \mu_3 e^{\lambda h} y\right)^{1-\gamma}}{1-\gamma} > e^{-\gamma K_M (T-s)} \cdot \frac{x^{1-\gamma}}{1-\gamma} = V^M(s,x).$$

Furthermore, it holds almost surely that

$$\pi_s^* = \frac{\theta}{\sigma\gamma} + \frac{\theta \mu_3 e^{\lambda h}}{\sigma\gamma} \cdot \frac{Y(s)}{X(s)} > \frac{\theta}{\sigma\gamma} = \pi_s^{M*}.$$

Proof. From (15) and (32), we see that for $\gamma > 1$,

$$K > K_M > 0 \quad \Rightarrow \quad e^{-\gamma K(T-s)} < e^{-\gamma K_M (T-s)}, \tag{34}$$

and that for $0 < \gamma < 1$,

$$K < K_M < 0 \quad \Rightarrow \quad e^{-\gamma K(T-s)} > e^{-\gamma K_M (T-s)}. \tag{35}$$

Also since $\mu_3 > 0$ is a positive constant, we see that for $\gamma > 1$,

$$\left(x + \mu_3 e^{\lambda h} y\right)^{1-\gamma} < x^{1-\gamma}, \tag{36}$$

and that for $0 < \gamma < 1$,

$$\left(x + \mu_3 e^{\lambda h} y\right)^{1-\gamma} > x^{1-\gamma}. \tag{37}$$

Thus (34) and (36) imply that for $\gamma > 1$,

$$e^{-\gamma K(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)^{1-\gamma} < e^{-\gamma K_M (T-s)} x^{1-\gamma},$$

and (35) and (37) imply that for $0 < \gamma < 1$,

$$e^{-\gamma K(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)^{1-\gamma} > e^{-\gamma K_M (T-s)} x^{1-\gamma}.$$

From the above results, for any $\gamma > 0$ ($\gamma \neq 1$),

$$V(s,x,y) = e^{-\gamma K(T-s)} \cdot \frac{\left(x + \mu_3 e^{\lambda h} y\right)^{1-\gamma}}{1-\gamma} > e^{-\gamma K_M (T-s)} \cdot \frac{x^{1-\gamma}}{1-\gamma} = V^M(s,x).$$

From Lemma 3.3, (16) and (33), we see that

$$\begin{aligned} \pi_s^* &= \frac{\theta(X(s) + \mu_3 e^{\lambda h} Y(s))}{\sigma\gamma X(s)} \\ &= \frac{\theta}{\sigma\gamma} + \frac{\theta \mu_3 e^{\lambda h}}{\sigma\gamma} \cdot \frac{Y(s)}{X(s)} \\ &> \frac{\theta}{\sigma\gamma} = \pi_s^{M*} \end{aligned}$$

almost surely. This completes the proof. □

From Proposition 3.10, we see that the investor with power-type utility and delays invests more in the risky asset than the CRRA investor without delays. The portfolio ratio π_s^* of the investor with power-type utility and delays depends on the ratio $Y(s)/X(s)$. As $Y(s)/X(s)$ increases, π_s^* also increases when the parameters are fixed. This shows that there is some delay effect for the investor with power-type utility and delays. Also, we see that as μ_3 increases, π_s^* increases as well holding other parameters fixed. As $Y(s)/X(s)$ decreases to zero, we see that π_s^* converges to π_s^{M*}.

3.3. **Numerical simulations with power-type utility.** In this subsection, we perform numerical simulations to illustrate our results in Subsections 3.1 and 3.2.

From (6) and (16), we obtain the optimal wealth process as follows: For $t \in [0, T]$,

$$\begin{aligned}
dX(t) &= \Big[rX(t) + (\mu_1 - r)\pi_t^* X(t) + \mu_2 Y(t) + \mu_3 Z(t)\Big]dt + \sigma\pi_t^* X(t) dB_t \\
&= \left[rX(t) + (\mu_1 - r)\frac{\theta(X(t) + \mu_3 e^{\lambda h} Y(t))}{\sigma\gamma X(t)} X(t) + \mu_2 Y(t) + \mu_3 Z(t)\right]dt \\
&\quad + \sigma\frac{\theta(X(s) + \mu_3 e^{\lambda h} Y(s))}{\sigma\gamma X(t)} X(t) dB_t \\
&= \left[rX(t) + \frac{\theta^2(X(t) + \mu_3 e^{\lambda h} Y(t))}{\gamma} + \mu_2 Y(t) + \mu_3 Z(t)\right]dt \\
&\quad + \frac{\theta(X(t) + \mu_3 e^{\lambda h} Y(t))}{\gamma} dB_t \\
&= \left[\left(r + \frac{\theta^2}{\gamma}\right) X(t) + \left(\mu_2 + \frac{\mu_3\theta^2 e^{\lambda h}}{\gamma}\right) Y(t) + \mu_3 Z(t)\right]dt \\
&\quad + \left[\frac{\theta}{\gamma} X(t) + \frac{\mu_3\theta e^{\lambda h}}{\gamma} Y(t)\right] dB_t. \qquad (38)
\end{aligned}$$

To simulate sample paths of $X(t)$, $Y(t)$ and $Z(t)$, we divide the time period $[0, T]$ into N intervals with the same length $\Delta t = T/N$. Benchmark parameters are given as follows:

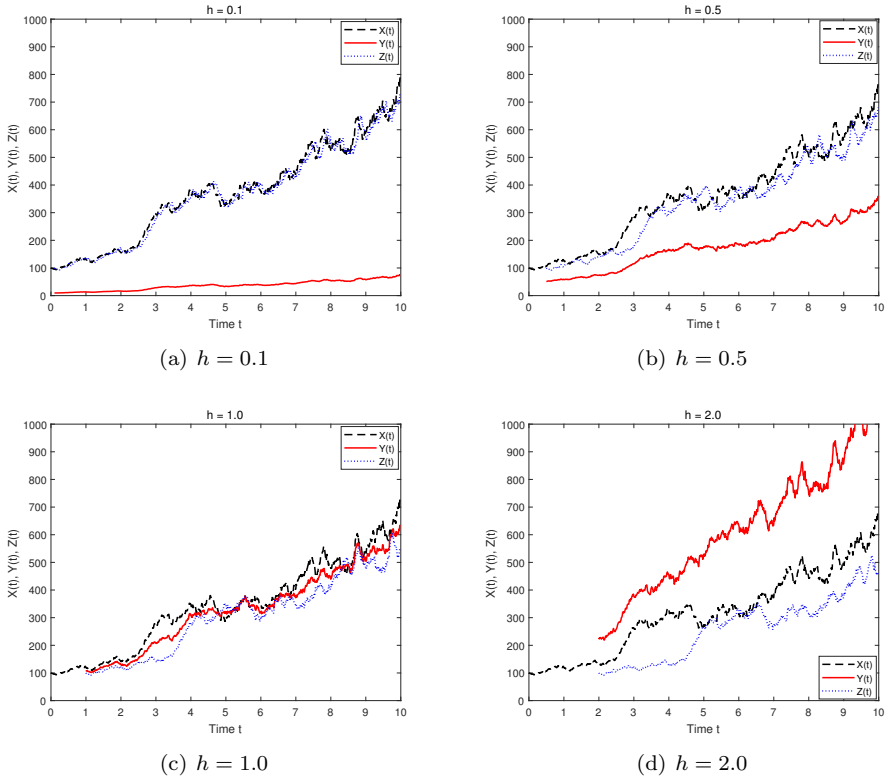
$$T = 10,\ N = 1000,\ X(0) = 100,\ r = 0.02,\ \mu_1 = 0.08,$$

$$\sigma = 0.2,\ \gamma = 2,\ \lambda = 0.01,\ h = 0.5,\ \mu_3 = 0.1.$$

We assume that $Y(t) = Z(t) \equiv 0$ for $t \in [0, h]$ and use the Trapezoidal rule to compute $Y(t)$ in (1). In light of (38), we approximate the optimal wealth process by

$$\begin{aligned}
X(t_i) &= X(t_{i-1}) + \left(r + \frac{\theta^2}{\gamma}\right) X(t_{i-1})\Delta t + \left(\mu_2 + \frac{\mu_3\theta^2 e^{\lambda h}}{\gamma}\right) Y(t_{i-1})\Delta t \\
&\quad + \mu_3 Z(t_{i-1})\Delta t \\
&\quad + \left[\frac{\theta}{\gamma} X(t_{i-1}) + \frac{\mu_3\theta e^{\lambda h}}{\gamma} Y(t_{i-1})\right]\sqrt{\Delta t}\,\epsilon_i,\ \epsilon_i \sim \mathcal{N}(0,1),\ i = 1, 2, \ldots, N.
\end{aligned}$$

For comparison purposes, the sample paths in all plots of each figure below are based on the same set of simulated random variables $\epsilon_i,\ i = 1, 2, \ldots, N$.



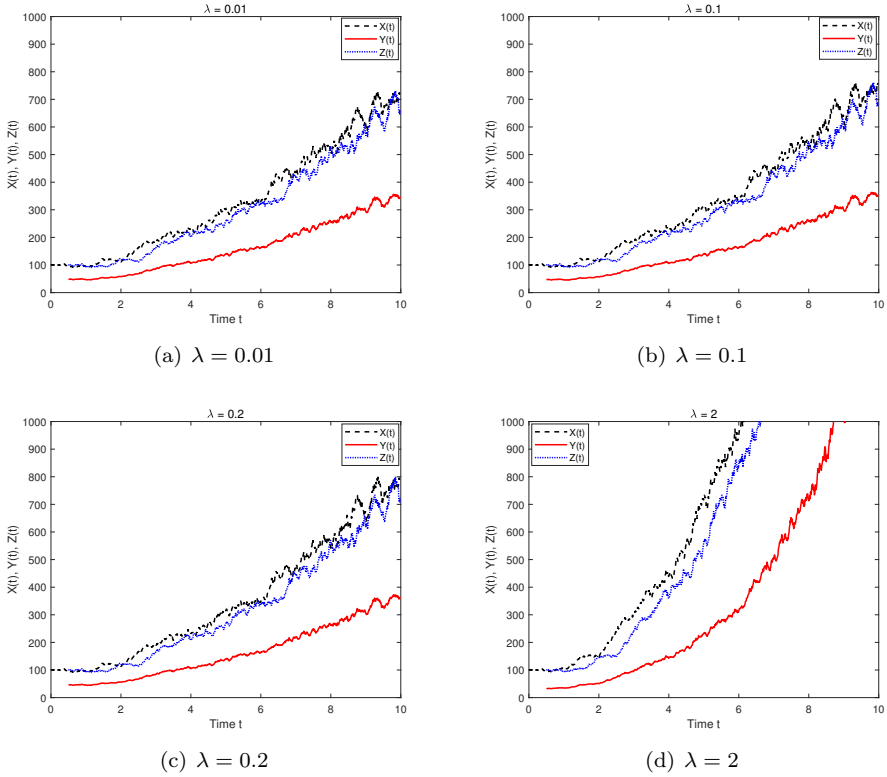
(a) $h = 0.1$ (b) $h = 0.5$

(c) $h = 1.0$ (d) $h = 2.0$

FIGURE 1. Sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of h under the power-type utility function (10)

In Figure 1, we simulate sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of h. From the figure, we observe that $Y(t)$ increases when h increases, which can be explained by its definition (1). Since the drift term in (38) is positive, $X(t)$ exhibits a growing trend in all four plots in Figure 1 as well as those in Figures 2 and 3. In Figure 2, we simulate sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of λ. We see that $Y(t)$ increases more quickly when $\lambda = 2$ comparing with other lower values, so does $X(t)$. In Figure 3, we simulate sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of μ_3. When μ_3 increases, both $X(t)$ and $Y(t)$ increase at a substantially faster rate. Overall, μ_3 has a much greater impact on the wealth $X(t)$ than h and λ.

In Figures 4–6, we simulate the optimal portfolio ratio π_t^* for different values of h, λ and μ_3, respectively. We observe that π_t^* is always greater than the Merton's portfolio π_t^{M*}, which coincides with Proposition 3.10. From Figure 4, we see that π_t^* increases as h increases. This is due to the expression (16) and that $Y(t)$ is significantly larger when h is larger (see Figure 1). On the other hand, π_t^* is not sensitive to changes in λ as seen in Figure 5. Lastly, we see from Figure 6 that π_t^* increases as μ_3 increases. Overall, h and μ_3 have larger effects on the optimal investment strategy π_t^* compared to λ.



(a) $\lambda = 0.01$ (b) $\lambda = 0.1$

(c) $\lambda = 0.2$ (d) $\lambda = 2$

FIGURE 2. Sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of λ under the power-type utility function (10)

3.4. **The optimization problem with exponential-type utility.** We now consider an exponential-type utility function defined below.

Definition 3.11 (Exponential-Type Utility Function)**.**
The utility function is given by

$$U(x,y) \equiv -\frac{e^{-\gamma\left(x+\mu_3 e^{\lambda h} y\right)}}{\gamma} = -\frac{1}{\gamma} e^{-\gamma x} e^{-\gamma \mu_3 e^{\lambda h} y}, \tag{39}$$

where $\gamma > 0$ is the risk aversion coefficient.

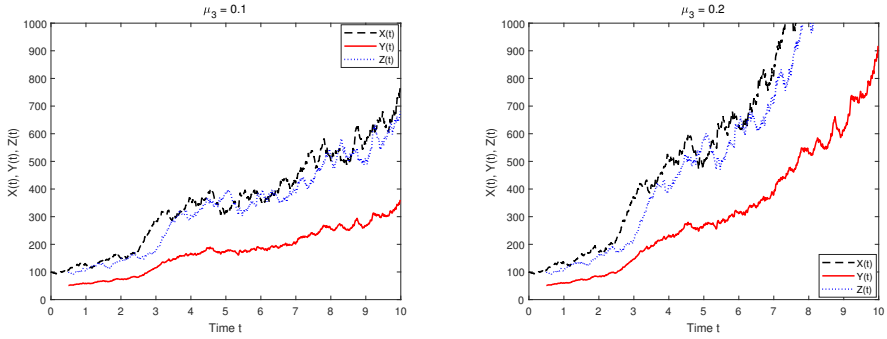
With the utility function specified to (39), the objective function (9) becomes

$$V(s,x,y) = \sup_{\pi \in \mathcal{A}} \mathbb{E}^{s,\varphi,\pi}\left[-\frac{1}{\gamma} e^{-\gamma X(T)} e^{-\gamma \mu_3 e^{\lambda h}} Y(T) \right]. \tag{40}$$

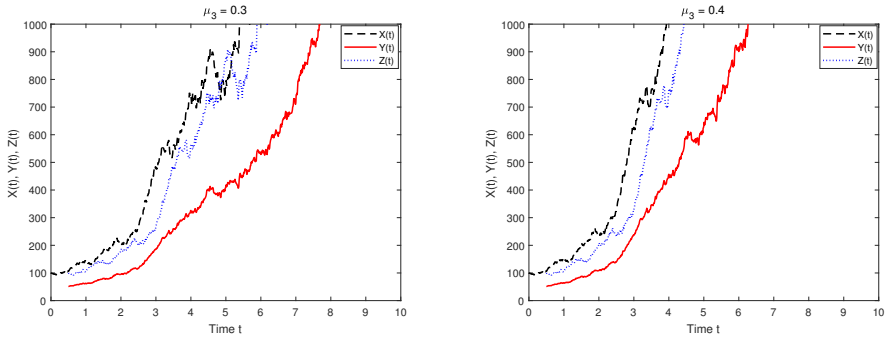
We next define the admissible set of investment strategies for the optimization problem (40) (cf. Definition 3.2).

Definition 3.12 (Admissible Set $\mathcal{A}$)**.**
An investment strategy $\pi \equiv \{\pi_t\}_{0 \le t \le T}$ is admissible if

(1) π_t is $\mathcal{F}_t$-measurable for all $t \in [0,T]$;



(a) $\mu_3 = 0.1$ (b) $\mu_3 = 0.2$



(c) $\mu_3 = 0.3$ (d) $\mu_3 = 0.4$

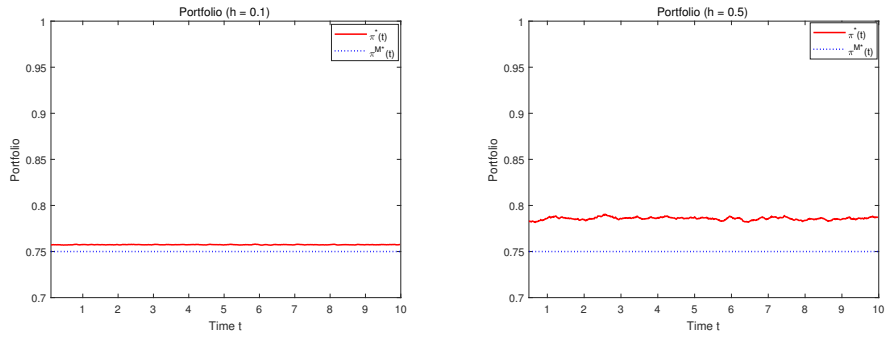
FIGURE 3. Sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of μ_3 under the power-type utility function (10)

(2) the following inequality holds:

$$\mathbb{E}\left[\int_0^T |\pi_t X(t)|^2 dt\right] < \infty.$$

Remark 3.13. We remark that the set of admissible control policies in Definition 3.2 is a subset of that in Definition 3.12. To see this, suppose (2) in Definition 3.2 holds. Then $|X(t)| \leq ||X_t||$ and (27) implies that

$$\begin{aligned}
\mathbb{E}\left[\int_0^T |\pi_t X(t)|^2 dt\right] &\leq \Lambda^2\, \mathbb{E}\left[\int_0^T \left|X(t) + \mu_3 e^{\lambda h} Y(t)\right|^2 dt\right] \\
&= \Lambda^2\, \mathbb{E}\left[\int_0^T \left(|X(t)|^2 + 2\mu_3 e^{\lambda h} |X(t)|\,|Y(t)| + \mu_3^2 e^{2\lambda h} |Y(t)|^2\right) dt\right] \\
&\leq \Lambda^2 \left(1 + h\mu_3 e^{\lambda h}\right)^2 \mathbb{E}\left[\int_0^T ||X_t||^2 dt\right] \\
&= \Lambda_2 \int_0^T \mathbb{E}||X_t||^2 dt
\end{aligned}$$



(a) $h = 0.1$ (b) $h = 0.5$

(c) $h = 1.0$ (d) $h = 2.0$

FIGURE 4. Optimal portfolio ratio π_t^* for different values of h under the power-type utility function (10)

$$\begin{aligned} &\leq \Lambda_2 \, T \, C_1 \left(1 + ||\varphi||^2\right) \\ &< \infty, \end{aligned}$$

where $\Lambda_2 \equiv \Lambda^2 \left(1 + h\mu_3 e^{\lambda h}\right)^2 > 0$ and $C_1 > 0$ are constants and the last step follows from Lemma 3.3.

We proceed to solve the HJB equation (12) with the exponential-type utility function (39).

Theorem 3.14. *The solution to the HJB equation* (12) *with the exponential-type terminal utility function*

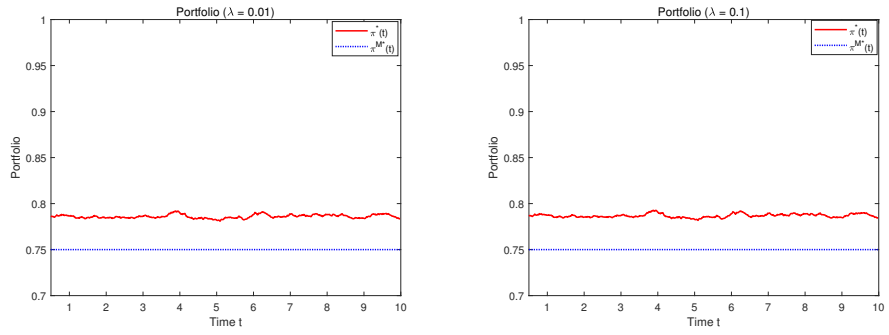
$$U(x, y) = -\frac{e^{-\gamma\left(x + \mu_3 e^{\lambda h} y\right)}}{\gamma}$$

is given by

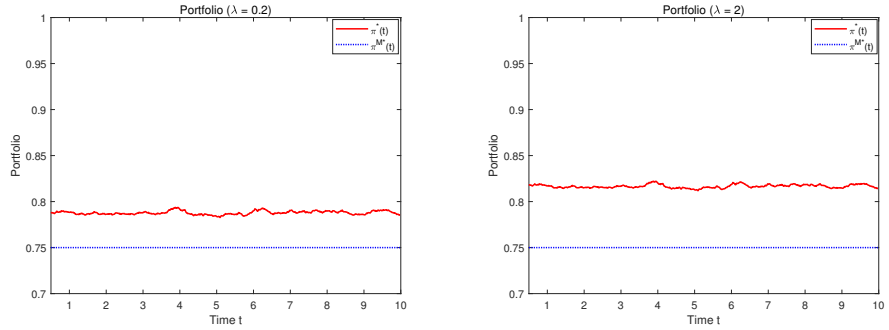
$$V(s, x, y) = -\frac{1}{\gamma} e^{-\frac{1}{2}\theta^2 (T-s)} \exp\left[-\gamma e^{\left(r + \mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)\right]. \tag{41}$$

Furthermore, the optimal portfolio ratio π_s^* at time s is given by

$$\pi_s^* = \frac{\theta}{\sigma\gamma X(s) e^{(r + \mu_3 e^{\lambda h})(T-s)}}. \tag{42}$$



(a) $\lambda = 0.01$ (b) $\lambda = 0.1$



(c) $\lambda = 0.2$ (d) $\lambda = 2$

FIGURE 5. Optimal portfolio ratio π_t^* for different values of λ under the power-type utility function (10)

Proof. First we consider the FOC with respect to π. From (12), we obtain

$$\pi^* = -\frac{(\mu_1 - r)V_x}{\sigma^2 x V_{xx}} = -\frac{\theta V_x}{\sigma x V_{xx}}. \tag{43}$$

Plugging (43) into (12), we have

$$0 = V_s + \{rx + \mu_2 y + \mu_3 z\} V_x + \left\{x - \lambda y - e^{-\lambda h} z\right\} V_y - \frac{1}{2}\theta^2 \frac{(V_x)^2}{V_{xx}}. \tag{44}$$

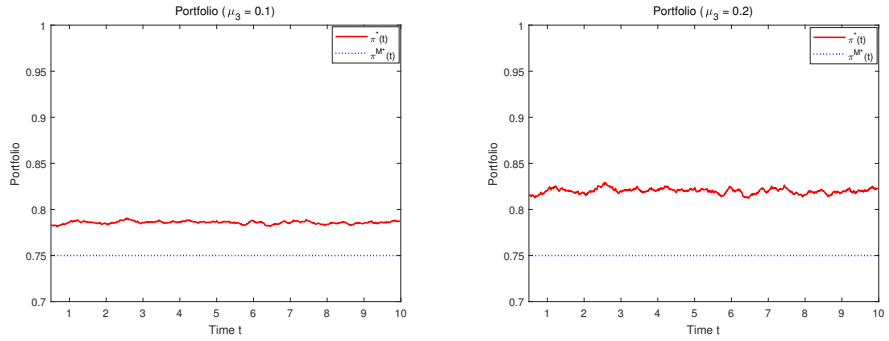
Conjecture the solution form as

$$V(s,x,y) = -\frac{1}{\gamma} A(s) \exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)\right]. \tag{45}$$

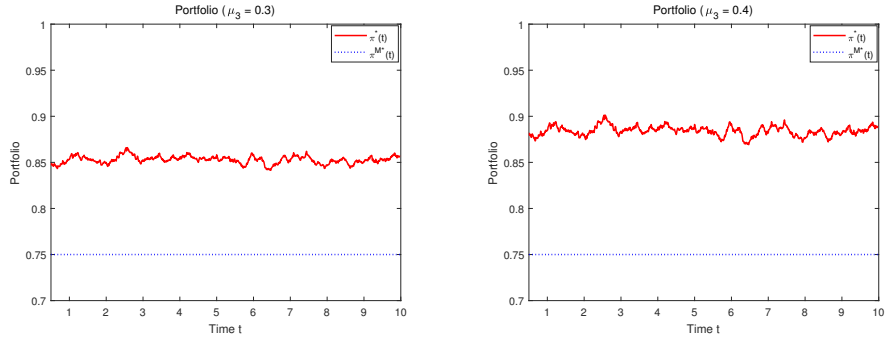
Then we have

$$V_s = -\frac{1}{\gamma}\left\{A'(s) + \gamma\left(r + \mu_3 e^{\lambda h}\right) e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right) A(s)\right\}$$
$$\times \exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)\right],$$

$$V_x = e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} A(s) \exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)\right],$$



(a) $\mu_3 = 0.1$ (b) $\mu_3 = 0.2$



(c) $\mu_3 = 0.3$ (d) $\mu_3 = 0.4$

FIGURE 6. Optimal portfolio ratio π_t^* for different values of μ_3 under the power-type utility function (10)

$$V_{xx} = -\gamma e^{2\left(r+\mu_3 e^{\lambda h}\right)(T-s)} A(s) \exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)\right],$$

$$V_y = e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \mu_3 e^{\lambda h} A(s) \exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)\right],$$

$$\frac{(V_x)^2}{V_{xx}} = -\frac{1}{\gamma} A(s) \exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right)\right].$$

Plugging the above equations into (44), we have

$$\begin{aligned} 0 = & -\frac{1}{\gamma}\left\{A'(s) + \gamma\left(r + \mu_3 e^{\lambda h}\right) e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \left(x + \mu_3 e^{\lambda h} y\right) A(s)\right\} \\ & + \left\{rx + \mu_2 y + \mu_3 z\right\} e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} A(s) \\ & + \left\{x - \lambda y - e^{-\lambda h} z\right\} e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} \mu_3 e^{\lambda h} A(s) + \frac{1}{2\gamma}\theta^2 A(s) \\ = & -\frac{1}{\gamma} A'(s) - \left(r + \mu_3 e^{\lambda h}\right)\left(x + \mu_3 e^{\lambda h} y\right) e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)} A(s) \end{aligned}$$

$$+\left\{\left(r+\mu_3 e^{\lambda h}\right)x+\left(\mu_2-\mu_3 e^{\lambda h}\lambda\right)y\right\}e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)}A(s)$$
$$+\frac{1}{2\gamma}\theta^2 A(s)$$
$$=-\frac{1}{\gamma}\left[A'(s)-\frac{1}{2}\theta^2 A(s)\right], \tag{46}$$

where the third equality is obtained from (8) in Assumption 2.4. Clearly, the solution to the ODE (46) is given by

$$A(s)=D_1 e^{\frac{1}{2}\theta^2 s},$$

for some constant D_1. From (13) and (45), we see that

$$V(T,x,y)=U(x,y)=-\frac{1}{\gamma}A(T)\exp\left[-\gamma\left(x+\mu_3 e^{\lambda h}y\right)\right]=A(T)U(x,y),$$

which implies $A(T)=1$. Consequently,

$$A(s)=e^{-\frac{1}{2}\theta^2(T-s)}.$$

Combining the above results,

$$V(s,x,y)=-\frac{1}{\gamma}e^{-\frac{1}{2}\theta^2(T-s)}\exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)}\left(x+\mu_3 e^{\lambda h}y\right)\right] \tag{47}$$

is the solution to HJB equation (12) with the exponential-type utility function (39) and the optimal portfolio ratio is

$$\pi^*=-\frac{\theta V_x}{\sigma x V_{xx}}=\frac{\theta}{\sigma\gamma x e^{(r+\mu_3 e^{\lambda h})(T-s)}}. \tag{48}$$

This completes the proof. □

To solve the optimization problem (40), we provide the following verification theorem, which can be proved in the same way as Theorem 3.7. Therefore, we omit the proof.

Theorem 3.15 (Verification Theorem)**.**
Let $X(t)$ be a strong solution to (6) with (7). Recall that $Y(t)$ and $Z(t)$ are defined by (1) and (2), respectively. Assume that $g(s,x,y)\in C^{1,2,1}([0,T]\times\mathbb{R}^2)$ is a solution to the HJB equation (12) with the terminal condition (13) and the exponential-type utility function (39), and that

$$\mathbb{E}\left[\int_0^T\left(\pi_t X(t)g_x(t,X(t),Y(t))\right)^2 dt\right]<\infty, \tag{49}$$

for any admissible control policy π. Then we have

$$g(s,x,y)\geq\sup_{\pi\in\mathcal{A}}\mathbb{E}^{s,\varphi,\pi}\left[-\frac{1}{\gamma}e^{-\gamma X(T)}e^{-\gamma\mu_3 e^{\lambda h}Y(T)}\right].$$

Let $\hat{\pi}=\{\hat{\pi}(t,X(t),Y(t))\}_{0\leq t\leq T}$, where

$$\hat{\pi}(t,x,y)=-\frac{\theta g_x(t,x,y)}{\sigma x g_{xx}(t,x,y)},\qquad (t,x,y)\in[0,T]\times\mathbb{R}^2. \tag{50}$$

If $\hat{\pi}$ is admissible, then $\hat{\pi}$ is the optimal control policy. Moreover, when $\hat{\pi}$ is taken as the control policy, we have

$$g(s,x,y) = \mathbb{E}^{s,\varphi,\hat{\pi}}\left[-\frac{1}{\gamma} e^{-\gamma X(T)} e^{-\gamma \mu_3 e^{\lambda h} Y(T)} \right].$$

Now, we are ready to solve the optimization problem (40).

Theorem 3.16. *Assume the utility function is given by (39) and the condition (8) holds. Then the function $V(s,x,y)$ given by (41) is a solution to the HJB equation (12) with (13) and is equal to the value function defined in (40). Moreover, the optimal control policy is given by (42).*

Proof. By Theorem 3.14, $V(s,x,y)$ given by (41) is a solution to the HJB equation (12) with (13) and the associated optimal portfolio ratio π^* is given by (42).

In what follows, we apply Theorem 3.15 to derive the solution of the optimization problem (40). We first verify that (49) holds for any admissible control policy π when the function $g(s,x,y) \equiv V(s,x,y)$. By Lemma 3.3, $X(t) > 0$ a.s. and $Y(t) > 0$ a.s. From (41),

$$\begin{aligned}
&|V_x(t, X(t), Y(t))| \\
&= e^{-\frac{1}{2}\theta^2 (T-t)} e^{\left(r+\mu_3 e^{\lambda h}\right)(T-t)} \exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-t)} \left(X(t) + \mu_3 e^{\lambda h} Y(t) \right) \right] \\
&\le L \exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-t)} \left(X(t) + \mu_3 e^{\lambda h} Y(t) \right) \right] \\
&\le L,
\end{aligned}$$

where

$$L \equiv \max\left\{ 1, e^{\left(r - \frac{1}{2}\theta^2 + \mu_3 e^{\lambda h}\right) T} \right\} > 0$$

is a constant that does not depend on t. By Definition 3.12, we have

$$\begin{aligned}
\mathbb{E}\left[\int_0^T (\pi_t X(t) V_x(t, X(t), Y(t)))^2 dt \right] &\le L^2\, \mathbb{E}\left[\int_0^T |\pi_t X(t)|^2 dt \right] \\
&< \infty.
\end{aligned}$$

Therefore, (49) holds for any admissible control policy π.

It is trivial that the control policy (42) satisfies (50) with $g(s,x,y) \equiv V(s,x,y)$. Next, we show that the control policy (42) is admissible. Clearly, π_t^* is $\mathcal{F}_t$-measurable for $t \in [0,T]$. Note that

$$\pi_t^* X(t) = \frac{\theta}{\sigma \gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-t)}}.$$

It is easy to see that the second condition in Definition 3.12 is also satisfied. Thus, π^* is admissible.

From the above results and Theorem 3.15, the function $V(s,x,y)$ given by (41) is equal to the value function defined in (40) and π^* is the optimal policy. □

3.5. **The value function and optimal portfolio ratio with exponential-type utility.** First we consider the original Merton problem (without delays). Our model becomes the original Merton model if we do not consider the effects of delays of the risky asset. More precisely, if we let $\mu_2 \equiv 0$ and $\mu_3 \equiv 0$ in the value function (40) and the wealth process (6), our model coincides with the original Merton model .

Remark 3.17 (Merton's Problem with CARA)**.**
The objective function of the investor is given by

$$V^M(s,x) = \sup_{\pi\in\mathcal{A}} \mathbb{E}^{s,X(s)=x}\left[-\frac{1}{\gamma}e^{-\gamma X(T)}\right]$$

subject to

$$dX(t) = \Big[rX(t) + (\mu_1 - r)\pi_t X(t)\Big]dt + \sigma\pi_t X(t)dB_t,\ X(s) = x,\ \forall\ t\in[s,T].$$

Then the associated HJB equation is derived as

$$\sup_{\pi}\left[V_s^M + \{rx + (\mu_1 - r)\pi x\}V_x^M + \frac{1}{2}\sigma^2\pi^2x^2V_{xx}^M\right] = 0, \tag{51}$$

with the terminal condition

$$V^M(T,x) = -\frac{1}{\gamma}e^{-\gamma x}. \tag{52}$$

If we consider the FOC with respect to π, then, from (51), we obtain

$$\pi^{M*} = -\frac{(\mu_1 - r)V_x^M}{\sigma^2 x V_{xx}^M} = -\frac{\theta V_x^M}{\sigma x V_{xx}^M}. \tag{53}$$

From (51)–(53), we obtain

$$V^M(s,x) = -\frac{1}{\gamma}e^{-\frac{1}{2}\theta^2(T-s)}e^{-\gamma e^{r(T-s)}x}. \tag{54}$$

Also using the FOC in (53) and the value function (54), we derive the optimal portfolio ratio at time s as

$$\pi_s^{M*} = \frac{\theta}{\sigma\gamma X(s)e^{r(T-s)}}. \tag{55}$$

We compare the value function (41) and optimal portfolio (42) with Merton's (54) and (55), respectively, in the next proposition.

Proposition 3.18. *For $0 \le s \le T$ and $x, y > 0$, we have*

$$\begin{aligned} V(s,x,y) &= -\frac{1}{\gamma}e^{-\frac{1}{2}\theta^2(T-s)}\exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)}\left(x + \mu_3 e^{\lambda h}y\right)\right] \\ &> -\frac{1}{\gamma}e^{-\frac{1}{2}\theta^2(T-s)}e^{-\gamma e^{r(T-s)}x} = V^M(s,x). \end{aligned}$$

Moreover,

$$\pi_s^*\cdot X(s) = \frac{\theta}{\sigma\gamma e^{(r+\mu_3 e^{\lambda h})(T-s)}} < \frac{\theta}{\sigma\gamma e^{r(T-s)}} = \pi_s^{M*}\cdot X(s).$$

Proof. Since $\mu_3 > 0$ is a positive constant, we see that

$$x + \mu_3 e^{\lambda h}y > x$$

and that

$$e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)}\left(x + \mu_3 e^{\lambda h}y\right) > e^{r(T-s)}x.$$

Therefore,

$$\exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)}\left(x+\mu_3 e^{\lambda h} y\right)\right] < e^{-\gamma e^{r(T-s)}x},$$

and

$$-\frac{1}{\gamma}e^{-\frac{1}{2}\theta^2(T-s)}\exp\left[-\gamma e^{\left(r+\mu_3 e^{\lambda h}\right)(T-s)}\left(x+\mu_3 e^{\lambda h} y\right)\right] > -\frac{1}{\gamma}e^{-\frac{1}{2}\theta^2(T-s)}e^{-\gamma e^{r(T-s)}x}.$$

From Lemma 3.3, (42) and (55), we see that

$$\begin{aligned}
\pi_s^* \cdot X(s) &= \frac{\theta}{\sigma\gamma e^{(r+\mu_3 e^{\lambda h})(T-s)}} \\
&= \frac{\theta}{\sigma\gamma}\cdot\frac{1}{e^{(r+\mu_3 e^{\lambda h})(T-s)}} \\
&< \frac{\theta}{\sigma\gamma e^{r(T-s)}} \\
&= \pi_s^{M*}\cdot X(s),
\end{aligned}$$

where the second last step holds because $e^{\mu_3 e^{\lambda h}(T-s)} > e^0 = 1$. This completes the proof. □

In Proposition 3.18, we compare the optimal amount of wealth invested in the risky asset for the investor with exponential-type utility and delays with that for the CARA investor without delays. From Proposition 3.18, we see that the investor with exponential-type utility and delays invests less in the risky asset than the CARA investor without delays. We also point out that, for the optimal portfolio ratios π_s^* and π_s^{M*} of two investors, we cannot say that one's portfolio ratio is greater than the other's.

3.6. **Numerical simulations with exponential-type utility.** In this subsection, we perform numerical simulations to illustrate our results in Subsections 3.4 and 3.5.

From (6) and (42), we obtain the optimal wealth process as follows: For $t \in [0,T]$,

$$\begin{aligned}
dX(t) &= \left[rX(t) + (\mu_1 - r)\pi_t^* X(t) + \mu_2 Y(t) + \mu_3 Z(t)\right]dt + \sigma\pi_t^* X(t)dB_t \\
&= \left[rX(t) + (\mu_1 - r)\frac{\theta}{\sigma\gamma X(t)e^{(r+\mu_3 e^{\lambda h})(T-t)}}X(t) + \mu_2 Y(t) + \mu_3 Z(t)\right]dt \\
&\quad + \sigma\frac{\theta}{\sigma\gamma X(t)e^{(r+\mu_3 e^{\lambda h})(T-t)}}X(t)dB_t \\
&= \left[rX(t) + \frac{\theta^2}{\gamma e^{(r+\mu_3 e^{\lambda h})(T-t)}} + \mu_2 Y(t) + \mu_3 Z(t)\right]dt \\
&\quad + \frac{\theta}{\gamma e^{(r+\mu_3 e^{\lambda h})(T-t)}}dB_t. \qquad (56)
\end{aligned}$$

To simulate sample paths of $X(t)$, $Y(t)$ and $Z(t)$, we divide the time period $[0,T]$ into N intervals with the same length $\Delta t = T/N$. Benchmark parameters are given as follows:

$$T = 10,\ N = 1000,\ X(0) = 1,\ r = 0.02,\ \mu_1 = 0.08,$$

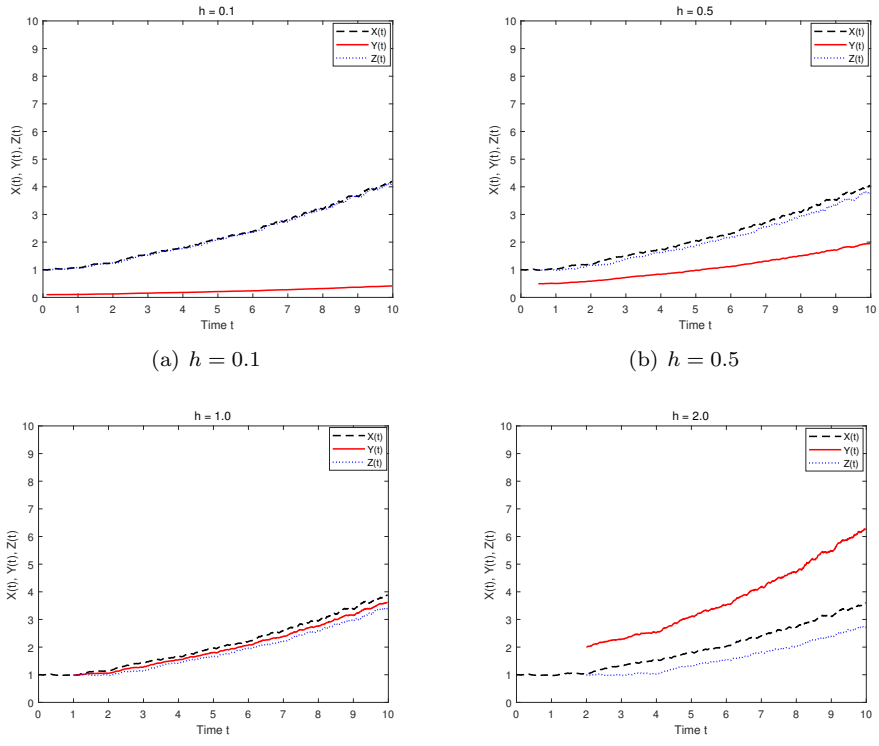
$$\sigma = 0.2,\ \gamma = 2,\ \lambda = 0.01,\ h = 0.5,\ \mu_3 = 0.1.$$

We assume that $Y(t) = Z(t) \equiv 0$ for $t \in [0, h]$ and use the Trapezoidal rule to compute $Y(t)$ in (1). The process (56) is approximated as

$$X_{t_i} = X_{t_{i-1}} + rX_{t_{i-1}}\Delta t + \frac{\theta^2}{\gamma\left(r+\mu_3 e^{\lambda h}\right)}\left[e^{-\left(r+\mu_3 e^{\lambda h}\right)(T-t_i)} - e^{-\left(r+\mu_3 e^{\lambda h}\right)(T-t_{i-1})}\right]$$

$$+ \mu_2 Y_{t_{i-1}}\Delta t + \mu_3 Z_{t_{i-1}}\Delta t + \frac{\theta}{\gamma e^{(r+\mu_3 e^{\lambda h})(T-t_{i-1})}}\sqrt{\Delta t}\,\epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0,1).$$

Again, for comparison purposes, we use the same set of simulated random variables $\epsilon_i, i = 1, 2, \ldots, N$, to generate the sample paths in all plots of each figure below.



(a) $h = 0.1$ (b) $h = 0.5$

(c) $h = 1.0$ (d) $h = 2.0$

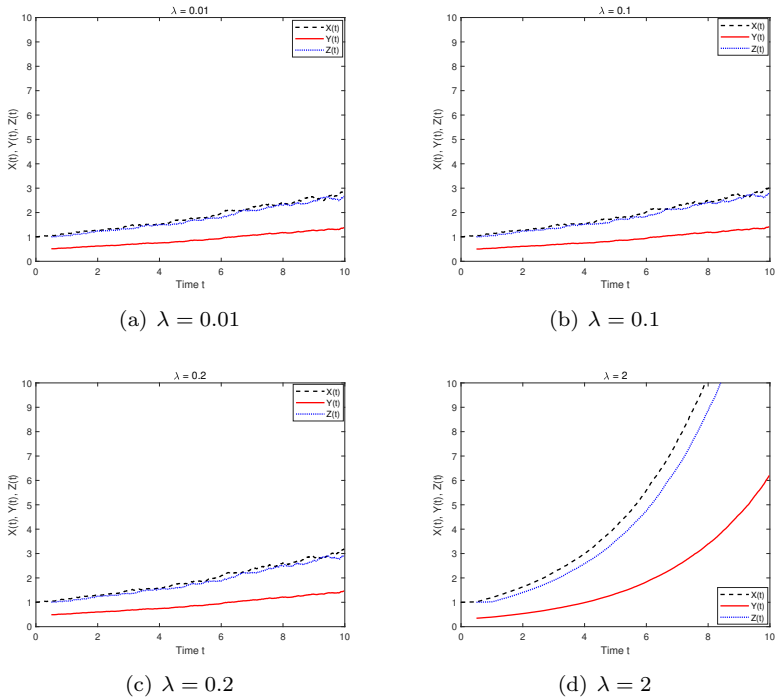
FIGURE 7. Sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of h under the exponential-type utility function (39)

In Figure 7, we simulate sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of h. As in Figure 1, we see that $Y(t)$ increases with h and $X(t)$ increases over time. In Figure 8, we simulate sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of λ. We see that $Y(t)$ exhibits a much faster rate of increase when $\lambda = 2$ comparing with other lower values, so does $X(t)$. In Figure 9, we simulate sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of μ_3. We see that the wealth $X(t)$

with larger μ_3 becomes relatively high as time passes. Moreover, the curves of all processes $X(t)$, $Y(t)$ and $Z(t)$ become smooth with larger μ_3. This is because the coefficient of dB_t in (56) becomes smaller as μ_3 increases. Overall, μ_3 has a greater impact on the wealth $X(t)$ than h and λ.

In Figures 10–12, we simulate the optimal portfolio ratio π_t^* for different values of h, λ and μ_3, respectively. We observe that π_t^* is less than the Merton's portfolio π_t^{M*} in these simulation results, which does not always holds. From Figures 10 and 11, we see that h and λ have relatively small effects on the portfolio π_t^*. On the other hand, we see from Figure 12 that π_t^* decreases significantly as μ_3 increases and becomes almost zero when $\mu_3 \geq 0.3$. This is because of the form of π_t^* in (42). Overall, we see that μ_3 has a greater impact on the portfolio π_t^* compared to h and λ. In Figure 13, we obtain the optimal portfolio for different initial endowment $X(0)$. In this subsection, we choose $X(0) = 1$ as the benchmark initial wealth. From Figure 13, we see that π_t^* decreases as $X(0)$ increases. This is because there are no wealth effects on investment in CARA utility and the investment amount in the risky asset is almost fixed. Therefore, as the initial wealth increases, the portfolio ratio decreases.

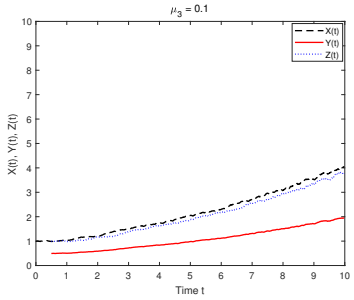
Remark 3.19. We see that μ_3 plays different roles in power-type and exponential type utilities. From (16) and Figure 6 with power-type utility, we see that π_t^* increases as μ_3 increases. However, from (42) and Figure 12 with exponential-type utility, we see that π_t^* decreases as μ_3 increases.



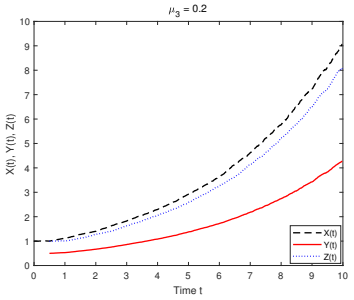
(a) $\lambda = 0.01$ (b) $\lambda = 0.1$

(c) $\lambda = 0.2$ (d) $\lambda = 2$

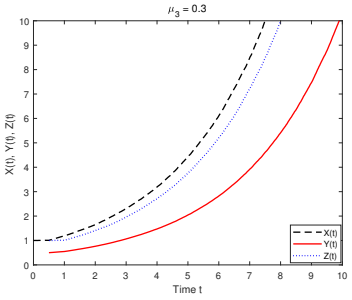
FIGURE 8. Sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of λ under the exponential-type utility function (39)



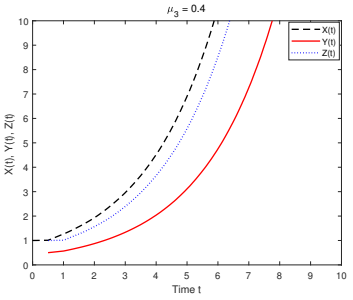
(a) $\mu_3 = 0.1$



(b) $\mu_3 = 0.2$

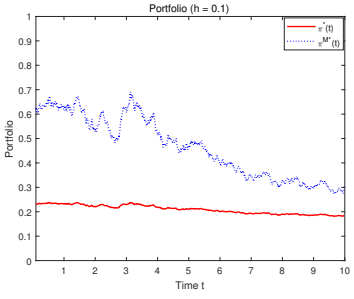


(c) $\mu_3 = 0.3$

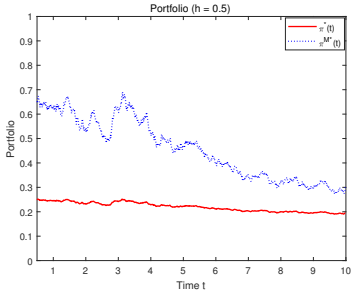


(d) $\mu_3 = 0.4$

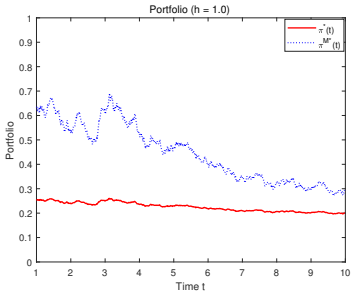
FIGURE 9. Sample paths of $X(t)$, $Y(t)$ and $Z(t)$ for different values of μ_3 under the exponential-type utility function (39)



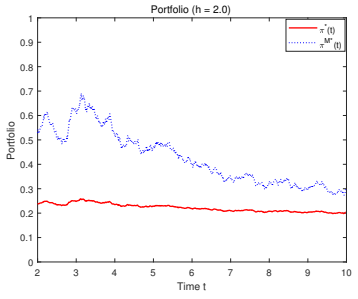
(a) $h = 0.1$



(b) $h = 0.5$

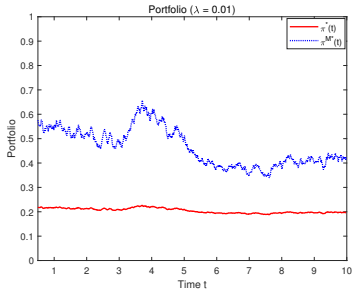


(c) $h = 1.0$

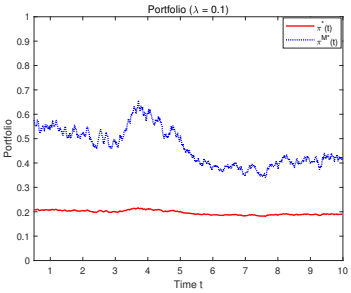


(d) $h = 2.0$

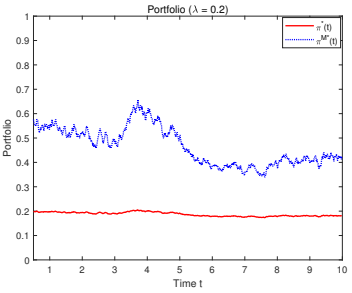
FIGURE 10. Optimal portfolio ratio π_t^* for different values of h under the exponential-type utility function (39)



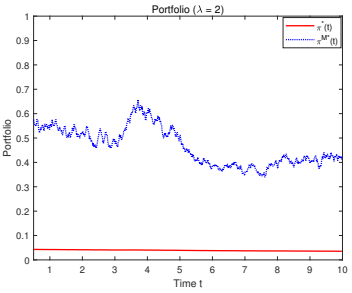
(a) $\lambda = 0.01$



(b) $\lambda = 0.1$

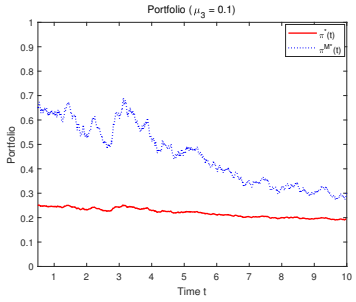


(c) $\lambda = 0.2$

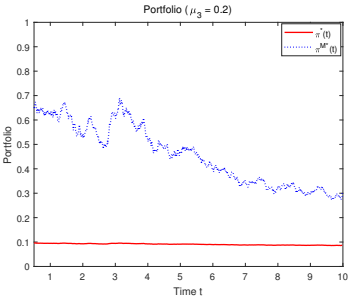


(d) $\lambda = 2$

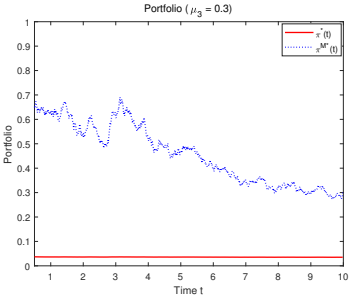
FIGURE 11. Optimal portfolio ratio π_t^* for different values of λ under the exponential-type utility function (39)



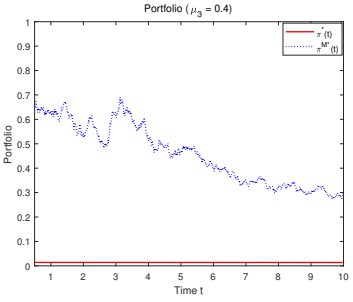
(a) $\mu_3 = 0.1$



(b) $\mu_3 = 0.2$

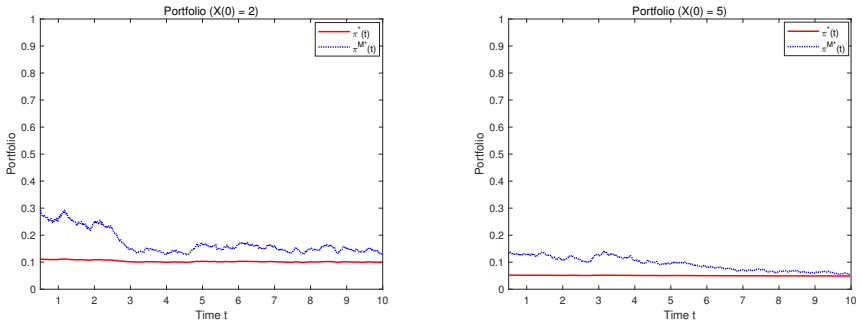


(c) $\mu_3 = 0.3$



(d) $\mu_3 = 0.4$

FIGURE 12. Optimal portfolio ratio π_t^* for different values of μ_3 under the exponential-type utility function (39)



(a) $X(0) = 2$ (b) $X(0) = 5$

FIGURE 13. Optimal portfolio ratio π_t^* for different initial wealth $X(0)$ under the exponential-type utility function (39)

4. **Concluding remarks.** We analyzed terminal utility maximization problems with delays, focusing on two types of utility functions: power-type and exponential-type. By assuming certain conditions, we derived the closed-form solution for each utility function and compared the resulting portfolios to those obtained under Merton's model without delays. Our findings showed that, in terms of the optimal portfolio ratio, an investor with power-type utility and delays tends to invest more in the risky asset compared to a CRRA investor without delays. On the other hand, in terms of the optimal portfolio amount, an investor with exponential-type utility and delays tends to invest less in the risky asset compared to a CARA investor without delays. Lastly, we provided some numerical simulation results to investigate the impacts of changes in various key parameters on the wealth process and the optimal investment strategy.

Acknowledgments. The authors would like to thank two anonymous referees for their useful comments.

REFERENCES

[1] C. A and Y. Shao, Portfolio optimization problem with delay under Cox-Ingersoll-Ross Model, *Journal of Mathematical Finance*, **7** (2017), 699-717.
[2] F. Biagini, Y. Hu, B. Oksendal and T. Zhang, *Stochastic Calculus for Fractional Brownian Motion and Applications*, Springer-Verlag London, Ltd., London, 2008.
[3] M.-H. Chang, T. Pang and Y. Yang, A stochastic portfolio optimization model with bounded memory, *Mathematics of Operations Research*, **36** (2011), 604-619.
[4] J. Gómez-González and N. Kiefer, Evidence of Non-Markovian behavior in the process of bank rating migrations, *Cuadernos de Economia*, **46** (2009), 33-50.
[5] K.-A. Lee, K. Lee and S.-H. Park, A financial market of a stochastic delay equation, *Bulletin of the Korean Mathematical Society*, **56** (2019), 1129-1141.
[6] L. Li and H. Mi, Optimal investment and consumption with stochastic factor and delay, *ANZIAM Journal*, **61** (2019), 99-117.
[7] R. C. Merton, Lifetime portfolio selection under uncertainty: The continuous-time case, *Review of Economics and Statistics*, **51** (1969), 247-257.
[8] R.C. Merton, Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory*, **3** (1971), 373-413.
[9] T. Pang and A. Hussain, An infinite time horizon portfolio optimization model with delays, *Mathematical Control and Related Fields*, **6** (2016), 629-651.

[10] T. Pang and A. Hussain, A stochastic portfolio optimization model with complete memory, *Stochastic Analysis and Applications*, **35** (2017), 742-766.
[11] X. Wang and L. Hu, A new method to solve the Hamilton-Jacobi-Bellman equation for a stochastic portfolio optimization model with boundary memory, *Journal of Industrial and Management Optimization*, **18** (2022), 3831-3845.

Received February 2023; 1st revision October 2023; 2nd revision January 2024; early access February 2024.

Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: www.tandfonline.com/journals/rquf20

An optimal retirement problem with job switching and unemployment risks under subsistence consumption constraints

Qi Li, Yong Hyun Shin & Ji-Hun Yoon

To cite this article: Qi Li, Yong Hyun Shin & Ji-Hun Yoon (2025) An optimal retirement problem with job switching and unemployment risks under subsistence consumption constraints, Quantitative Finance, 25:5, 795-815, DOI: 10.1080/14697688.2025.2498395

To link to this article: https://doi.org/10.1080/14697688.2025.2498395

Published online: 23 May 2025.

Submit your article to this journal

Article views: 25

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rquf20

Quantitative Finance, 2025
Vol. 25, No. 5, 795–815, https://doi.org/10.1080/14697688.2025.2498395

Routledge
Taylor & Francis Group
Check for updates

An optimal retirement problem with job switching and unemployment risks under subsistence consumption constraints

QI LI †, YONG HYUN SHIN ‡ and JI-HUN YOON *†§

†Department of Mathematics, Pusan National University, Busan 46241, Republic of Korea
‡Department of Mathematics & Research Institute of Natural Sciences, Sookmyung Women's University, Seoul 04310, Republic of Korea
§Institute of Mathematical Science, Pusan National University, Busan 46241, Republic of Korea

(*Received 17 September 2024; accepted 19 April 2025*)

This study presents for the first time a rigorous continuous-time optimal involuntary and voluntary retirement model that includes job-switching opportunities and subsistence consumption constraints. Two job states are allowed; J_1 and J_2. Job state J_1 provides a higher income but less leisure time than job state J_2. We examine the impacts of unemployment risk, job switching, and subsistence consumption constraints on consumption and investment decisions across three potential retirement statuses: voluntary retirement, involuntary retirement in job state J_1, and involuntary retirement in job state J_2. Using closed-form solutions derived from theoretical models by applying the Martingale approach and variational inequality method validated through comparative statics both analytically and numerically, we analyze how different unemployment possibilities, competitive insurance market conditions, and subsistence consumption levels influence individuals' economic decisions. Our findings indicate that subsistence consumption constraints and unemployment risk significantly reduce consumption and investment levels before voluntary retirement. Moreover, as the probability of unemployment increases, the wealth boundary for voluntary retirement decreases.

Keywords: Unemployment risk; Unemployment insurance; Involuntary retirement; Voluntary retirement; Job switching; Subsistence consumption constraints

JEL Classifications: E21, E24, G11,

1. Introduction

In retirement planning, the risk of forced unemployment poses a significant challenge that profoundly affects financial stability, consumption, and investment decisions. Traditional retirement models, such as extensions of the classical Merton's model (Merton 1969, 1971) exemplified by Choi *et al.* (2008), Farhi and Panageas (2007), Dybvig and Liu (2010, 2011), and Yang and Koo (2018), typically assume that individuals follow a predictable career trajectory, beginning their careers in the labor market and planning to retire voluntarily at a future date or mandatorily at a fixed retirement age. Although these models provide foundational frameworks, they fail to account for the uncertainty and disruption caused by exogenous unemployment shocks, which can force individuals to retire early and drastically alter their financial circumstances. This study considers the risk of unemployment in an optimal retirement problem with job-switching opportunities.

The intuition behind these findings is straightforward. The risk of involuntary retirement or forced unemployment is crucial for two reasons. First, it introduces a significant source of uncertainty in the retirement planning process. Unlike voluntary retirement, which can be planned and optimized, forced retirement often occurs unexpectedly, leading to the immediate cessation of labor income. This sudden loss of income can severely affect an individual's ability to maintain their standard of living, necessitating sufficient financial buffers and insurance mechanisms. Second, the risk of unemployment influences consumption and investment decisions, both before and after retirement. Faced with the possibility of involuntary unemployment, individuals are likely to adopt more conservative financial behaviors such as reducing consumption and high-risk investments, while increasing savings to mitigate potential income loss. This shift in behavior can have broader economic implications, affecting the overall demand and investment levels in the economy.

Some seminal works addressing unemployment risk are those of Salop (1979) and Shapiro and Stiglitz (1984) in

*Corresponding author. Email: ysci99@pusan.ac.kr

© 2025 Informa UK Limited, trading as Taylor & Francis Group

equilibrium. Specifically, Salop (1979) revisited the microfoundations of the natural unemployment rate by constructing a model that incorporates labor mobility and imperfect information, showing that market frictions and structural deficiencies in the labor market can lead to a natural unemployment rate and unavoidable structural unemployment. Shapiro and Stiglitz (1984) demonstrated how unemployment can serve as a disciplinary mechanism for workers, explaining that when employers can not monitor employee effort without cost, unemployment becomes a necessary equilibrium condition. Incorporating the risk of forced unemployment into retirement models builds on the foundational work of Merton (1969, 1971) and other classical retirement economic models (Farhi and Panageas 2007, Choi *et al.* 2008, Dybvig and Liu 2010, 2011, Yang and Koo 2018). Jang *et al.* (2013) and Bensoussan *et al.* (2016) considered the risk of forced unemployment in an optimal retirement model. Jang *et al.* (2013) examined optimal retirement strategies in the context of involuntary unemployment risks and borrowing limitations in a complete market, revealing that heightened unemployment risks and market frictions considerably influence consumption patterns and stockholdings approaching retirement. Bensoussan *et al.* (2016) investigated optimal retirement strategies for individuals who encounter unavoidable income risks from forced unemployment, showing that these risks result in reduced consumption, increased savings, and notable effects on investment decisions, particularly in incomplete market environments.

We adopt the framework of forced unemployment risks from Jang *et al.* (2013) and Bensoussan *et al.* (2016) and consider individuals purchasing unemployment insurance to hedge against the risk of unemployment, thereby completing the market. Unemployment insurance is an essential social safety net that offers temporary financial assistance to unemployed workers. It aims to ease the economic strain of job losses and promote stability for both individuals and the broader economy. Hopenhayn and Nicolini (1997) constructed an optimal unemployment insurance contract where benefits decrease over time and reemployment wages are taxed progressively, demonstrating that this approach significantly reduce costs and enhance consumption smoothing compared to traditional systems. Acemoglu and Shimer (2000) explored how unemployment insurance can enhance labor productivity by motivating workers to pursue more productive jobs and incentivizing firms to create these positions, ultimately leading to higher overall output and improved welfare, although it may result in increased unemployment rates. Jang *et al.* (2013) studied the impact of unemployment insurance in an optimal consumption-investment and retirement problem.

Contrary to Jang *et al.* (2013) and Bensoussan *et al.* (2016), we also introduce an opportunity for job switching, allowing individuals to switch freely between two job states; J_1 and J_2, before reaching the optimal retirement time. Consequently, our model encompasses three retirement scenarios: *Scenario One*, where individuals experience involuntary retirement owing to forced unemployment while in job state J_1; *Scenario Two*, where individuals experience involuntary retirement after switching to job state J_2; and *Scenario Three*, where individuals opt for voluntary retirement upon reaching the expected retirement age. The presence of job-switching opportunities (Shim and Shin 2014, Li *et al.* 2024, 2025) enhances labor market flexibility, allowing individuals to adapt more effectively to changing economic conditions. Specifically, individuals can switch their job states and choose between different career opportunities to achieve a higher income or more leisure time. This flexibility improves job satisfaction and significantly mitigates the adverse effects of unemployment risks, enabling adjustments in consumption and investment strategies. Notably, we focus on obtaining a closed-form solution for this model. Therefore, we assume that once individuals retire, whether voluntarily or involuntarily, they cannot reenter the labor market. This implies that post-retirement labor income is zero, and individuals should rely on unemployment insurance compensation and investment returns to maintain their basic standard of living.

We employ the Martingale method (Cox and Huang 1989, Choi *et al.* 2008, Shim and Shin 2014, Li *et al.* 2024, 2025), and variational inequalities (Karatzas and Wang 2000, Li *et al.* 2025) to derive a closed-form solution for the model. One of our innovations lies in providing comprehensive mathematical proofs for the derivation process and conducting a thorough theoretical analysis of the closed-form solution, analytically deriving its properties and corresponding propositions. Additionally, we verify the accuracy of the solution through numerical analysis and comparative statics and visually summarize the model's economic implications and policy insights using graphical representations. We also consider subsistence consumption constraints (SCCs) in our model. SCCs (Gong and Li 2006, Lakner and Ma Nygren 2006, Mai *et al.* 2024) define the minimum consumption level required to maintain a basic standard of living. Incorporating SCCs into optimal consumption-investment and retirement models ensures that individuals' basic needs are met, thus enhancing the realism and robustness of the model. This approach also provides accurate insights into public policy recommendations.

The main contributions of this study are as follows. First, this study pioneers the integration of unemployment risk into an optimal investment and consumption model that incorporates both voluntary retirement and job-switching opportunities, reflecting dual labor flexibility and introducing SCCs. Unlike Jang *et al.* (2013) and Bensoussan *et al.* (2016), our model allows individuals to flexibly choose their retirement age, while enabling them to freely select their labor income and leisure time before retirement, demonstrating that these variables are dynamic. By incorporating job-switching opportunities, our model better reflects the real-world conditions in which individuals may face unemployment in various job contexts, leading to significant differences in their pre-retirement consumption and investment decisions, depending on the job state at the time of forced retirement. Among the three retirement scenarios, individuals who retire involuntarily before job switching (Scenario 1) exhibit the highest consumption and lowest investment, whereas those who retire voluntarily (Scenario 3) have the lowest consumption and highest investment before retirement. This phenomenon is attributable to a reduction in consumption and an increase in investment when transitioning from high- to low-income states. These dynamics are illustrated in figures 6 and 7, along with the corresponding analytical proofs.

Second, comparing the consumption and investment strategies between the optimal retirement models with and without unemployment risks, figure 5 shows that unemployment risks lead to delayed job switching, earlier voluntary retirement, and a reduction in pre-retirement consumption and risk investments. Furthermore, as the probability of unemployment increases, figures 2(a) and 3(a) show that the wealth level boundary for voluntary retirement decreases.

Third, this study incorporates unemployment insurance as a hedge against unemployment risk. The availability of unemployment insurance alleviates financial pressure during forced unemployment, enabling individuals to more effectively adjust their consumption and investment strategies. Consequently, we assume a complete market. As demonstrated in figures 6 and 7, lower market competitiveness–reflected by a higher loading factor–leads to a smaller voluntary retirement wealth boundary and a larger job-switching wealth boundary. Moreover, in a low-competition insurance market, individuals' consumption and investment levels before retirement across all three retirement models are lower than those in a highly competitive market. The purchase of unemployment insurance effectively mitigates the financial losses associated with involuntary retirement at low wealth levels, significantly reducing financial strain.

Finally, we observe that the presence of SCCs requires individuals to allocate a minimum portion of their resources to meet basic living standards, which in turn results in lower consumption and investment levels in the period leading to voluntary retirement. Furthermore, as subsistence consumption level increases, individuals have less disposable wealth for discretionary spending and investment, leading to a corresponding decrease in the wealth boundary necessary for voluntary retirement. This dynamic highlights the significant impact that subsistence needs have on financial decision-making and retirement planning.

The remainder of the study is structured as follows. Section 2 presents a detailed explanation of the optimal consumption and investment problem by integrating decisions related to optimal retirement and job switching under unemployment risk and SCCs within a continuous-time framework. In Section 3, we derive closed-form solutions for the optimal consumption and investment strategies across two involuntary retirement scenarios and one voluntary retirement scenario by applying the Martingale approach and variational inequality. Section 4 explores the effects of varying the parameter values linked to unemployment risk, loading factors, and SCCs on the optimal strategies. Through a comparative static analysis, we numerically demonstrated the variations in optimal consumption, investment, job-switching wealth boundaries, and voluntary retirement wealth boundaries. Finally, Section 5 concludes the study. Comprehensive proofs and technical derivations are provided in the Appendix.

2. Model description and economic background

2.1. Economy

We develop a continuous-time economic model, assuming that individuals have infinite lifespans. Under the no transaction cost hypothesis, the model considers two underlying assets. The first asset is a risk-free money market account denoted by S_t^0, which appreciates a risk-free interest rate of $r > 0$. The second asset is a high-risk asset akin to a stock, denoted by S_t, with price evolution characterized by an expected return $\mu > 0$ and volatility $\sigma > 0$. This asset follows geometric Brownian motion, where B_t is a standard Brownian motion defined in probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$. The dynamics of these two assets can be described as follows.

$$\frac{\mathrm{d}S_t^0}{S_t^0} = r\mathrm{d}t,$$
$$\frac{\mathrm{d}S_t}{S_t} = \mu\mathrm{d}t + \sigma\mathrm{d}B_t.$$

We also consider a framework in which individuals can freely switch between two distinct job states; J_1 and J_2, before voluntary retirement. This reversible occupational choice allows individuals to dynamically respond to changing preferences and economic conditions. Each job state, J_i, is characterized by specific labor income, I_i, and a corresponding amount of leisure time, L_i, where $i \in \{1, 2\}$. Conditions $I_1 > I_2 > 0$ and $0 < L_1 < L_2 < 1$ hold, implying that job state J_1 provides a higher income but less leisure time than job state J_2. In addition, after retirement, individuals enjoy full leisure, valued at one, and without income.

Contrary to Farhi and Panageas (2007), Dybvig and Liu (2010), and Li *et al.* (2025), who assumed that individuals start their careers in the labor market and plan to retire voluntarily, this study incorporates the risk of involuntary unemployment. Specifically, we consider a scenario in which an individual may be forced to retire if an exogenous unemployment shock occurs before optimal voluntary retirement. The occurrence of a forced unemployment event is modeled as an exponential distribution with intensity parameter δ. This means that a shock can occur at any time $t \geq 0$, adding an element of uncertainty to an individual's career trajectory and retirement planning. The probability of this event is

$$P\{\tau \leqslant t\} = 1 - e^{-\delta t},$$

where τ is the time required for forced unemployment. It is assumed that the random variable τ is independent of the filtration $\{\mathcal{F}_t\}_{t\geqslant 0}$, which is generated by the Brownian motion B_t.

We also propose tools to hedge against the risk of involuntary unemployment, the most effective being unemployment insurance. Insurance is crucial for managing the financial risks associated with job loss. Individuals facing high involuntary unemployment risk may seek comprehensive insurance coverage to prevent income loss and ensure that they maintain a reasonable standard of living after retirement. We adopt the unemployment insurance framework of Jang *et al.* (2013) and incorporate the environmental factors of the insurance market, including the consideration of a loading factor identified as ϵ. In insurance pricing, the loading factor represents additional costs beyond the basic premium used by insurance companies to cover administrative, marketing, claims, and profit expenses. Specifically, an individual pays $(\delta(\bar{X}_t - X_t) + \epsilon)$ per time unit. Here, X_t is the wealth-level process.

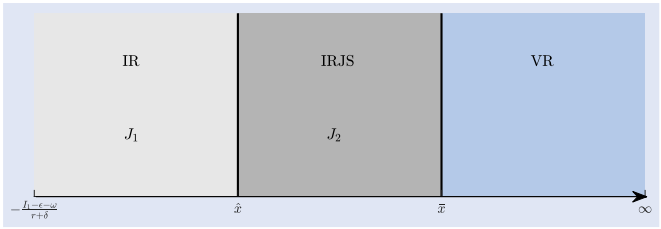


Figure 1. Structure for understanding when involuntary or voluntary retirement occurs in the wealth level domain.

$\bar{X}_t$ represents the potential retirement wealth. In the event of forced unemployment, individuals can receive a lump-sum payment of $(\bar{X}_t - X_t)$, thereby instantaneously elevating their wealth level to potential retirement wealth $\bar{X}_t$. This mechanism ensures that after experiencing forced unemployment, individuals can maintain a reasonable level of wealth, aiding in the stabilization of their consumption and quality of life. After involuntary retirement, individuals have the opportunity to optimize their wealth accumulation through prudent investment and consumption strategies. If no forced unemployment event occurs before the optimal retirement time τ_R, the individual will achieve the expected wealth level upon retirement voluntarily; that is, $\bar{X}_t = X_t$ for $t \geqslant \tau_R$. In this scenario, unemployment insurance serves a protective function without actual payouts, demonstrating its preventive role.

Consequently, the wealth level process of individuals can be characterized as follows:

$$\begin{aligned} \mathrm{d}X_t = \big[& rX_t + \pi_t(\mu - r) - c_t + \big\{ I_1 \mathbf{1}_{\{\Xi_t = J_1\}} + I_2 \mathbf{1}_{\{\Xi_t = J_2\}} \\ & - \big(\delta(\bar{X}_t - X_t) + \epsilon\big)\big\} \mathbf{1}_{\{0 \leqslant t < \tau \wedge \tau_R\}} \big] \mathrm{d}t + \pi_t \sigma \mathrm{d}B_t, \end{aligned} \tag{1}$$

with the initial endowment $X_0 = x$. $\mathcal{F}_t$-measurable processes for consumption c_t and investment π_t should adhere to the conditions $\int_0^\infty c_t \mathrm{d}t < \infty$ and $\int_0^\infty \pi_t^2 \mathrm{d}t < \infty$. To ensure the robustness of the model, accurately reflect real-life conditions, and safeguard basic living needs, we incorporate SCCs into the model.

$$c_t \geqslant \omega, \quad \text{for } 0 \leqslant t < \tau \wedge \tau_R. \tag{2}$$

By leveraging Lemma 3.1 from Gong and Li (2006), we determine the initial positive endowment wealth value as $x = -\frac{I_1 - \epsilon - \omega}{r+\delta}$ for $0 \leqslant t < \tau \wedge \tau_R$.

Figure 1 illustrates the occurrence of involuntary and voluntary retirement across different wealth levels. Here, the job-switching wealth boundary $\hat{x}$ and voluntary retirement wealth boundary $\bar{x}$ are determined in Section 3. Depending on the wealth level, involuntary and voluntary retirement may occur in the following three distinct regions:

(1) Involuntary Retirement Region (IR): This region corresponds to the wealth interval between $-\frac{I_1 - \epsilon - \omega}{r+\delta}$ and $\hat{x}$, where the individual is in the employment state J_1 before transitioning to another job. Within this range, individuals may be forced to retire involuntarily owing to external factors, such as layoffs and poor health conditions.
(2) Involuntary Retirement Post-Job-Switching Region (IRJS): After transitioning to employment state J_2 but before reaching the threshold for voluntary retirement, this region spans from $\hat{x}$ to $\bar{x}$. In this interval, the individual remains susceptible to involuntary retirement despite switching jobs.
(3) Voluntary Retirement Region (VR): This region is defined by wealth levels exceeding $\bar{x}$. Individuals in this wealth bracket have sufficient financial resources to retire voluntarily without further employment.

These regions collectively illustrate the potential retirement status that individuals with different wealth levels may face, highlighting the interplay between financial preparedness and retirement decisions. Understanding these dynamics provides valuable insights for individuals planning retirement and policymakers designing retirement-related policies.

2.2. *Stochastic control problem*

In this model, individuals gain utility from both leisure activities and consumption before retirement voluntarily or involuntarily. To capture these preferences, we employ a Cobb-Douglas utility function described as follows:

$$u_i(c_t, L_i) \equiv \frac{1}{\alpha} \frac{\left(c_t^\alpha L_i^{1-\alpha}\right)^{1-\gamma^*}}{1-\gamma^*}, \quad \text{with } i \in \{1, 2\}, \tag{3}$$

where γ^* denotes the coefficient of relative risk aversion and α is the elasticity of preference for consumption. The post-retirement utility function is $u_R(c_t) = \frac{1}{\alpha} \frac{c_t^{\alpha(1-\gamma^*)}}{1-\gamma^*}$. To simplify calculations, we define $\gamma = 1 - \alpha(1-\gamma^*) > 0$. Thus, the utility function in equation (3) can be reexpressed as

$$u_i(c_t, L_i) = L_i^{\gamma - \gamma^*} \frac{c_t^{1-\gamma}}{1-\gamma}.$$

We rewrite the post-retirement utility function as $u_R(c_t) = \frac{c_t^{1-\gamma}}{1-\gamma}$. The optimal consumption-leisure-portfolio problem, which includes the risk of unemployment and SCCs along with controls for job state, consumption, investment, and voluntary retirement timing, can be formulated as follows:

PROBLEM 1 *An individual seeks to maximize the expected utility derived from both consumption and leisure, as follows:*

$$\begin{aligned} V(x) \equiv & \sup_{(c,\pi,\Xi,\tau_R) \in \mathcal{A}(x)} \mathbb{E}\Bigg[\int_0^\infty e^{-\beta t} \Bigg(\Bigg(L_1^{\gamma-\gamma^*} \frac{c_t^{1-\gamma}}{1-\gamma} \mathbf{1}_{\{\Xi_t = J_1\}} \\ & + L_2^{\gamma-\gamma^*} \frac{c_t^{1-\gamma}}{1-\gamma} \mathbf{1}_{\{\Xi_t = J_2\}} \Bigg) \mathbf{1}_{\{0 \leqslant t < \tau \wedge \tau_R\}} \\ & + \frac{c_t^{1-\gamma}}{1-\gamma} \mathbf{1}_{\{t \geqslant \tau \wedge \tau_R\}} \Bigg) \mathrm{d}t \Bigg] \\ = & \sup_{(c,\pi,\Xi,\tau_R) \in \mathcal{A}(x)} \mathbb{E}\Bigg[\int_0^{\tau \wedge \tau_R} e^{-\beta t} \Bigg(L_1^{\gamma-\gamma^*} \frac{c_t^{1-\gamma}}{1-\gamma} \mathbf{1}_{\{\Xi_t = J_1\}} \\ & + L_2^{\gamma-\gamma^*} \frac{c_t^{1-\gamma}}{1-\gamma} \mathbf{1}_{\{\Xi_t = J_2\}} \Bigg) \mathrm{d}t \\ & + e^{-\beta(\tau \wedge \tau_R)} \bar{V}\left(X_{\tau \wedge \tau_R}\right) \Bigg]. \end{aligned} \tag{4}$$

Here, $\beta > 0$ is a subjective discount factor and

$$\bar{V}(X_{\tau\wedge\tau_R}) \equiv \sup_{(c,\pi)\in\mathcal{A}_1(x)} \mathbb{E}\left[\int_{\tau\wedge\tau_R}^{\infty} e^{-\beta(t-\tau\wedge\tau_R)}\frac{c_t^{1-\gamma}}{1-\gamma}\mathrm{d}t\right].$$

$\mathcal{A}(x)$ denotes the set of all admissible policies (c,π,Ξ,τ_R) associated with x and is defined as

$$\mathcal{A}(x) \equiv \Big\{(c,\pi,\Xi,\tau_R)\,|\,X_t > -\frac{I_1-\epsilon-\omega}{r+\delta} \quad \textit{for all}$$
$$0\leqslant t<\tau_R, X_t\geqslant 0 \textit{ for all}$$
$$\tau_R\leqslant t, \textit{ and } \tau_R\in\mathcal{T}\}, \tag{5}$$

where $\mathcal{T}$ represents the set of all $\mathcal{F}$-stopping times that can have values from 0 to ∞. $\mathcal{A}_1(x)$ is the set of all admissible strategies (c,π) and is defined as

$$\mathcal{A}_1(x) \equiv \{(c,\pi)\,|\,X_t\geqslant 0\}.$$

We define $s=\tau\wedge\tau_R$. Then, $\bar{V}$ can be rewritten as

$$\bar{V}(X_s) = \sup_{(c,\pi)\in\mathcal{A}_1(x)} E\left[\int_s^{\infty} e^{-\beta(t-s)}\frac{c_t^{1-\gamma}}{1-\gamma}\mathrm{d}t\right].$$

According to Merton (1969), the solution can be written directly as

$$\bar{V}(X_t) = \frac{X_t^{1-\gamma}}{(1-\gamma)K^{\gamma}}, \tag{6}$$

where $K\equiv r+\frac{\beta-r}{\gamma}+\frac{\gamma-1}{2\gamma^2}\theta^2$ and $\theta=\frac{\mu-r}{\sigma}>0$ is the market price of risk.

ASSUMPTION 1 *It is presumed that the following inequalities are consistently upheld.*

$$\beta\geqslant r, \quad K>0.$$

By substituting the intensity of unemployment risk into (4), we can rewrite (4) as follows:

$$\begin{aligned}
V(x) &= \sup_{(c,\pi,\Xi,\tau_R)\in\mathcal{A}(x)} \mathbb{E}\Big[\int_0^{\tau\wedge\tau_R} e^{-\beta t}U(c_t)\,\mathrm{d}t \\
&\quad + e^{-\beta(\tau\wedge\tau_R)}\bar{V}\left(X_{\tau\wedge\tau_R}\right)\Big] \\
&= \sup_{(c,\pi,\Xi,\tau_R)\in\mathcal{A}(x)} \mathbb{E}\Big[\int_0^{\infty}\delta e^{-\delta s}\int_0^{\tau_R\wedge s} e^{-\beta t}U(c_t)\,\mathrm{d}t\mathrm{d}s \\
&\quad + \int_0^{\infty}\delta e^{-\delta s}e^{-\beta(\tau_R\wedge s)}\bar{V}\left(X_{\tau_R\wedge s}\right)\mathrm{d}s\Big] \\
&= \sup_{(c,\pi,\Xi,\tau_R)\in\mathcal{A}(x)} \mathbb{E}\Big[\int_0^{\tau_R}\delta e^{-\delta s}\int_0^{s} e^{-\beta t}U(c_t)\,\mathrm{d}t\mathrm{d}s \\
&\quad + \int_{\tau_R}^{\infty}\delta e^{-\delta s}\int_0^{s} e^{-\beta t}U(c_t)\,\mathrm{d}t\mathrm{d}s \\
&\quad + \int_0^{\tau_R}\delta e^{-\delta s}e^{-\beta s}\bar{V}\left(\bar{X}_s\right)\mathrm{d}s \\
&\quad + \int_{\tau_R}^{\infty}\delta e^{-\delta s}e^{-\beta\tau_R}\bar{V}\left(X_{\tau_R}\right)\mathrm{d}s\Big] \\
&= \sup_{(c,\pi,\Xi,\tau_R)\in\mathcal{A}(x)} \mathbb{E}\Big[\int_0^{\tau_R} e^{-\beta t}U(c_t)\int_t^{\tau_R}\delta e^{-\delta s}\mathrm{d}s\mathrm{d}t \\
&\quad + \int_0^{\tau_R} e^{-\beta t}U(c_t)\int_{\tau_R}^{\infty}\delta e^{-\delta s}\mathrm{d}s\mathrm{d}t \\
&\quad + \int_0^{\tau_R} e^{-(\beta+\delta)s}\delta\bar{V}\left(\bar{X}_s\right)\mathrm{d}s + e^{-(\beta+\delta)\tau_R}\bar{V}\left(X_{\tau_R}\right)\Big] \\
&= \sup_{(c,\pi,\Xi,\tau_R)\in\mathcal{A}(x)} \mathbb{E}\Big[\int_0^{\tau_R} e^{-\beta t}U(c_t)\int_t^{\infty}\delta e^{-\delta s}\mathrm{d}s\mathrm{d}t \\
&\quad + \int_0^{\tau_R} e^{-(\beta+\delta)t}\delta\bar{V}\left(\bar{X}_t\right)\mathrm{d}t \\
&\quad + e^{-(\beta+\delta)\tau_R}\bar{V}\left(X_{\tau_R}\right)\Big] \\
&= \sup_{(c,\pi,\Xi,\tau_R)\in\mathcal{A}(x)} \mathbb{E}\Big[\int_0^{\tau_R} e^{-(\beta+\delta)t}\left\{U(c_t)+\delta\bar{V}\left(\bar{X}_t\right)\right\} \\
&\quad \mathrm{d}t + e^{-(\beta+\delta)\tau_R}\bar{V}\left(X_{\tau_R}\right)\Big],
\end{aligned} \tag{7}$$

where

$$U(c_t) \equiv L_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_1\}} + L_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_2\}}.$$

Given the discounted process $\zeta_t\equiv\exp\{-rt\}$ and the exponential Martingale process $Z_t\equiv\exp\{-\frac{1}{2}\theta^2 t-\theta B_t\}$, we can define the risk-neutral measure $\mathbb{Q}(A)\equiv\tilde{\mathbb{E}}[Z_T\mathbf{1}_A]$ for any $A\in\mathcal{F}_T$. By directly applying Girsanov's theorem, we transform the wealth process (1) (dynamic budget constraint) into a static budget constraint as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\int_0^{\tau\wedge\tau_R} H_t\Big\{c_t-\big(I_1\mathbf{1}_{\{\Xi_t=J_1\}}+I_2\mathbf{1}_{\{\Xi_t=J_2\}} \\
&\quad -\big(\delta(\bar{X}_t-X_t)+\epsilon\big)\big)\Big\}\,\mathrm{d}t + H_{\tau\wedge\tau_R}X_{\tau\wedge\tau_R}\Big] \leqslant x,
\end{aligned} \tag{8}$$

where H_t denotes the state price density (or stochastic discount factor), which is defined as follows:

$$H_t \equiv \zeta_t Z_t.$$

In the same manner of (7), the static budget constraint (8) can be characterized as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\int_0^{\tau_R} e^{-\delta t}H_t\left\{c_t-\left(I_1\mathbf{1}_{\{\Xi_t=J_1\}}+I_2\mathbf{1}_{\{\Xi_t=J_2\}}-\delta\bar{X}_t-\epsilon\right)\right\}\mathrm{d}t \\
&\quad + e^{-\delta\tau_R}H_{\tau_R}X_{\tau_R}\Big] \leqslant x.
\end{aligned} \tag{9}$$

3. Optimal solutions

The Martingale method, as outlined in Choi *et al.* (2008), Shim and Shin (2014), Li *et al.* (2024, 2025), and Cox and Huang (1989), is used to transform Problem 1 into its corresponding dual problem. This transformation is crucial for deriving an exact closed-form solution for both the value function and optimal strategies.

Subsequently, a Lagrangian framework is employed. For any set of variables (c,π,Ξ,τ_R) within the admissible set

$\mathcal{A}(x)$ as defined in (5), along with a positive multiplier λ, the Lagrangian $\mathcal{L}$ is constructed based on the budget constraint provided in (9) as follows:

$$\begin{aligned}\mathcal{L} \equiv \mathbb{E}\Bigg[\int_0^{\tau_R} & e^{-(\beta+\delta)t}\left\{U(c_t)+\delta\bar{V}\left(\bar{X}_t\right)-y_t c_t\right.\\ &\left.+y_t\left(I_1\mathbf{1}_{\{\Xi_t=J_1\}}+I_2\mathbf{1}_{\{\Xi_t=J_2\}}-\delta\bar{X}_t-\epsilon\right)\right\}\mathrm{d}t\\ &+e^{-(\beta+\delta)\tau_R}\left(\bar{V}(X_{\tau_R})-y_{\tau_R}X_{\tau_R}\right)\Bigg]+\lambda x\\ \leqslant \mathbb{E}\Bigg[\int_0^{\tau_R} & e^{-(\beta+\delta)t}\left\{\tilde{u}_1(y_t)\mathbf{1}_{\{\Xi_t=J_1\}}+\tilde{u}_2(y_t)\mathbf{1}_{\{\Xi_t=J_2\}}\right.\\ &\left.+\delta\tilde{U}_R(y_t)\right\}\mathrm{d}t\\ &+e^{-(\beta+\delta)\tau_R}\tilde{U}_R(y_{\tau_R})\Bigg]+\lambda x,\end{aligned}$$

where $y_t \equiv \lambda e^{\beta t}H_t$ is defined as the dual variable following the process given below.

$$\frac{\mathrm{d}y_t}{y_t}=(\beta-r)\mathrm{d}t-\theta\mathrm{d}B_t,\quad y_0=\lambda. \tag{10}$$

Assumption 2 *We clarify two boundaries $\tilde{y}$ and $\hat{y}$ which are given later:*

$$\begin{aligned}\tilde{y} &\equiv L_1^{\gamma-\gamma^*}\omega^{-\gamma},\\ \hat{y} &\equiv \left(\frac{\gamma}{1-\gamma}\frac{L_2^{\frac{\gamma-\gamma^*}{\gamma}}-L_1^{\frac{\gamma-\gamma^*}{\gamma}}}{I_1-I_2}\right)^{\gamma}.\end{aligned} \tag{11}$$

To examine the effects of SCCs and job-switching opportunities, we assume that the relationship

$$\hat{y}<\tilde{y}$$

always holds.

Remark 1 For $\tau_R \leqslant t$, the post-retirement problem (6) is independent of unemployment risk. The solutions for the dual-value function, optimal consumption, optimal investment, and wealth level are given by

$$\begin{aligned}\tilde{U}_R(y_t)&=\frac{\gamma}{K(1-\gamma)}y_t^{-\frac{1-\gamma}{\gamma}},\quad c_t^*=y_t^{-\frac{1}{\gamma}},\\ \pi_t^*&=\frac{\theta}{\sigma}\frac{1}{\gamma K}y_t^{-\frac{1}{\gamma}},\quad X_t^*=\frac{1}{K}y_t^{-\frac{1}{\gamma}}.\end{aligned}$$

Additionally, $\tilde{u}_1(y_t)$ and $\tilde{u}_2(y_t)$ are the conjugate utility functions expressed as

$$\begin{aligned}\tilde{u}_1(y_t) &\equiv \sup_{c_t\geqslant\omega}\left\{\left(L_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}-y_t c_t\right)\mathbf{1}_{\{0<y_t\leqslant\tilde{y}\}}\right.\\ &\quad\left.+\left(L_1^{\gamma-\gamma^*}\frac{\omega^{1-\gamma}}{1-\gamma}-y_t\omega\right)\mathbf{1}_{\{y_t>\tilde{y}\}}\right\}+y_t(I_1-\epsilon)\\ &=\frac{\gamma}{1-\gamma}L_1^{\frac{\gamma-\gamma^*}{\gamma}}y_t^{-\frac{1-\gamma}{\gamma}}\mathbf{1}_{\{0<y_t\leqslant\tilde{y}\}}+\left(L_1^{\gamma-\gamma^*}\frac{\omega^{1-\gamma}}{1-\gamma}-y_t\omega\right)\\ &\quad\mathbf{1}_{\{y_t>\tilde{y}\}}+y_t(I_1-\epsilon),\end{aligned} \tag{12}$$

$$\begin{aligned}\tilde{u}_2(y_t) &\equiv \sup_{c_t\geqslant\omega}\left\{\left(L_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}-y_t c_t\right)\mathbf{1}_{\{0<y_t\leqslant\tilde{y}\}}\right\}+y_t(I_2-\epsilon)\\ &=\frac{\gamma}{1-\gamma}L_2^{\frac{\gamma-\gamma^*}{\gamma}}y_t^{-\frac{1-\gamma}{\gamma}}\mathbf{1}_{\{0<y_t\leqslant\tilde{y}\}}+y_t(I_2-\epsilon).\end{aligned} \tag{13}$$

Applying the first-order conditions (FOCs) directly to (12) and (13) yields the following optimal consumption:

$$c^*(y_t)=\begin{cases}\omega, & \text{for } y_t\geqslant\tilde{y} \text{ and } \Xi_t=J_1,\\ L_1^{\frac{\gamma-\gamma^*}{\gamma}}y_t^{-\frac{1}{\gamma}}, & \text{for } 0<y_t\leqslant\tilde{y} \text{ and } \Xi_t=J_1,\\ L_2^{\frac{\gamma-\gamma^*}{\gamma}}y_t^{-\frac{1}{\gamma}}, & \text{for } 0<y_t\leqslant\tilde{y} \text{ and } \Xi_t=J_2.\end{cases} \tag{14}$$

For $0<y_t\leqslant\hat{y}$, $\tilde{u}_2(y_t)\geqslant\tilde{u}_1(y_t)$ holds, while for $y_t>\hat{y}$, we observe $\tilde{u}_2(y_t)<\tilde{u}_1(y_t)$. These inequalities allow us to clarify the optimal job state as follows:

$$\Xi^*(y_t)=\begin{cases}J_1, & \text{for } \tilde{y}<y_t,\\ J_1, & \text{for } \hat{y}<y_t\leqslant\tilde{y},\\ J_2, & \text{for } 0<y_t\leqslant\hat{y}.\end{cases} \tag{15}$$

Condition $\tilde{u}_1(\cdot)=\tilde{u}_2(\cdot)$ represents an equilibrium in which the utility values associated with both job states are identical. This equilibrium allows us to define the dual job-switching boundary, denoted by $\hat{y}$, as follows:

$$\hat{y}\equiv\left(\frac{\gamma}{1-\gamma}\frac{L_2^{\frac{\gamma-\gamma^*}{\gamma}}-L_1^{\frac{\gamma-\gamma^*}{\gamma}}}{I_1-I_2}\right)^{\gamma}.$$

The positivity of the free boundary $\hat{y}$ is established under two conditions. When $\gamma^*>1$, it follows that $\gamma^*>\gamma>1$, which leads to $\frac{\gamma}{1-\gamma}<0$ and $L_2^{\frac{\gamma-\gamma^*}{\gamma}}-L_1^{\frac{\gamma-\gamma^*}{\gamma}}<0$. Conversely, when $0<\gamma^*<1$, we have $0<\gamma^*<\gamma<1$. Thus, $\frac{\gamma}{1-\gamma}>0$ and $L_2^{\frac{\gamma-\gamma^*}{\gamma}}-L_1^{\frac{\gamma-\gamma^*}{\gamma}}>0$. Therefore, in all scenarios, $\frac{\gamma}{1-\gamma}(L_2^{\frac{\gamma-\gamma^*}{\gamma}}-L_1^{\frac{\gamma-\gamma^*}{\gamma}})$ remains positive.

Subsequently, the associated dual problem corresponding to Problem 1 is presented below.

Problem (Dual problem) *Starting with a positive initial value λ, the function $\phi(\lambda)$ is defined as follows:*

$$\begin{aligned}\phi(\lambda)=\sup_{\tau_R\in\mathcal{T}}\mathbb{E}\Bigg[\int_0^{\tau_R} & e^{-(\beta+\delta)t}\left\{\tilde{u}_1(y_t)\mathbf{1}_{\{\hat{y}<y_t\}}\right.\\ &\left.+\tilde{u}_2(y_t)\mathbf{1}_{\{0<y_t\leqslant\hat{y}\}}+\delta\tilde{U}_R(y_t)\right\}\mathrm{d}t\\ &+e^{-(\beta+\delta)\tau_R}\tilde{U}_R(y_{\tau_R})\Bigg].\end{aligned}$$

Without a loss of generality, we introduce the time-dependent dual function $\psi(t,y)$ as follows:

$$\psi(t,y)\equiv\sup_{\tau_R\in\mathcal{T}}\mathbb{E}^{y_t=y}\left[\int_t^{\tau_R}e^{-(\beta+\delta)s}\left\{\tilde{u}_1(y_s)\mathbf{1}_{\{\hat{y}<y_s\}}\right.\right.$$

$$+\tilde{u}_2(y_s)\mathbf{1}_{\{0<y_s\leqslant\hat{y}\}} + \delta\tilde{U}_R(y_s)\Big\}\,\mathrm{d}s$$
$$+e^{-(\beta+\delta)\tau_R}\tilde{U}_R(y_{\tau_R})\Big]. \tag{16}$$

3.1. Solutions to voluntary retirement problem with SCCs

First, we examine the VR region depicted in figure 1, where $\tau \geqslant \tau_R$. In this region, individuals retire voluntarily without unemployment. This is because unemployment is delayed until the individual opts for voluntary retirement. By applying the variational inequality framework of Karatzas and Wang (2000) and Li *et al.* (2025), function (16) can be expressed as the following variational inequality.

Remark 2 The optimization problem in (16) corresponds to a classical optimal stopping problem. We exploit the well-established connection between optimal stopping problems and variational inequalities, which allows us to reformulate the original Problem 1 in terms of a variational inequality. This transformation leverages the inherent structure of the problem, enabling a more tractable mathematical formulation. Using the dynamic programming principle, along with the methodologies presented in Dai and Kwok (2006), Peskir and Shiryaev (2006), and Yang and Koo (2018), we derive the following equality for equation (16), which holds for $y>0$:

$$\max\Big\{\mathcal{L}^*\psi + e^{-(\beta+\delta)t}\Big(\tilde{u}_1(y)\mathbf{1}_{\{\hat{y}<y\}} + \tilde{u}_2(y)\mathbf{1}_{\{0<y\leqslant\hat{y}\}} + \delta\tilde{U}_R(y)\Big), e^{-(\beta+\delta)t}\tilde{U}_R(y) - \psi(t,y)\Big\} = 0.$$

Variational Inequality 1 Define a free boundary $\bar{y}$, equivalent to y_{τ_R}, and a function $\psi(\cdot,\cdot) \in C^1((0,\infty)\times\mathbb{R}^+) \cap C^2((0,\infty)\times(\mathbb{R}^+\backslash\{\bar{y}\}))$ satisfying the given conditions:

$$\mathcal{L}^*\psi + e^{-(\beta+\delta)t}\Big(\tilde{u}_1(y)\mathbf{1}_{\{\hat{y}<y\}} + \tilde{u}_2(y)\mathbf{1}_{\{0<y\leqslant\hat{y}\}} + \delta\tilde{U}_R(y)\Big) = 0, \quad y>\bar{y}, \tag{17}$$

$$\mathcal{L}^*\psi + e^{-(\beta+\delta)t}\Big(\tilde{u}_1(y)\mathbf{1}_{\{\hat{y}<y\}} + \tilde{u}_2(y)\mathbf{1}_{\{0<y\leqslant\hat{y}\}} + \delta\tilde{U}_R(y)\Big) < 0, \quad 0<y\leqslant\bar{y}, \tag{18}$$

$$\psi(t,y) > e^{-(\beta+\delta)t}\tilde{U}_R(y), \quad y>\bar{y}, \tag{19}$$

$$\psi(t,y) = e^{-(\beta+\delta)t}\tilde{U}_R(y), \quad 0<y\leqslant\bar{y}, \tag{20}$$

where $\mathcal{L}^*$ is characterized by

$$\mathcal{L}^* \equiv \frac{\partial}{\partial t} + (\beta - r)y\frac{\partial}{\partial y} + \frac{1}{2}\theta^2 y^2\frac{\partial^2}{\partial y^2}.$$

Proposition 3.1 *Function $v(y)$ is expressed as follows:*

$$v(y) = \begin{cases} Cy^{n_1} + \frac{\delta\gamma}{(1-\gamma)K(K+\delta)}y^{\frac{\gamma-1}{\gamma}} & \\ \quad + \frac{I_1-\epsilon-\omega}{r+\delta}y + \frac{L_1^{\gamma-\gamma^*}\omega^{1-\gamma}}{(\beta+\delta)(1-\gamma)}, & \text{for } \tilde{y}<y \\ B_1y^{n_2} + B_2y^{n_1} + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) & \\ \frac{\gamma}{(1-\gamma)(K+\delta)}y^{\frac{\gamma-1}{\gamma}} + \frac{I_1-\epsilon}{r+\delta}y, & \text{for } \hat{y}<y\leqslant\tilde{y}, \\ A_1y^{n_2} + A_2y^{n_1} + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) & \\ \frac{\gamma}{(1-\gamma)(K+\delta)}y^{\frac{\gamma-1}{\gamma}} + \frac{I_2-\epsilon}{r+\delta}y, & \text{for } \bar{y}<y\leqslant\hat{y}, \\ \frac{\gamma}{K(1-\gamma)}y^{-\frac{1-\gamma}{\gamma}}, & \text{for } 0<y\leqslant\bar{y}, \end{cases} \tag{21}$$

where coefficients A_1, A_2, B_1, B_2, and C are given by

$$A_1 = B_1 + \frac{\left(\frac{L_2^{\frac{\gamma-\gamma^*}{\gamma}} - L_1^{\frac{\gamma-\gamma^*}{\gamma}}}{1-\gamma}\right)^{\gamma}\left(\frac{\gamma}{I_1-I_2}\right)^{\gamma-1}\left(-\frac{\gamma n_1+1-\gamma}{K+\delta} + \gamma\frac{n_1-1}{r+\delta}\right)}{(n_1-n_2)\hat{y}^{n_2}} \tag{22}$$

$$A_2 = \frac{\frac{\gamma n_2+1-\gamma}{1-\gamma}\bar{y}^{-\frac{1}{\gamma}}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right] + \frac{\epsilon-I_2}{r+\delta}(n_2-1)}{(n_2-n_1)\bar{y}^{n_1-1}}, \tag{23}$$

$$B_1 = \frac{\frac{n_1}{(\beta+\delta)(1-\gamma)} - \frac{\gamma n_1+1-\gamma}{(K+\delta)(1-\gamma)} + \frac{1-n_1}{r+\delta}}{n_1-n_2}L_1^{(1-n_2)(\gamma-\gamma^*)}\omega^{1-\gamma+\gamma n_2}, \tag{24}$$

$$B_2 = A_2 + \left(\frac{L_2^{\frac{\gamma-\gamma^*}{\gamma}} - L_1^{\frac{\gamma-\gamma^*}{\gamma}}}{1-\gamma}\right)^{\gamma-\gamma n_1}\left(\frac{\gamma}{I_1-I_2}\right)^{\gamma-1-\gamma n_1}\frac{\frac{\gamma n_2+1-\gamma}{K+\delta} + \gamma\frac{1-n_2}{r+\delta}}{n_2-n_1}, \tag{25}$$

$$C = B_2 + \frac{\frac{n_2}{(\beta+\delta)(1-\gamma)} - \frac{1-\gamma+\gamma n_2}{(K+\delta)(1-\gamma)} + \frac{1-n_2}{r+\delta}}{n_1-n_2} L_1^{(1-n_1)(\gamma-\gamma^*)}\omega^{1-\gamma+\gamma n_1}. \tag{26}$$

$n_2>1$ and $n_1<0$ are the real roots of the following quadratic equation:

$$\mathcal{G}(n) \equiv \frac{1}{2}\theta^2n^2 + \left(\beta - r - \frac{1}{2}\theta^2\right)n - (\beta+\delta) = 0.$$

The free boundary $\bar{y}$ is the root of the following equation:

$$\frac{1}{1-\gamma}\frac{\gamma n_1+1-\gamma}{n_1-n_2}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right]\bar{y}^{-\frac{1}{\gamma}} - \frac{I_2-\epsilon}{r+\delta}\frac{1-n_1}{n_2-n_1} = A_1\bar{y}^{n_2-1}. \tag{27}$$

Subsequently, $\psi(t,y) = \mathrm{e}^{-(\beta+\delta)t}v(y)$ is a solution to Variational Inequality 1.

Proof See Appendix 1. ■

LEMMA 3.2 $B_1 < 0$.

Proof See Appendix 2. ■

REMARK 3 The sign of $\frac{1}{1-\gamma}[\frac{1}{K} - (L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K})\frac{1}{K+\delta}]$ is positive. Furthermore, it is difficult to verify the existence of a solution to equation (27) when $A_1 < 0$.

Thus, we make the following assumption to ensure the existence of $\bar{y}$ in equation (27).

ASSUMPTION 3 *We assume* $A_1 > 0$.

REMARK 4 When there are no SCCs, $B_1 = \omega = 0$ and $\lim_{\omega \downarrow 0} \tilde{y} = \infty$. Then, the wealth boundary at the time of voluntary retirement without SCCs takes the form

$$\bar{x}^N = \frac{1}{K}\left(\bar{y}^N\right)^{-\frac{1}{\gamma}},$$

where $\bar{y}^N$ is the root of the following algebraic equation:

$$\frac{1}{1-\gamma}\frac{\gamma n_1 + 1 - \gamma}{n_1 - n_2}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right]\left(\bar{y}^N\right)^{-\frac{1}{\gamma}} - \frac{I_2 - \epsilon}{r+\delta}\frac{1-n_1}{n_2-n_1} = \mathscr{H}\left(\bar{y}^N\right)^{n_2-1}, \tag{28}$$

with

$$\mathscr{H} = \frac{\left(\frac{L_2^{\frac{\gamma-\gamma^*}{\gamma}} - L_1^{\frac{\gamma-\gamma^*}{\gamma}}}{1-\gamma}\right)^{\gamma}\left(\frac{\gamma}{I_1 - I_2}\right)^{\gamma-1}\left(-\frac{\gamma n_1 + 1 - \gamma}{K+\delta} + \gamma\frac{n_1 - 1}{r+\delta}\right)}{(n_1 - n_2)\hat{y}^{n_2}} > 0.$$

PROPOSITION 3.3 *From Assumption 3, the voluntary retirement wealth threshold without SCCs is greater than that with SCCs; that is,*

$$\bar{x}^N > \bar{x}, \tag{29}$$

where $\bar{x}$ *is the optimal voluntary retirement wealth boundary with SCCs; basically,*

$$\bar{x} = \frac{1}{K}\bar{y}^{-\frac{1}{\gamma}}. \tag{30}$$

Proof See Appendix 3. ■

LEMMA 3.4 *We define a new function* $\mathscr{B}(z)$ *according to* (27) *as follows:*

$$\mathscr{B}(z) = z^{n_2-1}\mathscr{C}(z),$$

where

$$\mathscr{C}(z) = \frac{\gamma}{1-\gamma}\frac{n_1 + \frac{1-\gamma}{\gamma}}{n_1 - n_2}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right] z^{1-n_2-\frac{1}{\gamma}} - \frac{I_2 - \epsilon}{r+\delta}\frac{1-n_1}{n_2-n_1}z^{1-n_2} - A_1.$$

Define

$$\mathcal{Z} = \left[\frac{(n_2-1)(I_2-\epsilon)(1-n_1)(1-\gamma)}{(r+\delta)\left(n_1 + \frac{1-\gamma}{\gamma}\right)\left(\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right)(\gamma - \gamma n_2 - 1)}\right]^{-\gamma}.$$

From Assumption 3, the following condition

$$\mathcal{Z} > \underline{\mathcal{Z}} = \left[\frac{(I_2-\epsilon)(1-n_1)(1-\gamma)}{(r+\delta)\left(\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right)(\gamma - \gamma n_1 - 1)}\right]^{-\gamma}$$

holds, which implies that equation (27) *has two solutions in* $(0, \infty)$.

Proof See Appendix 4. ■

From the proof of Proposition 3.3 and the assumption that $A_1 > 0$, we can verify that $\mathscr{B}(\bar{y}^N) > 0$. Lemma 3.4 states that the two solutions to (27), denoted by $\bar{y}_1$ and $\bar{y}_2$, should satisfy $\bar{y}_1 < \mathcal{Z} < \bar{y}_2$. Thus, we derive the relationship $\bar{y}^N < \bar{y}_1 < \mathcal{Z} < \bar{y}_2$. Because $\lim_{\omega \downarrow 0} \bar{y} = \bar{y}^N$, $\bar{y}_1$ should be selected as the value of $\bar{y}$.

REMARK 5 From Lemma 3.4, we can easily verify that

$$\left[\frac{(n_2-1)(I_2-\epsilon)(1-\gamma)}{(r+\delta)\left(\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right)(\gamma n_2 - \gamma + 1)}\right]^{-\gamma} > \mathcal{Z} > \bar{y} = \bar{y}_1.$$

Hence, $A_2 > 0$.

REMARK 6 We restate the facts supported by Li *et al.* (2024) that:

$$\frac{\frac{1-\gamma}{\gamma} + n_1}{K+\delta} + \frac{1-n_1}{r+\delta} > 0, \quad \frac{\frac{1-\gamma}{\gamma} + n_2}{K+\delta} + \frac{1-n_2}{r+\delta} > 0.$$

Then, from Remark 5, $B_2 > 0$.

REMARK 7 Following a similar reasoning as in the proof of Lemma 3.2, it can be established that

$$\frac{n_2}{(\beta+\delta)(1-\gamma)} - \frac{1-\gamma+\gamma n_2}{(K+\delta)(1-\gamma)} + \frac{1-n_2}{r+\delta} < 0.$$

Consequently, based on Remark 6, it follows that $C > 0$.

It is evident that the equalities $\phi(\lambda) = \psi(0, \lambda) = v(\lambda)$ hold. The value function $V(x)$ and the function $\phi(\lambda)$ are related by the following equation:

$$V(x) = \inf_{\lambda > 0}\left[\phi(\lambda) + \lambda x\right]. \tag{31}$$

PROPOSITION 3.5 *The dual value function $\phi(\lambda)$ satisfies the following properties: $\phi(\lambda)$ is strictly convex for $\lambda > 0$, which implies that $\phi'(\lambda)$ is monotone increasing in $(0,\infty)$. $\lim_{\lambda\to 0+}\phi'(\lambda) = -\infty$ and $\lim_{\lambda\to\infty}\phi'(\lambda) = \frac{I_1-\epsilon-\omega}{r+\delta}$. Then, for the minimization problem in (31), there exists a unique solution λ^* such that*

$$x = -\phi'(\lambda^*). \tag{32}$$

Proof See Appendix 5. ∎

Consequently, the value function $V(x)$ can be derived directly.

THEOREM 3.6 (Main theorem) *For $(c^*, \pi^*, \Xi^*, \tau_R^*) \in \mathcal{A}(x)$, the value function corresponding to the optimal stopping problem in* (4) *is*

$$V(x) = \begin{cases} C(1-n_1)\left(\lambda_1^*\right)^{n_1} \\ +\frac{\delta}{K(K+\delta)(1-\gamma)}\left(\lambda_1^*\right)^{\frac{\gamma-1}{\gamma}} \\ +\frac{L_1^{\gamma-\gamma^*}\omega^{1-\gamma}}{(\beta+\delta)(1-\gamma)}, & for\ -\frac{I_1-\epsilon-\omega}{r+\delta} < x < \tilde{x} \text{ with } t < \tau_R^*, \\ B_1(1-n_2)\left(\lambda_2^*\right)^{n_2} \\ +B_2(1-n_1)\left(\lambda_2^*\right)^{n_1} \\ +\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \\ \frac{1}{(1-\gamma)(K+\delta)}\left(\lambda_2^*\right)^{-\frac{1-\gamma}{\gamma}}, & for\ \tilde{x} \leqslant x < \hat{x} \text{ with } t < \tau_R^*, \\ A_1(1-n_2)\left(\lambda_3^*\right)^{n_2} \\ +A_2(1-n_1)\left(\lambda_3^*\right)^{n_1} \\ +\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \\ \frac{1}{(1-\gamma)(K+\delta)}\left(\lambda_3^*\right)^{-\frac{1-\gamma}{\gamma}}, & for\ \hat{x} \leqslant x < \bar{x} \text{ with } t < \tau_R^*, \\ \frac{x^{1-\gamma}}{(1-\gamma)K^\gamma}, & for\ \tau_R^* \leqslant t, \end{cases}$$

where the values of λ_1^*, λ_2^*, λ_3^*, and λ_4^* are obtained using the following algebraic equations:

$$\begin{aligned} x &= -n_1 C\left(\lambda_1^*\right)^{n_1-1} + \frac{\delta}{K(K+\delta)}\left(\lambda_1^*\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta}, \\ x &= -n_2 B_1\left(\lambda_2^*\right)^{n_2-1} - n_1 B_2\left(\lambda_2^*\right)^{n_1-1} \\ &\quad + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\left(\lambda_2^*\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon}{r+\delta}, \\ x &= -n_2 A_1\left(\lambda_3^*\right)^{n_2-1} - n_1 A_2\left(\lambda_3^*\right)^{n_1-1} \\ &\quad + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\left(\lambda_3^*\right)^{-\frac{1}{\gamma}} - \frac{I_2-\epsilon}{r+\delta}, \\ x &= \frac{1}{K}\left(\lambda_4^*\right)^{-\frac{1}{\gamma}}. \end{aligned}$$

The optimal retirement time τ_R^* is identified as

$$\tau_R^* = \inf\left\{t \geqslant 0 \mid 0 < y_t^* \leqslant \bar{y}\right\}.$$

$\tilde{x}$ represents the wealth threshold associated with the subsistence consumption level described in (2)*; that is,*

$$\tilde{x} = -n_1 C\tilde{y}^{n_1-1} + \frac{\delta}{K(K+\delta)}\tilde{y}^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta}.$$

$\hat{x}$ is the job-switching wealth boundary; that is,

$$\begin{aligned} \hat{x} &= -n_2 A_1\hat{y}^{n_2-1} - n_1 A_2\hat{y}^{n_1-1} \\ &\quad + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\hat{y}^{-\frac{1}{\gamma}} - \frac{I_2-\epsilon}{r+\delta}. \end{aligned}$$

Given that $y_t^{\lambda_p^*}$ solves the stochastic differential equation (10) with the initial condition λ_p^* for $p \in \{1,2,3,4\}$, the optimal wealth level X_t^* can be directly obtained using (32) as follows:

$$X_t^* = \begin{cases} -n_1 C\left(y_t^{\lambda_1^*}\right)^{n_1-1} \\ +\frac{\delta}{K(K+\delta)}\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta}, & \text{for } \tilde{y} < y_t^{\lambda_1^*} \text{ with } t < \tau_R^*, \\ -n_2 B_1\left(y_t^{\lambda_2^*}\right)^{n_2-1} \\ -n_1 B_2\left(y_t^{\lambda_2^*}\right)^{n_1-1} \\ +\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \\ \frac{1}{K+\delta}\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon}{r+\delta}, & \text{for } \hat{y} < y_t^{\lambda_2^*} \leqslant \tilde{y} \text{ with } t < \tau_R^*, \\ -n_2 A_1\left(y_t^{\lambda_3^*}\right)^{n_2-1} \\ -n_1 A_2\left(y_t^{\lambda_3^*}\right)^{n_1-1} \\ +\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \\ \frac{1}{K+\delta}\left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}} - \frac{I_2-\epsilon}{r+\delta}, & \text{for } \bar{y} < y_t^{\lambda_3^*} \leqslant \hat{y} \text{ with } t < \tau_R^*, \\ \frac{1}{K}\left(y_t^{\lambda_4^*}\right)^{-\frac{1}{\gamma}}, & \text{for } \tau_R^* \leqslant t. \end{cases} \tag{33}$$

THEOREM 3.7 (Main theorem) *The optimal solutions for the job state and consumption processes, $(\Xi_t^*)_{t\geqslant 0}$ and $(c_t^*)_{t\geqslant 0}$, are established by*

$$\Xi_t^* = \begin{cases} J_1, & for\ \tilde{y} < y_t^{\lambda_1^*} \text{ with } t < \tau_R^* \\ J_1, & for\ \hat{y} < y_t^{\lambda_2^*} \leqslant \tilde{y} \text{ with } t < \tau_R^*, \\ J_2, & for\ \bar{y} < y_t^{\lambda_3^*} \leqslant \hat{y} \text{ with } t < \tau_R^*, \end{cases}$$

$$c_t^* = \begin{cases} \omega, & for\ \tilde{y} < y_t^{\lambda_1^*} \text{ with } t < \tau_R^*, \\ L_1^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}, & for\ \hat{y} < y_t^{\lambda_2^*} \leqslant \tilde{y} \text{ with } t < \tau_R^*, \\ L_2^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}}, & for\ \bar{y} < y_t^{\lambda_3^*} \leqslant \hat{y} \text{ with } t < \tau_R^*, \\ \left(y_t^{\lambda_4^*}\right)^{-\frac{1}{\gamma}}, & for\ \tau_R^* \leqslant t. \end{cases}$$

The optimal portfolio process $(\pi_t^*)_{t\geqslant 0}$ *is given by*

$$\pi_t^* = \begin{cases} \frac{\theta}{\sigma}\left[n_1(n_1-1)C\left(y_t^{\lambda_1^*}\right)^{n_1-1} + \frac{\delta}{\gamma K(K+\delta)}\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}}\right], & \text{for } \tilde{y} < y_t^{\lambda_1^*} \text{ with } t < \tau_R^*, \\ \frac{\theta}{\sigma}\left[n_2(n_2-1)B_1\left(y_t^{\lambda_2^*}\right)^{n_2-1} + n_1(n_1-1)B_2\left(y_t^{\lambda_2^*}\right)^{n_1-1} + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{\gamma(K+\delta)}\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}\right], & \text{for } \hat{y} < y_t^{\lambda_2^*} \leqslant \tilde{y} \text{ with } t < \tau_R^*, \\ \frac{\theta}{\sigma}\left[n_2(n_2-1)A_1\left(y_t^{\lambda_3^*}\right)^{n_2-1} + n_1(n_1-1)A_2\left(y_t^{\lambda_3^*}\right)^{n_1-1} + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{\gamma(K+\delta)}\left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}}\right], & \text{for } \bar{y} < y_t^{\lambda_3^*} \leqslant \hat{y} \text{ with } t < \tau_R^*, \\ \frac{\theta}{\sigma}\frac{1}{\gamma K}\left(y_t^{\lambda_4^*}\right)^{-\frac{1}{\gamma}}, & \text{for } \tau_R^* \leqslant t. \end{cases}$$

Proof See Appendix 6. ■

3.2. Solutions to involuntary retirement with SCCs

We now explore the scenario in which an unemployment event occurs before planned voluntary retirement; that is, $\tau < \tau_R^*$. In the context of our study, which includes job-switching opportunities, it is crucial to consider the risk of unemployment in each job state along an individual's career path. This consideration is particularly important given that job transitions can significantly impact retirement planning and financial stability.

Specifically, when involuntary retirement occurs in job state J_1 before reaching the switching boundary $\hat{x}$, the objective function is expressed as

$$V(x) = V^{IR}(x) \equiv \sup_{(c,\pi)\in\mathcal{A}_2(x)} \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(L_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma} + \delta\bar{V}\left(\bar{X}_t\right)\right)\mathrm{d}t\right],$$

where $\mathcal{A}_2(x)$ is the set of all admissible strategies (c,π) and is defined as

$$\mathcal{A}_2(x) \equiv \left\{(c,\pi)\,|\,X_t > -\frac{I_1-\epsilon-\omega}{r+\delta}\right\}.$$

LEMMA 3.8 *If involuntary retirement occurs before switching to job state* J_2*, then, subject to the SCCs (2), the solutions for the value function* $V^{IR}(x)$*, the optimal consumption process* c_t^{**}*, and the optimal portfolio process* π_t^{**} *with* $(c^{**},\pi^{**}) \in \mathcal{A}_2(x)$*, are determined as follows:*

$$V^{IR}(x) = \begin{cases} \tilde{C}(1-n_1)\left(\tilde{\lambda}_1^*\right)^{n_1} + \frac{\delta}{K(K+\delta)(1-\gamma)}\left(\tilde{\lambda}_1^*\right)^{\frac{\gamma-1}{\gamma}} + \frac{L_1^{\gamma-\gamma^*}\omega^{1-\gamma}}{(\beta+\delta)(1-\gamma)}, & \text{for } -\frac{I_1-\epsilon-\omega}{r+\delta} < x < \tilde{x}^{IR}, \\ \tilde{B}_1(1-n_2)\left(\tilde{\lambda}_2^*\right)^{n_2} + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{(1-\gamma)(K+\delta)}\left(\tilde{\lambda}_2^*\right)^{-\frac{1-\gamma}{\gamma}}, & \text{for } \tilde{x}^{IR} \leqslant x, \end{cases}$$

$$c_t^{**} = \begin{cases} \omega, & \text{for } \tilde{y} < y_t^{\tilde{\lambda}_1^*}, \\ L_1^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\tilde{\lambda}_2^*}\right)^{-\frac{1}{\gamma}}, & \text{for } 0 < y_t^{\tilde{\lambda}_2^*} \leqslant \tilde{y}, \end{cases}$$

$$\pi_t^{**} = \begin{cases} \frac{\theta}{\sigma}\left[n_1(n_1-1)\tilde{C}\left(y_t^{\tilde{\lambda}_1^*}\right)^{n_1-1} + \frac{\delta}{\gamma K(K+\delta)}\left(y_t^{\tilde{\lambda}_1^*}\right)^{-\frac{1}{\gamma}}\right], & \text{for } \tilde{y} < y_t^{\tilde{\lambda}_1^*}, \\ \frac{\theta}{\sigma}\left[n_2(n_2-1)\tilde{B}_1\left(y_t^{\tilde{\lambda}_2^*}\right)^{n_2-1} + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{\gamma(K+\delta)}\left(y_t^{\tilde{\lambda}_2^*}\right)^{-\frac{1}{\gamma}}\right], & \text{for } 0 < y_t^{\tilde{\lambda}_2^*} \leqslant \tilde{y}, \end{cases}$$

with

$$\tilde{B}_1 = B_1, \quad \tilde{C} = \frac{\frac{n_2}{(\beta+\delta)(1-\gamma)} - \frac{1-\gamma+\gamma n_2}{(K+\delta)(1-\gamma)} + \frac{1-n_2}{r+\delta}}{n_1-n_2} L_1^{(1-n_1)(\gamma-\gamma^*)}\omega^{1-\gamma+\gamma n_1},$$

$$\tilde{x}^{IR} = -n_1\tilde{C}\tilde{y}^{n_1-1} + \frac{\delta}{K(K+\delta)}\tilde{y}^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta}.$$

$\tilde{\lambda}_1^*$ *and* $\tilde{\lambda}_2^*$ *are determined using the following algebraic equations:*

$$\begin{aligned} x &= -n_1\tilde{C}\left(\tilde{\lambda}_1^*\right)^{n_1-1} + \frac{\delta}{K(K+\delta)}\left(\tilde{\lambda}_1^*\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta}, \\ x &= -n_2\tilde{B}_1\left(\tilde{\lambda}_2^*\right)^{n_2-1} + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\left(\tilde{\lambda}_2^*\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon}{r+\delta}. \end{aligned} \tag{34}$$

Proof See Appendix 7. ■

When involuntary retirement occurs in job state J_2 after job switching before voluntary retirement, the objective function is characterized as:

$$V(x) = V^{IRJS}(x) \equiv \sup_{(c,\pi,\Xi)\in\mathcal{A}_3(x)} \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(U(c_t) + \delta\bar{V}\left(\bar{X}_t\right)\right)\mathrm{d}t\right],$$

where $\mathcal{A}_3(x)$ is the set of all admissible strategies (c,π,Ξ) and is defined as

$$\mathcal{A}_3(x) \equiv \left\{(c,\pi,\Xi)\,|\,X_t > -\frac{I_1-\epsilon-\omega}{r+\delta}\right\}.$$

LEMMA 3.9 *If involuntary retirement occurs after job switching but before voluntary retirement, then, subject to the SCCs (2), the solutions for the value function $V^{IRJS}(x)$, the optimal consumption process c_t^{***}, and the optimal portfolio process π_t^{***} with $(c^{***}, \pi^{***}) \in \mathcal{A}_3(x)$, are deduced as follows:*

$$V^{IRJS}(x) =$$

$$\begin{cases} \hat{C}(1-n_1)\left(\hat{\lambda}_1^*\right)^{n_1} + \frac{\delta}{K(K+\delta)(1-\gamma)}\left(\lambda_1^*\right)^{\frac{\gamma-1}{\gamma}} \\ +\frac{L_1^{\frac{\gamma-\gamma^*}{\gamma}}\omega^{1-\gamma}}{(\beta+\delta)(1-\gamma)}, & for\ -\frac{I_1-\epsilon-\omega}{r+\delta} < x < \tilde{x}^{IRJS}, \\ \hat{B}_1(1-n_2)\left(\hat{\lambda}_2^*\right)^{n_2} + \hat{B}_2(1-n_1)\left(\hat{\lambda}_2^*\right)^{n_1} \\ +\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{(1-\gamma)(K+\delta)}\left(\hat{\lambda}_2^*\right)^{-\frac{1-\gamma}{\gamma}}, & for\ \tilde{x}^{IRJS} \leqslant x < \hat{x}^{IRJS}, \\ \hat{A}_1(1-n_2)\left(\hat{\lambda}_3^*\right)^{n_2} \\ +\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{(1-\gamma)(K+\delta)}\left(\hat{\lambda}_3^*\right)^{-\frac{1-\gamma}{\gamma}}, & for\ \hat{x}^{IRJS} \leqslant x, \end{cases}$$

$$c_t^{***} =$$

$$\begin{cases} \omega, & for\ \tilde{y} < y_t^{\hat{\lambda}_1^*}, \\ L_1^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\hat{\lambda}_2^*}\right)^{-\frac{1}{\gamma}}, & for\ \hat{y} < y_t^{\hat{\lambda}_2^*} \leqslant \tilde{y}, \\ L_2^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\hat{\lambda}_3^*}\right)^{-\frac{1}{\gamma}}, & for\ 0 < y_t^{\hat{\lambda}_3^*} \leqslant \hat{y}, \end{cases}$$

$$\pi_t^{***} =$$

$$\begin{cases} \frac{\theta}{\sigma}\left[n_1(n_1-1)\hat{C}\left(y_t^{\hat{\lambda}_1^*}\right)^{n_1-1}\right. \\ \left.+\frac{\delta}{\gamma K(K+\delta)}\left(y_t^{\hat{\lambda}_1^*}\right)^{-\frac{1}{\gamma}}\right], & for\ \tilde{y} < y_t^{\hat{\lambda}_1^*}, \\ \frac{\theta}{\sigma}\left[n_2(n_2-1)\hat{B}_1\left(y_t^{\hat{\lambda}_2^*}\right)^{n_2-1}\right. \\ +n_1(n_1-1)\hat{B}_2\left(y_t^{\hat{\lambda}_2^*}\right)^{n_1-1} \\ \left.+\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{\gamma(K+\delta)}\left(y_t^{\hat{\lambda}_2^*}\right)^{-\frac{1}{\gamma}}\right], & for\ \hat{y} < y_t^{\hat{\lambda}_2^*} \leqslant \tilde{y}, \\ \frac{\theta}{\sigma}\left[n_2(n_2-1)\hat{A}_1\left(y_t^{\hat{\lambda}_3^*}\right)^{n_2-1}\right. \\ \left.+\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{\gamma(K+\delta)}\left(y_t^{\hat{\lambda}_3^*}\right)^{-\frac{1}{\gamma}}\right], & for\ 0 < y_t^{\hat{\lambda}_3^*} \leqslant \hat{y}, \end{cases}$$

with

$$\hat{A}_1 = A_1, \quad \hat{B}_1 = B_1,$$

$$\hat{B}_2 = \left(\frac{L_2^{\frac{\gamma-\gamma^*}{\gamma}} - L_1^{\frac{\gamma-\gamma^*}{\gamma}}}{1-\gamma}\right)^{\gamma-\gamma n_1}\left(\frac{\gamma}{I_1-I_2}\right)^{\gamma-1-\gamma n_1} \frac{\frac{\gamma n_2+1-\gamma}{K+\delta} + \gamma\frac{1-n_2}{r+\delta}}{n_2-n_1},$$

$$\hat{C} = \hat{B}_2 + \frac{\frac{n_2}{(\beta+\delta)(1-\gamma)} - \frac{1-\gamma+\gamma n_2}{(K+\delta)(1-\gamma)} + \frac{1-n_2}{r+\delta}}{n_1-n_2} L_1^{\frac{(1-n_1)(\gamma-\gamma^*)}{\gamma}}\omega^{1-\gamma+\gamma n_1},$$

$$\tilde{x}^{IRJS} = -n_1\hat{C}\tilde{y}^{n_1-1} + \frac{\delta}{K(K+\delta)}\tilde{y}^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta},$$

$$\hat{x}^{IRJS} = -n_2\hat{A}_1\hat{y}^{n_2-1} + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\hat{y}^{-\frac{1}{\gamma}} - \frac{I_2-\epsilon}{r+\delta}.$$

$\hat{\lambda}_1^*$, $\hat{\lambda}_2^*$, and $\hat{\lambda}_3^*$ are determined using the following algebraic equations:

$$x = -n_1\hat{C}\left(\hat{\lambda}_1^*\right)^{n_1-1} + \frac{\delta}{K(K+\delta)}\left(\hat{\lambda}_1^*\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta},$$

$$\begin{aligned} x = &-n_2\hat{B}_1\left(\hat{\lambda}_2^*\right)^{n_2-1} - n_1\hat{B}_2\left(\hat{\lambda}_2^*\right)^{n_1-1} \\ &+\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\left(\hat{\lambda}_2^*\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon}{r+\delta}, \end{aligned} \tag{35}$$

$$\begin{aligned} x = &-n_2\hat{A}_1\left(\hat{\lambda}_3^*\right)^{n_2-1} + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\left(\hat{\lambda}_3^*\right)^{-\frac{1}{\gamma}} \\ &-\frac{I_2-\epsilon}{r+\delta}. \end{aligned} \tag{36}$$

Proof See Appendix 8. ■

LEMMA 3.10 *The following relationship holds for the corresponding wealth levels:*

$$\tilde{x}^{IR} < \tilde{x}^{IRJS} < \tilde{x}.$$

Proof See Appendix 9. ■

4. Economic implications

In this section, we perform a comprehensive analysis using the closed-form solutions derived from Theorems 3.6 and 3.7, as well as Lemmas 3.8 and 3.9. Our primary objective is to examine the impact of unemployment risk, loading factors, and subsistence consumption levels on individuals' decisions regarding job-switching and voluntary retirement. Additionally, we investigate consumption and investment choices prior to retirement, whether they are voluntary or involuntary. To achieve these aims, we analytically provide comparative statics to demonstrate the properties of the solutions. We also conduct numerical simulations using the following baseline parameter values:

$$\begin{aligned} &\beta = 0.03; \quad \gamma^* = 2; \quad r = 0.03; \quad I_1 = 1; \\ &I_2 = 0.3; \quad L_1 = 0.3; \quad L_2 = 0.8; \\ &\alpha = 0.5; \quad \mu = 0.08; \quad \sigma = 0.2; \\ &\delta = 0.01; \quad \epsilon = 0.2; \quad \omega = 0.6. \end{aligned} \tag{37}$$

The values in (37) are set by referring to various studies, such as Jang *et al.* (2013), Bensoussan *et al.* (2016), and Li *et al.* (2025). Among these, δ and ϵ, representing the forced unemployment possibility and the loading factor, respectively, are set to reflect the market values, according to Jang *et al.* (2013), Bensoussan *et al.* (2016), Acemoglu and Shimer (2000), and Cocco *et al.* (2005). Various values are used in subsequent analyses.

4.1. Voluntary retirement

PROPOSITION 4.1 *From Assumption 3, free boundary $\bar{y}$ increases monotonically with the subsistence consumption*

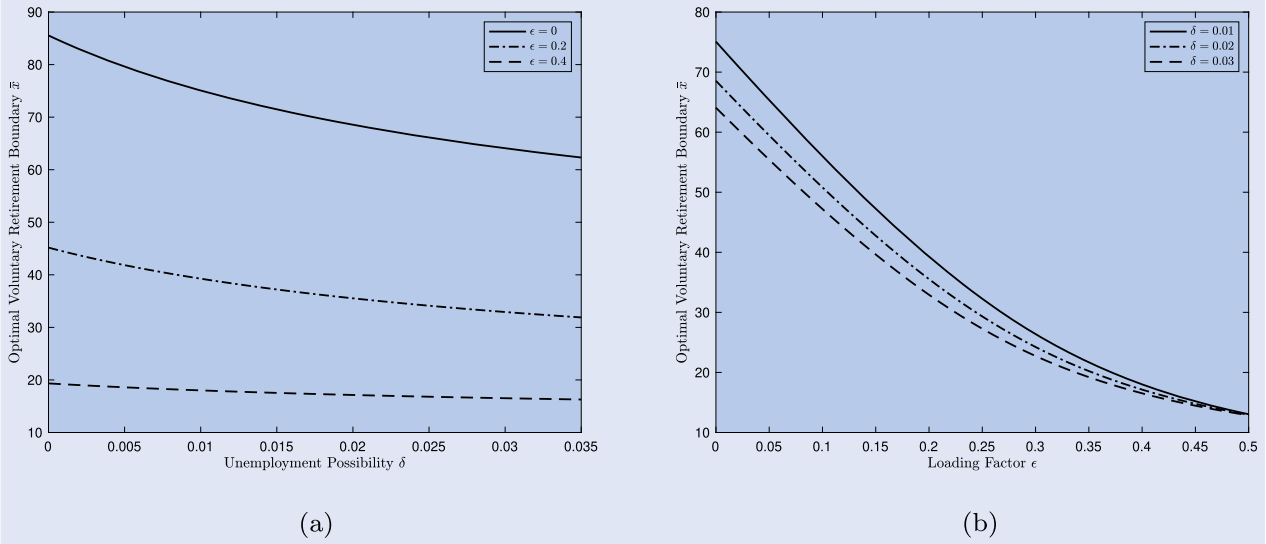


Figure 2. Sensitivity analysis on optimal voluntary retirement boundary. (a) Sensitivity w.r.t. δ and (b) Sensitivity w.r.t. ϵ.

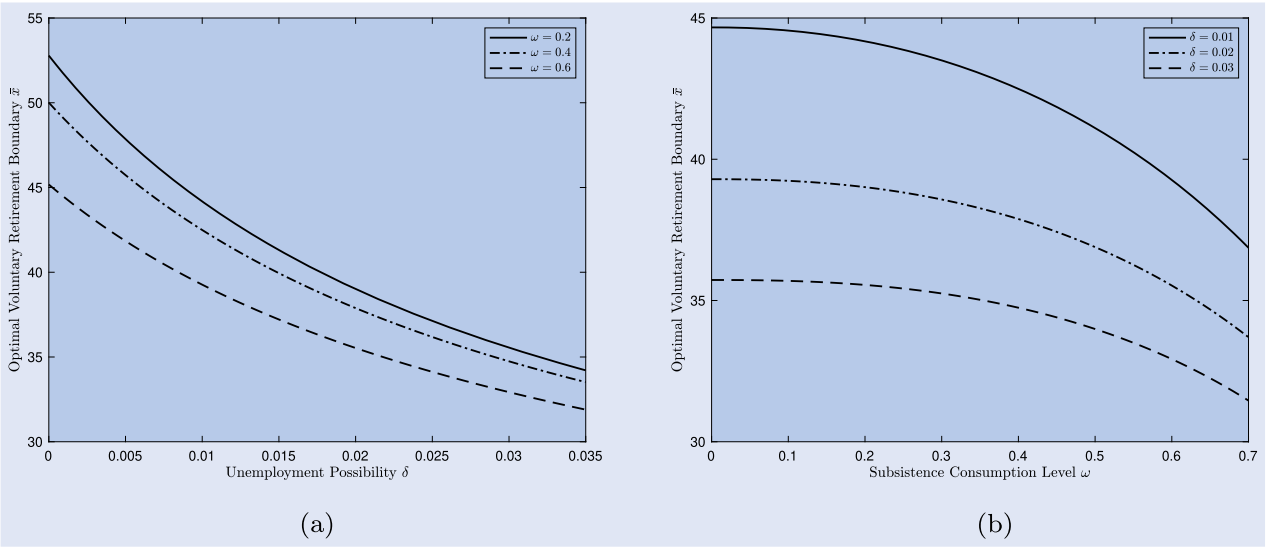


Figure 3. Sensitivity analysis on optimal voluntary retirement boundary. (a) Sensitivity w.r.t. δ and (b) Sensitivity w.r.t. ω.

level ω. The corresponding voluntary retirement critical $\bar{x}$ decreases with ω.

Proof See Appendix 10. ■

PROPOSITION 4.2 *From Assumption 3, the free boundary $\bar{y}$ monotonically increases with the loading factor ϵ. The corresponding voluntary retirement critical $\bar{x}$ decreases with ϵ.*

Proof See Appendix 11. ■

This subsection delves into how variations in unemployment intensity, loading factor, and subsistence consumption levels affect the optimal voluntary retirement wealth boundary, consumption, and investment strategies. Figure 2 illustrates the sensitivity of the optimal voluntary retirement threshold to changes in the probability of forced unemployment and the loading factor.

A key observation is that an increase in either the probability of forced unemployment or the loading factor reduces the optimal voluntary retirement wealth boundary. This finding aligns with previous research, particularly the results presented in Jang *et al.* (2013). When the likelihood of unemployment increases, individuals face significant income uncertainty, which makes them worried about their ability to pay for essential future consumption. Individuals opt for early retirement to mitigate the financial risks associated with unemployment. Moreover, the loading factor represents costs beyond the basic premium. An increased loading factor signifies a higher expense. This leads to higher retirement costs and reduced disposable income. To avoid the higher costs and negative financial impact of a high-loading factor, individuals choose to retire earlier. This finding aligns with Proposition 4.2.

The impact of changes in the subsistence consumption level on the optimal voluntary retirement boundary is well illustrated in figure 3. As shown in figure 3, an increase in subsistence consumption leads to a decrease in the optimal voluntary retirement boundary. When subsistence consumption levels increase, individuals should accumulate more wealth

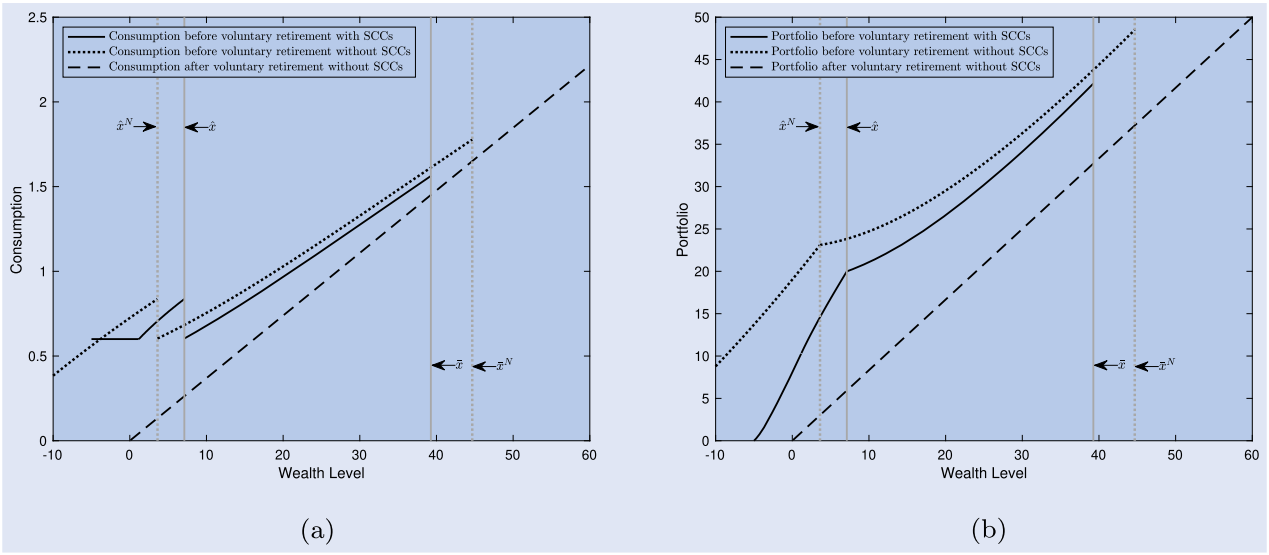


Figure 4. Impacts of SCCs on the optimal strategies of voluntary retirement problem with unemployment risks and job switching. (a) Consumption and (b) Portfolio.

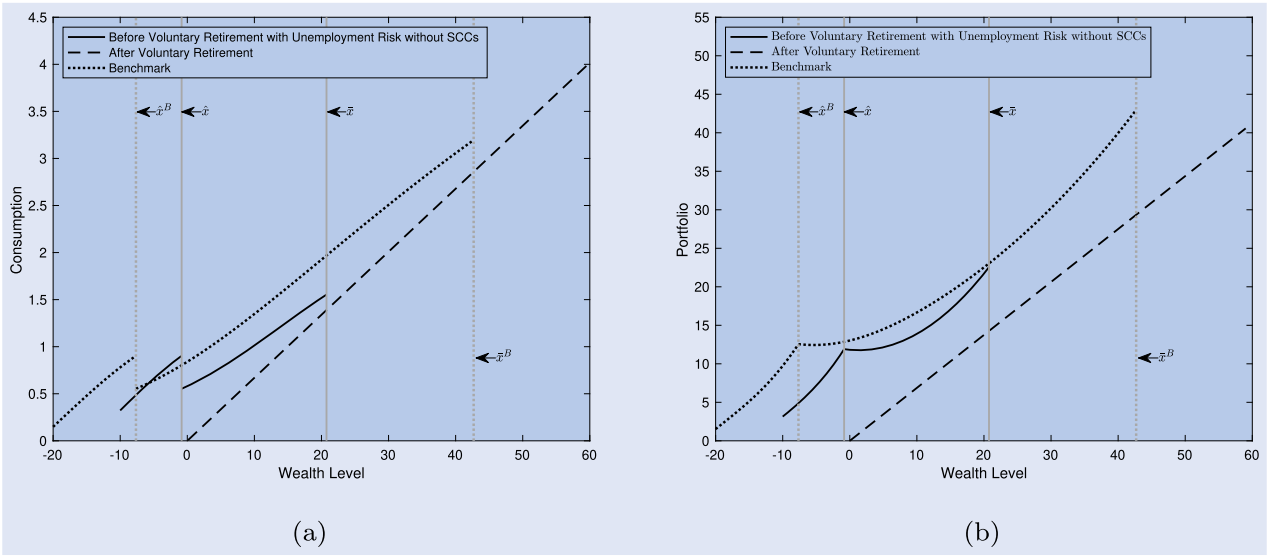


Figure 5. Impacts of unemployment risks on the optimal strategies of voluntary retirement problem. ($\beta = 0.07; \gamma^* = 3; r = 0.045; I_1 = 1; I_2 = 0.3; L_1 = 0.3; L_2 = 0.8; \alpha = 0.5; \mu = 0.1; \sigma = 0.2; \epsilon = 0.2; \omega = 0; \delta = 0.01;$). (a) Consumption and (b) Portfolio.

during their working years to meet the rising costs of living. To avoid the potential unemployment risks and uncertainties that could result in additional costs, they may choose to retire early at a lower level of wealth. This result is consistent with that of Proposition 4.1.

Figure 4 illustrates the impact of subsistence consumption levels on consumption and investment decisions. In the optimal voluntary retirement model, which incorporates job switching and unemployment risk, the presence of SCCs results in delayed job switching and an inclination toward earlier voluntary retirement, which is consistent with the result from Proposition 3.3. In addition, SCCs lead to reduced consumption and investments before retirement. These conclusions align with the existing literature findings (e.g. Gong and Li 2006, Lakner and Ma Nygren 2006, Mai *et al.* 2024). When individuals face SCCs, they tend to remain in their current high-income jobs for extended periods to ensure income stability and minimize the risk of income fluctuations associated with job changes. However, SCCs significantly limit individuals' financial flexibility, compelling them to allocate most of their income to basic survival needs, thereby reducing their ability to engage in higher levels of consumption and investment. To alleviate the pressures in daily life, individuals prefer to opt for earlier voluntary retirement.

These findings provide a crucial theoretical foundation for policymakers to consider strategies to mitigate the impact of labor market risks on individual economic behavior. First, by implementing policies aimed at reducing the risk of unemployment and facilitating smooth job transitions, policymakers can minimize negative economic behaviors triggered by subsistence consumption levels. Second, policymakers should consider enacting social security policies to ensure that individuals' basic living needs are met during unemployment or job-switching periods. This can alleviate the suppressive effects of subsistence consumption levels on consumption and investment. Programs such as unemployment benefits and basic living allowances can maintain a certain level of consumption, even amid economic uncertainty.

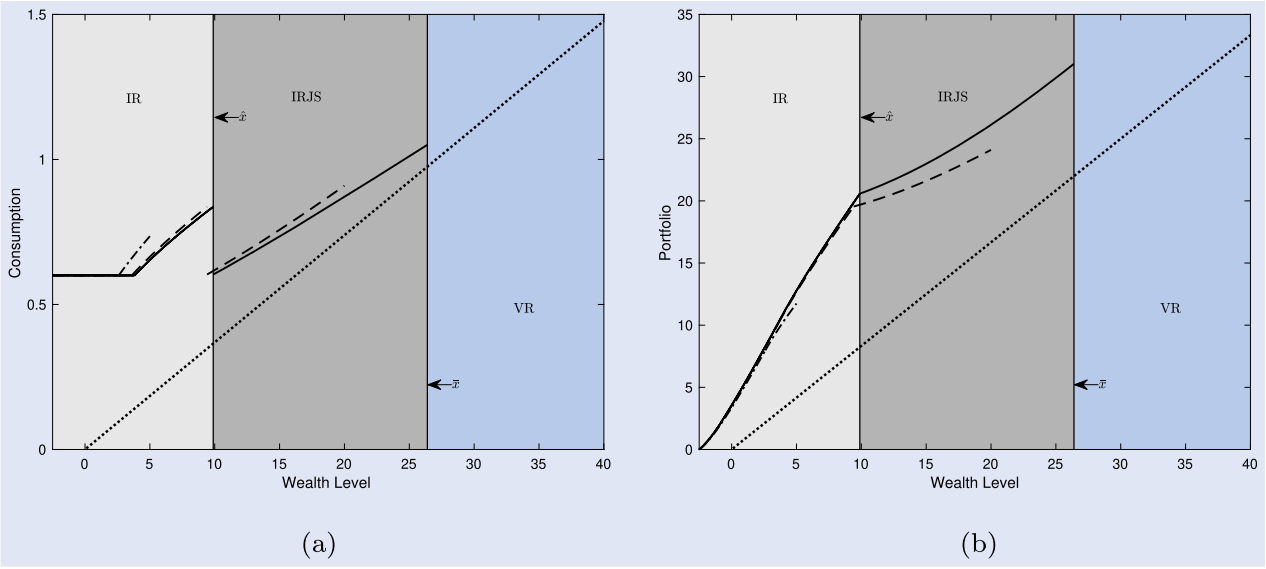


Figure 6. Optimal strategies with a highly competitive insurance market with $\epsilon = 0.3$ among two involuntary retirement cases and a voluntary retirement. The solid line represents voluntary retirement whereas the dash-dotted line denotes the involuntary retirement occurring at job state J_1. The dashed line depicts the involuntary retirement occurring at job state J_2. The dotted line indicates the case after retirement both voluntarily and involuntarily. (a) Consumption and (b) Portfolio.

As shown in figure 5, the benchmark model represents the voluntary retirement model with job switching, excluding unemployment risks and SCCs. This model is thoroughly examined in Problem 3 in Li *et al.* (2025), with the corresponding solutions derived directly from Theorem 4.2 in the same reference. To focus specifically on the effects of unemployment risk, SCCs are not considered in this analysis. From figure 5, we observe the following. When unemployment risk and unemployment insurance are considered, individuals' consumption and investment levels before voluntary retirement are reduced compared with the benchmark model. Furthermore, individuals exposed to unemployment risk and purchasing unemployment insurance tend to delay job switching but advance the timing of voluntary retirement.

These findings highlight the significant influence of unemployment risk and insurance on consumption and investment behaviors. Unemployment insurance provides a financial safety net that allows individuals to manage employment risks more effectively, thereby influencing their economic decisions before and after retirement. Thus, policymakers should consider these dynamics when designing social security systems. Unemployment risks impact pre-retirement economic decisions and play a crucial role in determining the timing of job switching and voluntary retirement. Understanding these interactions is essential for developing policies that support financial stability and well-being throughout an individual's career and retirement.

4.2. *Involuntary retirement*

Proposition 4.3 *The consumption rates of the three cases have the following relationship:*

$$\begin{cases} c_t^{**} > c_t^{***} > c_t^*, & \text{if } X_t^* \in \left(\tilde{x}, \hat{x}^{IRJS}\right), \\ c_t^{***} > c_t^*, & \text{if } X_t^* \in \left(\hat{x}, \bar{x}\right). \end{cases}$$

Proof See Appendix 12. ■

Proposition 4.4 *The investment satisfies the following relationship:*

$$\pi_t^{***} < \pi_t^*, \quad \text{if } X_t^* \in \left(\hat{x}, \bar{x}\right).$$

Proof See Appendix 13. ■

In this subsection, we analyze the differences in consumption and investment decisions among the three potential retirement statuses that individuals might encounter: two scenarios of involuntary retirement and one of voluntary retirement, as detailed in figure 1. We use different line styles to represent each scenario: the solid line represents voluntary retirement and the dashed-dotted line denotes involuntary retirement occurring in job state J_1. The dashed line depicts involuntary retirement occurring in job state J_2. The dotted line indicates the situation after retirement, both voluntarily and involuntarily.

Figure 6 illustrates a highly competitive insurance market with a loading factor of 30%. As described in Section 2 and illustrated in figure 2, the loading factor indicates the costs beyond the basic unemployment insurance premium. A lower loading factor indicates lower additional expense. Compared with figure 7, which considers a low-competition insurance market with a loading factor of 40%, the voluntary retirement boundary in figure 6 is larger. This finding is examined analytically in Proposition 4.2 and numerically in figure 2. Moreover, the job-switching boundary in figure 6 is smaller than that in figure 7.

Regarding consumption and investment decisions among the three potential retirement scenarios, the following observations are made irrespective of the extra unemployment insurance premium. Figures 6 and 7 show that in the IR region, the consumption level is the highest when unemployment occurs in job state J_1 compared to the other scenarios. However, the consumption level is the lowest in the case of

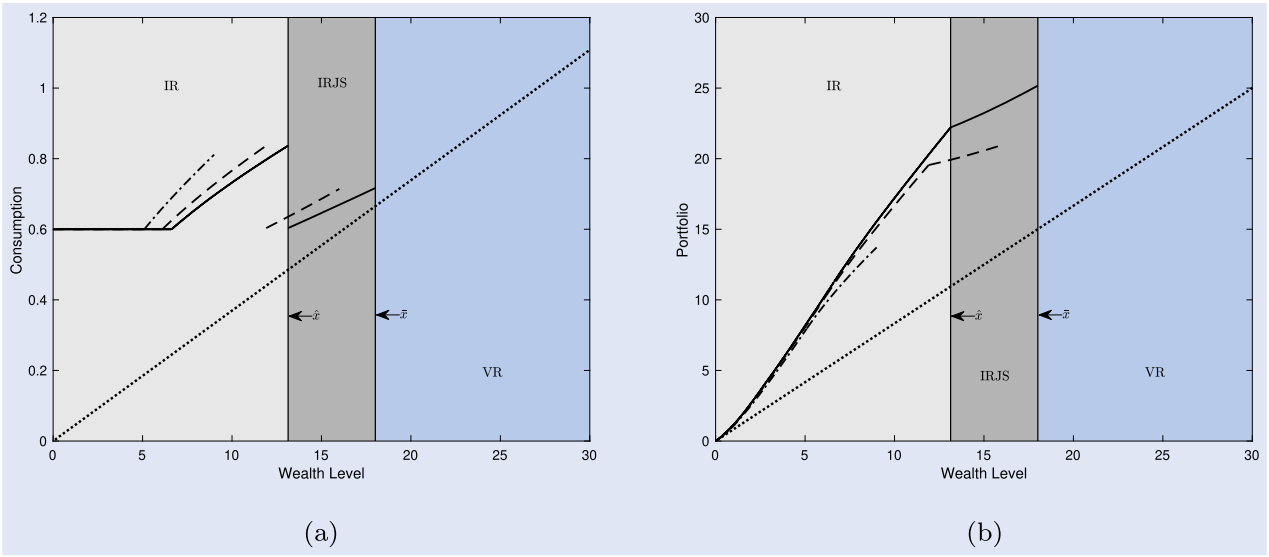


Figure 7. Optimal strategies with a lowly competitive insurance market with $\epsilon = 0.4$ among two involuntary retirement cases and a voluntary retirement. The solid line represents voluntary retirement whereas the dash-dotted line denotes the involuntary retirement occurring at job state J_1. The dashed line depicts the involuntary retirement occurring at job state J_2. The dotted line indicates the case after retirement both voluntarily and involuntarily. (a) Consumption and (b) Portfolio.

voluntary retirement. In the IRJS region, when unemployment occurs in the job state J_2, the consumption level is higher than in the voluntary retirement scenario. However, the investment strategy in both regions follows a trend opposite to that of consumption. Propositions 4.3 and 4.4 support these findings.

In cases where involuntary retirement occurs at low levels of individual wealth (i.e. in the IR or IRJS region), unemployment insurance partially offsets the financial losses resulting from forced unemployment and plays a crucial role in maintaining the financial stability of the unemployed. Moreover, consistent with the conclusions drawn by Hopenhayn and Nicolini (1997) in their empirical analysis, this study further confirms the crucial role of unemployment insurance in facilitating consumption smoothing and risk sharing through theoretical modeling.

Notably, our model assumes that once individuals retire, whether voluntarily or involuntarily, they do not reenter the workforce. Involuntary retirement is typically caused by deteriorating health or layoffs owing to company closure. The assumption of treating involuntary retirement as a permanent state aligns with the assumptions found in authoritative studies such as Cocco *et al.* (2005). From an empirical perspective, Smith (2006) provided strong support for this hypothesis. Specifically, by analyzing data from the British Household Panel Survey, Smith (2006) found that over 70% of involuntary retirements are caused by health issues and layoffs, with nearly 50% attributed to illness and over 20% attributed to layoffs. Morissette (2004) and Picot (1992) emphasized the critical role of company size in determining layoff rates. Their researches indicate that permanent layoffs or involuntary retirements resulting from layoffs are more likely to occur in smaller companies. Additionally, Dorn and Sousa-Poza (2010) used international microdata from 19 industrialized countries to explore the phenomena of voluntary and involuntary early retirement, providing empirical evidence for this study. The findings indicate that involuntary early retirement is particularly common in Continental Europe, particularly in countries that have experienced economic downturns and have enforced strict employment protection laws, where the proportion of involuntary retirees is significantly higher than that of early retirees.

These findings provide valuable insights into how unemployment risk and insurance market conditions influence consumption and investment decisions before and after retirement. Understanding the interplay between unemployment risk, insurance premiums, and individual economic behaviors is crucial for developing effective policies that enhance financial stability and support economic well-being throughout an individual's working life and retirement. For example, insurance companies can enhance unemployment insurance coverage by lowering the loading factor and broadening eligibility criteria. Policymakers can strengthen social safety nets such as unemployment benefits and basic living allowances.

5. Conclusion

In this study, we explored an optimal retirement model with unemployment risk and job switching under SCCs using a rigorous continuous-time optimal portfolio selection framework. Our analysis encompassed both voluntary and involuntary retirement scenarios. We framed voluntary retirement as an optimal stopping problem, and incorporated job switching to model three realistic retirement scenarios. By employing the Martingale method and variational inequality, we derived closed-form solutions and performed comprehensive comparative statics, both analytically and numerically.

The primary findings of this study are summarized as follows. First, it introduces the integration of unemployment risk into an optimal investment and consumption model that includes both voluntary retirement and job-switching opportunities, reflecting dual labor flexibility and incorporating SCCs. Unlike previous studies, this model allows individuals

to flexibly choose their retirement date and manage their labor income and leisure time dynamically before retirement, better reflecting the different retirement scenarios in which unemployment risk affects pre-retirement consumption and investment decisions. Additionally, this study compares consumption and investment strategies between voluntary retirement models with and without unemployment risk, revealing that unemployment risk leads to delayed job switching, earlier voluntary retirement, and reduced pre-retirement consumption and investment. Furthermore, it explores the role of unemployment insurance as a hedge against unemployment risk and it mitigates financial pressure during forced unemployment. Particularly in low-competition insurance markets, the wealth boundary for voluntary retirement is reduced and the levels of consumption and investment before retirement are low. Moreover, this study highlights the impact of SCCs, which require individuals to allocate resources to meet basic living standards, thereby lowering consumption and investment levels before retirement and decreasing the necessary wealth threshold for voluntary retirement.

The study's limitation is its exclusive focus on fixed unemployment risk, which restricts its ability to accurately model broader labor and financial markets. This limitation arises because of the objective of deriving a closed-form solution. However, incorporating time-varying unemployment risks (where the unemployment intensity follows a stochastic or deterministic process) makes deriving such solutions more difficult. Consequently, addressing time-varying unemployment risks while achieving closed-form solutions will be the focus of future research.

Acknowledgments

The authors would like to express their sincere gratitude to the two anonymous referees for their insightful and constructive comments, which have greatly improved the quality of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Yong Hyun Shin was supported by the National Research Foundation of Korea (NRF) Grant [Grant No. RS-2023-00240574], and Ji-Hun Yoon was supported by the NRF Grant [Grant No. 2022R1A5A1033624 and No. RS-2023-NR076791].

ORCID

Qi Li http://orcid.org/0000-0003-1785-8591
Yong Hyun Shin http://orcid.org/0000-0002-0299-1536
Ji-Hun Yoon http://orcid.org/0000-0002-2943-4027

References

Acemoglu, D. and Shimer, R., Productivity gains from unemployment insurance. *Eur. Econ. Rev.*, 2000, **44**, 1195–1224.
Bensoussan, A., Jang, B.G. and Park, S., Unemployment risks and optimal retirement in an incomplete market. *Oper. Res.*, 2016, **64**, 1015–1032.
Choi, K.J., Shim, G. and Shin, Y.H., Optimal portfolio, consumption-leisure and retirement choice problem with CES utility. *Math. Finance Int. J. Math. Stat. Financ. Econ.*, 2008, **18**, 445–472.
Cocco, J.F., Gomes, F.J. and Maenhout, P.J., Consumption and portfolio choice over the life cycle. *Rev. Financ. Stud.*, 2005, **18**, 491–533.
Cox, J.C. and Huang, C.F., Optimal consumption and portfolio policies when asset prices follow a diffusion process. *J. Econ. Theory.*, 1989, **49**, 33–83.
Dai, M. and Kwok, Y.K., Characterization of optimal stopping regions of American Asian and lookback options. *Math. Finance Int. J. Math. Stat. Financ. Econ.*, 2006, **16**, 63–82.
Dorn, D. and Sousa-Poza, A., 'Voluntary' and 'involuntary' early retirement: An international analysis. *Appl. Econ.*, 2010, **42**, 427–438.
Dybvig, P.H. and Liu, H., Lifetime consumption and investment: Retirement and constrained borrowing. *J. Econ. Theory.*, 2010, **145**, 885–907.
Dybvig, P.H. and Liu, H., Verification theorems for models of optimal consumption and investment with retirement and constrained borrowing. *Math. Oper. Res.*, 2011, **36**, 620–635.
Farhi, E. and Panageas, S., Saving and investing for early retirement: A theoretical analysis. *J.Financ.Econ.*, 2007, **83**, 87–121.
Gong, N. and Li, T., Role of index bonds in an optimal dynamic asset allocation model with real subsistence consumption. *Appl. Math. Comput.*, 2006, **174**, 710–731.
Hopenhayn, H.A. and Nicolini, J.P., Optimal unemployment insurance. *J. Political Econ.*, 1997, **105**, 412–438.
Jang, B.G., Park, S. and Rhee, Y., Optimal retirement with unemployment risks. *J. Bank. Financ.*, 2013, **37**, 3585–3604.
Karatzas, I. and Wang, H., Utility maximization with discretionary stopping. *SIAM J. Control Opt.*, 2000, **39**, 306–329.
Lakner, P. and Ma Nygren, L., Portfolio optimization with downside constraints. *Math. Finance Int. J. Math. Stat. Financ. Econ.*, 2006, **16**, 283–299.
Li, Q., Shin, Y.H. and Yoon, J.H., Job switching, mortality risk, and bequest motive in an optimal consumption–investment model under inflation risk. Working Paper, 2024.
Li, Q., Shin, Y.H. and Yoon, J.H., The effects of inflation risk on voluntary retirement and job switching by a Martingale approach [online]. American Institute of Mathematical Sciences. Available online at: https://www.aimsciences.org/article/id/67ab05783942fc6603063b96 (accessed February 2025), 2025.
Mai, R., Yang, Z., Lai, Y. and Lin, J., Portfolio-consumption choice problem with voluntary retirement and consumption constraints. *J. Comput. Appl. Math.*, 2024, **445**, 115839.
Merton, R.C., Lifetime portfolio selection under uncertainty: The continuous-time case. *Rev. Econ. Stat.*, 1969, **51**, 247–257.
Merton, R.C., Optimum consumption and portfolio rules in a continuous-time model. *J. Econ. Theory.*, 1971, **3**, 373–413.
Morissette, R., *Permanent layoff rates*, Statistics Canada, 2004.
Peskir, G. and Shiryaev, A., *Optimal Stopping and Free-Boundary Problems*, Lectures in Mathematics ETH Zürich, 2006 (Birkhäuser Verlag: Basel).
Picot, G., Workers on the move: Permanent layoffs. *Perspectives on Labour and Income, Autumn, Statistics Canada, Ottawa*, 1992.
Salop, S.C., A model of the natural rate of unemployment. *Am. Econ. Rev.*, 1979, **69**, 117–125.

Shapiro, C. and Stiglitz, J.E., Equilibrium unemployment as a worker discipline device. *Am. Econ. Rev.*, 1984, **74**, 433–444.
Shim, G. and Shin, Y.H., An optimal job, consumption/leisure, and investment policy. *Oper. Res. Lett.*, 2014, **42**, 145–149.
Smith, S., The retirement-consumption puzzle and involuntary early retirement: Evidence from the British Household Panel Survey. *Econ. J.*, 2006, **116**, C130–C148.
Yang, Z. and Koo, H.K., Optimal consumption and portfolio selection with early retirement option. *Math. Oper. Res.*, 2018, **43**, 1378–1404.

Appendices

Appendix 1. Proof of Proposition 3.1

For $y > \bar{y}$, we redefine function $\psi(t,y)$ in the following manner:

$$\psi(t,y) \equiv \begin{cases} \psi_1(t,y), & \text{for } \tilde{y} < y, \\ \psi_2(t,y), & \text{for } \hat{y} < y \leqslant \tilde{y}, \\ \psi_3(t,y), & \text{for } \bar{y} < y \leqslant \hat{y}, \end{cases}$$

which serves as the solution to the partial differential equation within Variational Inequality 1:

$$\begin{cases} \frac{\partial \psi_1}{\partial t} + (\beta - r) y \frac{\partial \psi_1}{\partial y} + \frac{1}{2}\theta^2 y^2 \frac{\partial^2 \psi_1}{\partial y^2} \\ + \mathrm{e}^{-(\beta+\delta)t} \left[L_1^{\gamma-\gamma^*} \frac{\omega^{1-\gamma}}{1-\gamma} \right. \\ \left. + y(I_1 - \epsilon - \omega) + \frac{\delta\gamma}{K(1-\gamma)} y^{-\frac{1-\gamma}{\gamma}} \right] = 0, & \text{for } \tilde{y} < y, \\ \frac{\partial \psi_2}{\partial t} + (\beta - r) y \frac{\partial \psi_2}{\partial y} + \frac{1}{2}\theta^2 y^2 \frac{\partial^2 \psi_2}{\partial y^2} \\ + \mathrm{e}^{-(\beta+\delta)t} \left[\frac{\gamma}{1-\gamma} \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) y^{-\frac{1-\gamma}{\gamma}} \right. \\ \left. + (I_1 - \epsilon) y \right] = 0, & \text{for } \hat{y} < y \leqslant \tilde{y}, \\ \frac{\partial \psi_3}{\partial t} + (\beta - r) y \frac{\partial \psi_3}{\partial y} + \frac{1}{2}\theta^2 y^2 \frac{\partial^2 \psi_3}{\partial y^2} \\ + \mathrm{e}^{-(\beta+\delta)t} \left[\frac{\gamma}{1-\gamma} \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) y^{-\frac{1-\gamma}{\gamma}} \right. \\ \left. + (I_2 - \epsilon) y \right] = 0, & \text{for } \bar{y} < y \leqslant \hat{y}. \end{cases} \tag{A1}$$

We propose a trial solution of the form $\psi(t,y) = e^{-(\beta+\delta)t} v(y)$ and substitute it into (A1) to yield the following ordinary differential equation:

$$\begin{cases} \frac{1}{2}\theta^2 y^2 v''(y) + (\beta - r) y v'(y) \\ - (\beta + \delta) v(y) \\ + L_1^{\gamma-\gamma^*} \frac{\omega^{1-\gamma}}{1-\gamma} \\ + y(I_1 - \epsilon - \omega) + \frac{\delta\gamma}{K(1-\gamma)} y^{-\frac{1-\gamma}{\gamma}} = 0, & \text{for } \tilde{y} < y, \\ \frac{1}{2}\theta^2 y^2 v''(y) + (\beta - r) y v'(y) \\ - (\beta + \delta) v(y) \\ + \frac{\gamma}{1-\gamma} \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) y^{-\frac{1-\gamma}{\gamma}} \\ + (I_1 - \epsilon) y = 0, & \text{for } \hat{y} < y \leqslant \tilde{y}, \\ \frac{1}{2}\theta^2 y^2 v''(y) + (\beta - r) y v'(y) \\ - (\beta + \delta) v(y) \\ + \frac{\gamma}{1-\gamma} \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) y^{-\frac{1-\gamma}{\gamma}} \\ + (I_2 - \epsilon) y = 0, & \text{for } \bar{y} < y \leqslant \hat{y}. \end{cases} \tag{A2}$$

(A2) takes a classical form of the Cauchy-Euler equation. For the case where $\tilde{y} < y$, the particular solution to the first equation in (A2) can be conjectured as follows:

$$v_p(y) = \mathscr{K}_1 y + \mathscr{K}_2 y^{-\frac{1-\gamma}{\gamma}} + \mathscr{K}_3. \tag{A3}$$

By substituting (A3) into the first equation of (A2), the coefficients $\mathscr{K}_1$, $\mathscr{K}_2$, and $\mathscr{K}_3$ can be determined as follows:

$$\mathscr{K}_1 = \frac{I_1 - \epsilon - \omega}{r + \delta}, \quad \mathscr{K}_2 = \frac{\delta\gamma}{(1-\gamma) K (K + \delta)},$$
$$\mathscr{K}_3 = \frac{L_1^{\gamma-\gamma^*} \omega^{1-\gamma}}{(\beta + \delta)(1-\gamma)}.$$

The general solution to the first equation in (A2) can be conjectured below:

$$v_g(y) = C y^{n_1} + D y^{n_2}.$$

Since the model is well-defined, the coefficient D is set to zero, while the coefficient C is determined later by using the boundary conditions. Hence, the solution to the first equation in (A2) is given by

$$v(y) = v_g(y) + v_p(y), \quad \text{for } \tilde{y} < y.$$

Applying the same approach to the second and third equations in (A2) results in the solution to (A2) as follows:

$$v(y) = \begin{cases} C y^{n_1} + \frac{\delta\gamma}{(1-\gamma)K(K+\delta)} y^{\frac{\gamma-1}{\gamma}} \\ + \frac{I_1 - \epsilon - \omega}{r+\delta} y + \frac{L_1^{\gamma-\gamma^*}\omega^{1-\gamma}}{(\beta+\delta)(1-\gamma)}, & \text{for } \tilde{y} < y, \\ B_1 y^{n_2} + B_2 y^{n_1} + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) \\ \frac{\gamma}{(1-\gamma)(K+\delta)} y^{\frac{\gamma-1}{\gamma}} + \frac{I_1 - \epsilon}{r+\delta} y, & \text{for } \hat{y} < y \leqslant \tilde{y}, \\ A_1 y^{n_2} + A_2 y^{n_1} + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) \\ \frac{\gamma}{(1-\gamma)(K+\delta)} y^{\frac{\gamma-1}{\gamma}} + \frac{I_2 - \epsilon}{r+\delta} y, & \text{for } \bar{y} < y \leqslant \hat{y}. \end{cases}$$

By applying the boundary conditions at $\hat{y}$ and $\bar{y}$, we obtain $v(\hat{y}-) = v(\hat{y}+)$, $v'(\hat{y}-) = v'(\hat{y}+)$, $v(\bar{y}-) = v(\bar{y}+)$, and $v'(\bar{y}-) = v'(\bar{y}+)$. When the consumption rate reaches the subsistence level as specified by constraints (2), the associated dual variable $\tilde{y} = L_1^{\gamma-\gamma^*}\omega^{-\gamma}$ ensures that $v(\tilde{y}-) = v(\tilde{y}+)$ and $v'(\tilde{y}-) = v'(\tilde{y}+)$. Consequently, the following conclusions are drawn:

$$\begin{aligned} & C\tilde{y}^{n_1} + \frac{\delta\gamma}{(1-\gamma)K(K+\delta)} \tilde{y}^{\frac{\gamma-1}{\gamma}} \\ & \quad + \frac{I_1 - \epsilon - \omega}{r+\delta} \tilde{y} + \frac{L_1^{\gamma-\gamma^*}\omega^{1-\gamma}}{(\beta+\delta)(1-\gamma)} \\ & \quad = B_1 \tilde{y}^{n_2} + B_2 \tilde{y}^{n_1} + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) \frac{\gamma}{(1-\gamma)(K+\delta)} \tilde{y}^{\frac{\gamma-1}{\gamma}} \\ & \quad + \frac{I_1 - \epsilon}{r+\delta} \tilde{y}, \end{aligned} \tag{A4}$$

$$\begin{aligned} & n_1 C\tilde{y}^{n_1-1} - \frac{\delta}{K(K+\delta)} \tilde{y}^{-\frac{1}{\gamma}} + \frac{I_1 - \epsilon - \omega}{r+\delta} \\ & \quad = n_2 B_1 \tilde{y}^{n_2-1} + n_1 B_2 \tilde{y}^{n_1-1} - \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) \frac{1}{K+\delta} \tilde{y}^{-\frac{1}{\gamma}} \\ & \quad + \frac{I_1 - \epsilon}{r+\delta}, \end{aligned} \tag{A5}$$

$$\begin{aligned} & B_1 \hat{y}^{n_2} + B_2 \hat{y}^{n_1} + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) \frac{\gamma}{(1-\gamma)(K+\delta)} \hat{y}^{\frac{\gamma-1}{\gamma}} + \frac{I_1 - \epsilon}{r+\delta} \hat{y} \\ & \quad = A_1 \hat{y}^{n_2} + A_2 \hat{y}^{n_1} + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) \frac{\gamma}{(1-\gamma)(K+\delta)} \hat{y}^{\frac{\gamma-1}{\gamma}} \\ & \quad + \frac{I_2 - \epsilon}{r+\delta} \hat{y}, \end{aligned} \tag{A6}$$

$$n_2 B_1 \hat{y}^{n_2-1} + n_1 B_2 \hat{y}^{n_1-1} - \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K} \right) \frac{1}{K+\delta} \hat{y}^{-\frac{1}{\gamma}} + \frac{I_1 - \epsilon}{r+\delta}$$

$$= n_2 A_1 \hat{y}^{n_2-1} + n_1 A_2 \hat{y}^{n_1-1} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\hat{y}^{-\frac{1}{\gamma}} + \frac{I_2-\epsilon}{r+\delta}, \tag{A7}$$

$$\frac{\gamma}{(1-\gamma)K}\bar{y}^{\frac{\gamma-1}{\gamma}} = A_1\bar{y}^{n_2} + A_2\bar{y}^{n_1} + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{\gamma}{(1-\gamma)(K+\delta)}\bar{y}^{\frac{\gamma-1}{\gamma}} + \frac{I_2-\epsilon}{r+\delta}\bar{y}, \tag{A8}$$

$$-\frac{1}{K}\bar{y}^{-\frac{1}{\gamma}} = n_2 A_1 \bar{y}^{n_2-1} + n_1 A_2 \bar{y}^{n_1-1} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\bar{y}^{-\frac{1}{\gamma}} + \frac{I_2-\epsilon}{r+\delta}. \tag{A9}$$

By utilizing (A4) and (A5), B_1 is derived as shown in (24). From the expressions in (11), (A6), and (A7), A_1 is obtained as indicated in (22). Similarly, A_2 is determined from (A8) and (A9) as presented in (23). Substituting the values of B_1, A_1, and A_2 into (A6) allows us to determine B_2 as shown in (25). Furthermore, by inserting B_1 and B_2 into (A4), we calculate C as outlined in (26). Inserting A_1 and A_2 into (A8) results in an algebraic equation, where free boundary $\bar{y}$ is determined as the root, as specified in (27). When incorporating the solution to the post-retirement problem mentioned in Remark 1, function $\psi(t,y) = \mathrm{e}^{-(\beta+\delta)t}v(y)$ emerges as a solution to (17) and (20). Additionally, inequalities (18) and (19) in Variational Inequality 1 are satisfied by following the methodology described in Li *et al.* (2025).

Appendix 2. Proof of Lemma 3.2

It is clear that $n_1 - n_2 < 0$. The first inequality in Assumption 1 states that $\beta - r - (\beta+\delta)\gamma \geqslant -(r+\delta)\gamma$ with $\gamma^* \in (0,1)$. Assumption 1 states that $K + \delta > r + \delta + \frac{\beta-r}{\gamma}$ with $\gamma^* \in (1,\infty)$. $\mathscr{G}(-\frac{1-\gamma}{\gamma}) = -K - \delta < 0$ yields that $n_1 < -\frac{1-\gamma}{\gamma} < 1 < n_2$, which implies that $\frac{1}{K+\delta}(1 + \frac{\gamma n_1}{1-\gamma}) < 0$ with $\gamma^* \in (0,1)$.

Then, when $0 < \gamma^* < 1$,

$$\begin{aligned}
&\frac{n_1}{(\beta+\delta)(1-\gamma)} - \frac{\gamma n_1 + 1 - \gamma}{(K+\delta)(1-\gamma)} + \frac{1-n_1}{r+\delta} \\
&= -\frac{1}{K+\delta}\left(1 + \frac{\gamma n_1}{1-\gamma}\right) + \frac{1}{r+\delta} + \frac{r-\beta+(\beta+\delta)\gamma}{(\beta+\delta)(r+\delta)(1-\gamma)}n_1 \\
&> \frac{1}{r+\delta} + \frac{\gamma}{(\beta+\delta)(1-\gamma)}n_1 > 0.
\end{aligned}$$

When $\gamma^* > 1$,

$$\begin{aligned}
&\frac{n_1}{(\beta+\delta)(1-\gamma)} - \frac{\gamma n_1 + 1 - \gamma}{(K+\delta)(1-\gamma)} + \frac{1-n_1}{r+\delta} \\
&= -\frac{1}{K+\delta}\left(1 + \frac{\gamma n_1}{1-\gamma} - \frac{1-n_1}{r+\delta}(K+\delta) - \frac{n_1}{(\beta+\delta)(1-\gamma)}(K+\delta)\right) \\
&> -\frac{1}{K+\delta}\left(1 + \frac{\gamma n_1}{1-\gamma} - \frac{1-n_1}{r+\delta}\left(r+\delta+\frac{\beta-r}{\gamma}\right) - \frac{n_1}{(\beta+\delta)(1-\gamma)}\left(r+\delta+\frac{\beta-r}{\gamma}\right)\right) \\
&= -\frac{1}{K+\delta}\left(-\frac{\beta-r}{\gamma(r+\delta)} + \frac{n_1(\beta-r)^2}{(\beta+\delta)\gamma(r+\delta)} - \frac{n_1(r+\delta)}{(\beta+\delta)(1-\gamma)}\right) > 0.
\end{aligned}$$

Consequently, $B_1 < 0$ always holds.

Appendix 3. Proof of Proposition 3.3

From (27) and (28), we deduce that

$$\begin{aligned}
&\frac{1}{1-\gamma}\frac{\gamma n_1 + 1 - \gamma}{n_1 - n_2}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right]\left(\bar{y}^N\right)^{-\frac{1}{\gamma}} - \mathscr{H}\left(\bar{y}^N\right)^{n_2-1} \\
&= \frac{1}{1-\gamma}\frac{\gamma n_1 + 1 - \gamma}{n_1 - n_2}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right]\bar{y}^{-\frac{1}{\gamma}} - (\mathscr{H} + B_1)\bar{y}^{n_2-1}.
\end{aligned} \tag{A10}$$

Define a new function $f(y) \equiv \frac{1}{1-\gamma}\frac{\gamma n_1+1-\gamma}{n_1-n_2}[\frac{1}{K} - (L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K})\frac{1}{K+\delta}]y^{-\frac{1}{\gamma}} - \mathscr{H}y^{n_2-1}$. Thus, it is clear that $f(\cdot)$ decreased. From (A10), we have $f(\bar{y}^N) - f(\bar{y}) = -B_1\bar{y}^{n_2-1} > 0$. Thus, the following equation holds.

$$\bar{y}^N < \bar{y},$$

which generates inequality (29).

Appendix 4. Proof of Lemma 3.4

$$\mathscr{C}'(z) = \mathscr{D}(z)z^{-n_2},$$

where

$$\mathscr{D}(z) = \left(1 - n_2 - \frac{1}{\gamma}\right)\frac{\gamma}{1-\gamma}\frac{n_1 + \frac{1-\gamma}{\gamma}}{n_1 - n_2}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right]z^{-\frac{1}{\gamma}} - (1-n_2)\frac{I_2-\epsilon}{r+\delta}\frac{1-n_1}{n_2-n_1}.$$

Since $\mathscr{D}'(z) = (1 - n_2 - \frac{1}{\gamma})\frac{1}{\gamma-1}\frac{n_1+\frac{1-\gamma}{\gamma}}{n_1-n_2}[\frac{1}{K} - (L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K})\frac{1}{K+\delta}]z^{-\frac{1}{\gamma}} > 0$ implies that $\mathscr{D}(\cdot)$ is monotonically increasing in $z \in (0,\infty)$. Then, there is a unique $\mathcal{Z}$ making $\mathscr{D}(\cdot) = 0$ because $\lim_{z\downarrow 0}\mathscr{D}(z) = -\infty$ and $\lim_{z\uparrow\infty}\mathscr{D}(z) = -(1-n_2)\frac{I_2-\epsilon}{r+\delta}\frac{1-n_1}{n_2-n_1} > 0$. Thus, for $0 < z < \mathcal{Z}$, we have $\mathscr{C}'(z) < 0$ which implies $\mathscr{C}(\cdot)$ is decreasing monotonically. For $z \geqslant \mathcal{Z}$, we have $\mathscr{C}'(z) \geqslant 0$ which implies $\mathscr{C}(\cdot)$ increases monotonically. Subsequently, $\mathscr{C}(\cdot)$ reaches its minimum at $\mathcal{Z}$. It is trivial to verify that $\mathscr{C}(\mathcal{Z}) < 0$ because $A_1 > 0$ is assumed. Hence, we have $\mathscr{C}(\mathcal{Z}) < 0$ and equation (27) has two solutions in $(0,\infty)$.

Appendix 5. Proof of Proposition 3.5

For $\lambda > \tilde{y}$,

$$\phi''(\lambda) = n_1(n_1-1)C\lambda^{n_1-2} + \frac{\delta}{K(K+\delta)\gamma}\lambda^{-\frac{1}{\gamma}-1} > 0,$$

since $C > 0$ which is described in Remark 7.

For $\hat{y} < \lambda \leqslant \tilde{y}$,

$$\phi''(\lambda) = n_2(n_2-1)B_1\lambda^{n_2-2} + n_1(n_1-1)B_2\lambda^{n_1-2}$$

$$+\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{1}{(K+\delta)\gamma}\lambda^{-\frac{1}{\gamma}-1}$$
$$=\lambda^{n_2-2}\mathscr{E}(\lambda),$$

where

$$\mathscr{E}(\lambda)=n_2(n_2-1)B_1+n_1(n_1-1)B_2\lambda^{n_1-n_2}$$
$$+\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{1}{(K+\delta)\gamma}\lambda^{-\frac{1}{\gamma}-n_2+1}.$$

$B_2>0$ stated in Remark 6 says that

$$\mathscr{E}'(\lambda)=(n_1-n_2)n_1(n_1-1)B_2\lambda^{n_1-n_2-1}$$
$$+\left(-\frac{1}{\gamma}-n_2+1\right)\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{1}{(K+\delta)\gamma}\lambda^{-\frac{1}{\gamma}-n_2}<0,$$

then, the function $\mathscr{E}(\lambda)$ is monotone decreasing in $(\hat{y},\tilde{y}]$. $\phi''(\tilde{y})=\lambda^{n_2-2}\mathscr{E}(\tilde{y})>0$ is clear, which implies that $\mathscr{E}(\tilde{y})>0$. Then, we have $\mathscr{E}(\lambda)>\mathscr{E}(\tilde{y})>0$. Hence, for any $\hat{y}<\lambda\leqslant\tilde{y}$, $\phi''(\lambda)>0$ holds.

For $\bar{y}<\lambda\leqslant\hat{y}$,

$$\phi''(\lambda)=n_2(n_2-1)A_1\lambda^{n_2-2}+n_1(n_1-1)A_2\lambda^{n_1-2}$$
$$+\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{1}{(K+\delta)\gamma}\lambda^{-\frac{1}{\gamma}-1}>0,$$

since $A_1>0$ described in Assumption 3 and $A_2>0$ stated in Remark 5.

For $0<\lambda\leqslant\bar{y}$,

$$\phi''(\lambda)=\frac{1}{K\gamma}\lambda^{-\frac{1}{\gamma}-1}>0.$$

Hence, we conclude that $\phi(\lambda)$ is strictly convex for any $\lambda>0$, generating that $\phi'(\lambda)$ is monotone increasing in $(0,\infty)$.

Taking the first-order conditions to (31) with respect to λ generates the equation (32). Since

$$\lim_{\lambda\to0+}\phi'(\lambda)=-\infty,\ \lim_{\lambda\to\infty}\phi'(\lambda)=\frac{I_1-\epsilon-\omega}{r+\delta},$$

$-\phi'(\lambda)\in(-\frac{I_1-\epsilon-\omega}{r+\delta},\infty)$ is monotone decreasing in $(0,\infty)$. This means that there is a unique $\lambda^*>0$ such that

$$x=-\phi'\left(\lambda^*\right),$$

for given $x>-\frac{I_1-\epsilon-\omega}{r+\delta}$.

Appendix 6. Proof of Theorem 3.7

Consequently, the optimal job state process Ξ_t^* can be obtained by (15). The optimal consumption c_t^* is determined using (14). By applying Itô's formula to the optimal wealth level in (33), we find that for $\bar{y}<y_t^{\lambda_3^*}\leqslant\hat{y}$ and $t<\tau_R^*$, the following holds:

$$\mathrm{d}X_t^*=\left[-n_2(n_2-1)A_1\left(y_t^{\lambda_3^*}\right)^{n_2-2}-n_1(n_1-1)A_2\left(y_t^{\lambda_3^*}\right)^{n_1-2}\right.$$
$$\left.-\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{1}{\gamma(K+\delta)}\left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}-1}\right]\mathrm{d}y_t^{\lambda_3^*}$$
$$+\frac{1}{2}\left(-n_2(n_2-1)(n_2-2)A_1\left(y_t^{\lambda_3^*}\right)^{n_2-3}\right.$$
$$-n_1(n_1-1)(n_1-2)A_2\left(y_t^{\lambda_3^*}\right)^{n_1-3}$$
$$\left.+\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{1+\gamma}{\gamma^2(K+\delta)}\left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}-2}\right)\mathrm{d}\left\langle y^{\lambda_3^*}\right\rangle_t$$
$$=\left[\left(-n_2(n_2-1)A_1\left(y_t^{\lambda_3^*}\right)^{n_2-1}-n_1(n_1-1)A_2\left(y_t^{\lambda_3^*}\right)^{n_1-1}\right.\right.$$
$$\left.-\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{1}{\gamma(K+\delta)}\left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}}\right)(\beta-r)$$
$$+\frac{1}{2}\theta^2\left(-n_2(n_2-1)(n_2-2)A_1\left(y_t^{\lambda_3^*}\right)^{n_2-1}\right.$$
$$-n_1(n_1-1)(n_1-2)A_2\left(y_t^{\lambda_3^*}\right)^{n_1-1}$$
$$\left.\left.+\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{1+\gamma}{\gamma^2(K+\delta)}\left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}}\right)\right]\mathrm{d}t$$
$$-\theta\left[-n_2(n_2-1)A_1\left(y_t^{\lambda_3^*}\right)^{n_2-1}\right.$$
$$-n_1(n_1-1)A_2\left(y_t^{\lambda_3^*}\right)^{n_1-1}-\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)$$
$$\left.\frac{1}{\gamma(K+\delta)}\left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}}\right]\mathrm{d}B_t. \tag{A11}$$

Next, we compare the diffusion coefficients in (A11) and (1) to determine the optimal portfolio in the range $\bar{y}<y_t^{\lambda_3^*}\leqslant\hat{y}$. Similarly, the optimal portfolio process for other specified intervals, as defined in Theorem 3.7, can be derived.

Appendix 7. Proof of Lemma 3.8

By setting $B_2=0$ in (21), the solution for $v^{IR}(y)$ is obtained as follows:

$$v(y)=v^{IR}(y)=$$
$$\begin{cases}\tilde{C}y^{n_1}+\frac{\delta\gamma}{(1-\gamma)K(K+\delta)}y^{\frac{\gamma-1}{\gamma}}+\frac{I_1-\epsilon-\omega}{r+\delta}y\\ +\frac{L_1^{\gamma-\gamma^*}\omega^{1-\gamma}}{(\beta+\delta)(1-\gamma)}, & \text{for } \tilde{y}<y,\\ \tilde{B}_1y^{n_2}+\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\frac{\gamma}{(1-\gamma)(K+\delta)}y^{\frac{\gamma-1}{\gamma}}\\ +\frac{I_1-\epsilon}{r+\delta}y, & \text{for } 0<y\leqslant\tilde{y}.\end{cases}$$

Making use of the methodology described in Section 3.1 yields solutions to the value function and control processes in Lemma 3.8.

Appendix 8. Proof of Lemma 3.9

By considering $A_2=0$ in (21), we obtain the following solution for $v^{IRJS}(y)$

$$v(y)=v^{IRJS}(y)=$$
$$\begin{cases}\hat{C}y^{n_1}+\frac{\delta\gamma}{(1-\gamma)K(K+\delta)}y^{\frac{\gamma-1}{\gamma}}\\ +\frac{I_1-\epsilon-\omega}{r+\delta}y+\frac{L_1^{\gamma-\gamma^*}\omega^{1-\gamma}}{(\beta+\delta)(1-\gamma)}, & \text{for } \tilde{y}<y\\ \hat{B}_1y^{n_2}+\hat{B}_2y^{n_1}+\left(L_1^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\\ \frac{\gamma}{(1-\gamma)(K+\delta)}y^{\frac{\gamma-1}{\gamma}}+\frac{I_1-\epsilon}{r+\delta}y, & \text{for } \hat{y}<y\leqslant\tilde{y},\\ \hat{A}_1y^{n_2}+\left(L_2^{\frac{\gamma-\gamma^*}{\gamma}}+\frac{\delta}{K}\right)\\ \frac{\gamma}{(1-\gamma)(K+\delta)}y^{\frac{\gamma-1}{\gamma}}+\frac{I_2-\epsilon}{r+\delta}y, & \text{for } 0<y\leqslant\hat{y}.\end{cases}$$

By applying the methodology outlined in Section 3.1, we derive solutions for the value function and control processes, as presented in Lemma 3.9.

Appendix 9. Proof of Lemma 3.10

$$\begin{aligned}
&\tilde{x} - \tilde{x}^{IRJS} \\
&= -n_1 C\tilde{y}^{n_1-1} + \frac{\delta}{K(K+\delta)}\tilde{y}^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta} \\
&\quad - \left[-n_1\hat{C}\tilde{y}^{n_1-1} + \frac{\delta}{K(K+\delta)}\tilde{y}^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta}\right] \\
&= -n_1 A_2 \tilde{y}^{n_1-1} > 0, \\
&\tilde{x}^{IRJS} - \tilde{x}^{IR} \\
&= -n_1\hat{C}\tilde{y}^{n_1-1} + \frac{\delta}{K(K+\delta)}\tilde{y}^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta} \\
&\quad - \left[-n_1\tilde{C}\tilde{y}^{n_1-1} + \frac{\delta}{K(K+\delta)}\tilde{y}^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon-\omega}{r+\delta}\right] \\
&= -n_1\hat{B}_2\tilde{y}^{n_1-1} > 0,
\end{aligned}$$

since $A_2 > 0$ and $\hat{B}_2 > 0$.

Appendix 10. Proof of Proposition 4.1

By taking the differential of (27) with respect to ω, we can verify that:

$$\frac{\partial\bar{y}}{\partial\omega} = \frac{\frac{\frac{n_1}{(\beta+\delta)(1-\gamma)} - \frac{\gamma n_1+1-\gamma}{(K+\delta)(1-\gamma)} + \frac{1-n_1}{r+\delta}}{n_1-n_2}}{-\frac{1}{\gamma(1-\gamma)}\frac{\gamma n_1+1-\gamma}{n_1-n_2}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right]\bar{y}^{-\frac{1}{\gamma}-1} - A_1(n_2-1)\bar{y}^{n_2-2}} > 0.$$

Subsequently, from (30), the following differential is derived:

$$\frac{\partial\bar{x}}{\partial\omega} = -\frac{1}{\gamma K}\bar{y}^{-\frac{1}{\gamma}-1}\frac{\partial\bar{y}}{\partial\omega} < 0.$$

Appendix 11. Proof of Proposition 4.2

Taking the differential of (27) with respect to ϵ, we obtain

$$\frac{\partial\bar{y}}{\partial\epsilon} = \frac{-\frac{1-n_1}{(r+\delta)(n_2-n_1)}}{-\frac{1}{\gamma(1-\gamma)}\frac{\gamma n_1+1-\gamma}{n_1-n_2}\left[\frac{1}{K} - \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\right]\bar{y}^{-\frac{1}{\gamma}-1} - A_1(n_2-1)\bar{y}^{n_2-2}} > 0,$$

which results directly in $\frac{\partial\bar{x}}{\partial\epsilon} < 0$.

Appendix 12. Proof of Proposition 4.3

For $X_t^* \in (\tilde{x}, \hat{x}^{IRJS})$, $y_t^{\hat{\lambda}_2^*}$ is substituted into (33) to generate the following function:

$$\begin{aligned}
X_t^*|_{y_t^{\lambda_2^*}=y_t^{\hat{\lambda}_2^*}} &= -n_2 B_1\left(y_t^{\hat{\lambda}_2^*}\right)^{n_2-1} - n_1 B_2\left(y_t^{\hat{\lambda}_2^*}\right)^{n_1-1} \\
&\quad + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\left(y_t^{\hat{\lambda}_2^*}\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon}{r+\delta}. \qquad \text{(A12)}
\end{aligned}$$

Based on (35), we rewrite (A12) as follows:

$$\begin{aligned}
X_t^*|_{y_t^{\lambda_2^*}=y_t^{\hat{\lambda}_2^*}} &= -n_2 B_1\left(y_t^{\hat{\lambda}_2^*}\right)^{n_2-1} - n_1 B_2\left(y_t^{\hat{\lambda}_2^*}\right)^{n_1-1} \\
&\quad + n_2\hat{B}_1\left(y_t^{\hat{\lambda}_2^*}\right)^{n_2-1} + n_1\hat{B}_2\left(y_t^{\hat{\lambda}_2^*}\right)^{n_1-1} + X_t^* \\
&= -n_1 A_2\left(y_t^{\hat{\lambda}_2^*}\right)^{n_1-1} + X_t^* \\
&> X_t^*,
\end{aligned}$$

because $\hat{B}_1 = B_1 < 0$ and $\hat{B}_2 - B_2 = -A_2 < 0$. Thus, we conclude that

$$y_t^{\lambda_2^*} > y_t^{\hat{\lambda}_2^*},$$

because X_t^* is a monotonically decreasing function; hence, we can obtain

$$c_t^{***} = L_1^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\hat{\lambda}_2^*}\right)^{-\frac{1}{\gamma}} > L_1^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}} = c_t^*.$$

Next, we plug $y_t^{\tilde{\lambda}_2^*}$ into (35) to generate the following function:

$$\begin{aligned}
X_t^*|_{y_t^{\hat{\lambda}_2^*}=y_t^{\tilde{\lambda}_2^*}} &= -n_2\hat{B}_1\left(y_t^{\tilde{\lambda}_2^*}\right)^{n_2-1} - n_1\hat{B}_2\left(y_t^{\tilde{\lambda}_2^*}\right)^{n_1-1} \\
&\quad + \left(L_1^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\left(y_t^{\tilde{\lambda}_2^*}\right)^{-\frac{1}{\gamma}} - \frac{I_1-\epsilon}{r+\delta}. \qquad \text{(A13)}
\end{aligned}$$

From (34), equation (A13) can be rewritten as

$$\begin{aligned}
X_t^*|_{y_t^{\hat{\lambda}_2^*}=y_t^{\tilde{\lambda}_2^*}} &= -n_2\hat{B}_1\left(y_t^{\tilde{\lambda}_2^*}\right)^{n_2-1} - n_1\hat{B}_2\left(y_t^{\tilde{\lambda}_2^*}\right)^{n_1-1} \\
&\quad + n_2\tilde{B}_1\left(y_t^{\tilde{\lambda}_2^*}\right)^{n_2-1} + X_t^* \\
&= -n_1\hat{B}_2\left(y_t^{\tilde{\lambda}_2^*}\right)^{n_1-1} + X_t^* \\
&> X_t^*,
\end{aligned}$$

because $\tilde{B}_1 = \hat{B}_1 = B_1$ and $\hat{B}_2 > 0$. Then, we conclude that

$$\begin{aligned}
&y_t^{\hat{\lambda}_2^*} > y_t^{\tilde{\lambda}_2^*}, \\
&c_t^{**} = L_1^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\tilde{\lambda}_2^*}\right)^{-\frac{1}{\gamma}} > L_1^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\hat{\lambda}_2^*}\right)^{-\frac{1}{\gamma}} = c_t^{***}.
\end{aligned}$$

For $X_t^* \in (\hat{x}, \bar{x})$, substituting $y_t^{\hat{\lambda}_3^*}$ into (33) yields:

$$\begin{aligned}
X_t^*|_{y_t^{\lambda_3^*}=y_t^{\hat{\lambda}_3^*}} &= -n_2 A_1\left(y_t^{\hat{\lambda}_3^*}\right)^{n_2-1} - n_1 A_2\left(y_t^{\hat{\lambda}_3^*}\right)^{n_1-1} \\
&\quad + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right)\frac{1}{K+\delta}\left(y_t^{\hat{\lambda}_3^*}\right)^{-\frac{1}{\gamma}} - \frac{I_2-\epsilon}{r+\delta}. \qquad \text{(A14)}
\end{aligned}$$

From (36) and (A14) can be reorganized as follows:

$$\begin{aligned}
X_t^*|_{y_t^{\lambda_3^*}=y_t^{\hat{\lambda}_3^*}} &= -n_2 A_1\left(y_t^{\hat{\lambda}_3^*}\right)^{n_2-1} - n_1 A_2\left(y_t^{\hat{\lambda}_3^*}\right)^{n_1-1} \\
&\quad + n_2\hat{A}_1\left(y_t^{\hat{\lambda}_3^*}\right)^{n_2-1} + X_t^* \\
&= -n_1 A_2\left(y_t^{\hat{\lambda}_3^*}\right)^{n_1-1} + X_t^* \\
&> X_t^*,
\end{aligned}$$

which implies

$$y_t^{\lambda_3^*} > y_t^{\hat{\lambda}_3^*}, \quad c_t^{***} > c_t^*.$$

Appendix 13. Proof of Proposition 4.4

For $X_t^* \in (\hat{x}, \bar{x})$, we equate the wealth level process from (33) and (36) as follows:

$$
\begin{aligned}
&- n_2 A_1 \left(y_t^{\lambda_3^*}\right)^{n_2-1} - n_1 A_2 \left(y_t^{\lambda_3^*}\right)^{n_1-1} \\
&\quad + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \frac{1}{K+\delta} \left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}} - \frac{I_2-\epsilon}{r+\delta} \\
&\quad = -n_2 \hat{A}_1 \left(y_t^{\hat{\lambda}_3^*}\right)^{n_2-1} + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \frac{1}{K+\delta} \left(y_t^{\hat{\lambda}_3^*}\right)^{-\frac{1}{\gamma}} \\
&\quad - \frac{I_2-\epsilon}{r+\delta}. \qquad \text{(A15)}
\end{aligned}
$$

Multiplying both sides of (A15) by $n_1 - 1$ and reorganizing it generates:

$$
\begin{aligned}
&- \frac{\sigma}{\theta}\pi_t^* + n_2 A_1 \left(y_t^{\lambda_3^*}\right)^{n_2-1} (n_2 - n_1) \\
&\quad + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \frac{1}{K+\delta} \left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}} \left(n_1 + \frac{1-\gamma}{\gamma}\right) \\
&\quad = -\frac{\sigma}{\theta}\pi_t^{***} + n_2 \hat{A}_1 \left(y_t^{\hat{\lambda}_3^*}\right)^{n_2-1} (n_2 - n_1) \\
&\qquad + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \frac{1}{K+\delta} \left(y_t^{\hat{\lambda}_3^*}\right)^{-\frac{1}{\gamma}} \left(n_1 + \frac{1-\gamma}{\gamma}\right).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\pi_t^* - \pi_t^{***} \\
&\quad = -\frac{\theta}{\sigma} \Bigg[n_2 A_1 (n_2 - n_1) \left(\left(y_t^{\hat{\lambda}_3^*}\right)^{n_2-1} - \left(y_t^{\lambda_3^*}\right)^{n_2-1} \right) \\
&\qquad + \left(L_2^{\frac{\gamma-\gamma^*}{\gamma}} + \frac{\delta}{K}\right) \frac{1}{K+\delta} \left(n_1 + \frac{1-\gamma}{\gamma}\right) \\
&\qquad \left(\left(y_t^{\hat{\lambda}_3^*}\right)^{-\frac{1}{\gamma}} - \left(y_t^{\lambda_3^*}\right)^{-\frac{1}{\gamma}} \right) \Bigg] > 0,
\end{aligned}
$$

because $\hat{A}_1 = A_1$ and $n_1 + \frac{1-\gamma}{\gamma} < 0$.

Economic Modelling 153 (2025) 107307

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling

Job switching and bequest motives in an optimal consumption–investment model under inflation and mortality risks

Qi Li [a,b], Yong Hyun Shin [c,*], Ji-Hun Yoon [a,d]

[a] Department of Mathematics, Pusan National University, Busan 46241, Republic of Korea
[b] Department of Mathematics & Information Technology, The Education University of Hong Kong, Tai Po, N.T., Hong Kong
[c] Department of Mathematics & Research Institute of Natural Sciences, Sookmyung Women's University, Seoul 04310, Republic of Korea
[d] Institute of Mathematical Science, Pusan National University, Busan 46241, Republic of Korea

ARTICLE INFO

Dataset link: Replication Package for "Job Switching and Bequest Motives in an Optimal Consumption--Investment Model under Inflation and Mortality Risks" (Original data)

Keywords:
Life insurance
Reversible job switching
Inflation risk
Bequest motive

ABSTRACT

This study presents a unified model for analyzing optimal consumption, investment, and life-insurance decisions under the combined effects of inflation risk, mortality risk, and reversible job-switching opportunities. In the model, individuals dynamically choose between two job states with distinct income–leisure trade-offs while allocating wealth across a complete financial market that includes inflation-linked bonds, stocks, and bank deposits. Using the martingale approach, we derive closed-form solutions and conduct comparative statics to examine how life expectancy, bequest motives, inflation, and leisure preferences jointly influence individual financial decision-making. The results show that mortality risk and bequest motives have opposing effects on the wealth threshold for job switching, offering new theoretical insights beyond prior studies, which typically consider these factors in isolation. Furthermore, we demonstrate that job-switching flexibility can significantly enhance consumption and investment outcomes, especially in inflationary environments. The model is further extended to incorporate stochastic labor income and job-switching costs, capturing more realistic labor market frictions and income uncertainty. Although these extensions do not yield closed-form solutions, the theoretical frameworks provide a foundation for future research using partial differential equation methods. Overall, the proposed framework delivers strong explanatory power and meaningful policy implications for retirement planning, life-insurance design, inflation-hedging strategies, and the development of flexible labor market policies.

1. Introduction

Since the seminal work of Merton (1969), the optimal consumption and investment problem has been a cornerstone of financial economics. Subsequent studies have expanded this foundational framework to include various real-world factors influencing individual financial decisions (e.g., mortality risk, labor market dynamics, and inflation). Mortality risk affects consumption and investment behavior through its impact on life expectancy, while labor supply flexibility shapes the trade-off between income and leisure, thereby influencing financial choices. Inflation alters the real value of both assets and income. Despite the importance of these factors, most prior research has examined them in isolation, with limited focus on their combined effects on intertemporal decision-making (e.g., Shim and Shin (2014); Kwak and Lim (2014); Zhang et al. (2023); Park et al. (2023)). In reality, individuals are exposed to these risks simultaneously. This study addresses this gap by analyzing an optimal consumption–investment problem that jointly incorporates job-switching opportunities, life insurance, and bequest motives within an inflation-adjusted market environment.

Labor supply flexibility generally involves two key dimensions: the option to switch jobs and the opportunity to retire from the workforce. Shim and Shin (2014) developed an optimal consumption and investment model that includes job-switching opportunities, showing that such flexibility can either raise or lower the consumption rate compared with models with a fixed job state. Building on this work, Lee et al. (2019) investigated how borrowing constraints affect optimal consumption and portfolio decisions when job choice flexibility is present. In a related study, Jeon and Oh (2023) extended the analysis to a consumption–leisure–investment–retirement model by integrating life insurance into the decision-making process. Separately, Zhang et al. (2023) and Li et al. (2025) explored asset pricing models that incorporate job-switching behavior under inflation

* Corresponding author.
E-mail address: yhshin@sookmyung.ac.kr (Y.H. Shin).

https://doi.org/10.1016/j.econmod.2025.107307
Received 12 April 2024; Received in revised form 9 September 2025; Accepted 10 September 2025
Available online 27 September 2025
0264-9993/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Table 1
A review of influential literature concerning inflation risk, job switching, and life insurance.

	Life-insurance payment	Job-switching option	Inflation risk management
Shim and Shin (2014), Lee et al. (2019)	×	✓	×
Kwak and Lim (2014), Park et al. (2023)	✓	×	✓
Jeon and Oh (2023)	✓	✓	×
Zhang et al. (2023), Li et al. (2025)	×	✓	✓
This paper	✓	✓	✓

risk. However, these studies often fail to capture key interactions: some omit inflation risk (Jeon and Oh, 2023), others exclude life insurance (Zhang et al. (2023); Li et al. (2025)), and several omit both elements (Shim and Shin (2014); Lee et al. (2019)). This lack of integration limits a full understanding of how these elements together influence individual financial strategies.

A major source of market incompleteness in traditional models is the presence of stochastic labor income, which limits individuals' ability to smooth consumption and manage financial risk (e.g., Duffie et al. (1997); Cocco et al. (2005); Wang et al. (2016); Ahn et al. (2019); Bensoussan and Park (2024)). In contrast, the present study assumes that labor income is fixed within each job state, preserving the structure of a complete market. Although stochastic income adds further constraints to market participation, our focus is on how individuals optimally allocate resources when job-switching opportunities, mortality risk, and inflation-hedging instruments are present. This modeling choice enables a clearer examination of financial decision-making in environments where inflation risk can be effectively managed using available hedging strategies.

Meanwhile, a widely adopted approach to mitigate mortality risk is through life insurance. Richard (1975) was among the first to integrate life insurance into an optimal consumption and portfolio selection model in a continuous-time framework. Subsequent research has built on this foundation. For example, Wei et al. (2020) studied the case of a wage-earning couple, Chen et al. (2022) examined the effects of an incomplete life-insurance market, and Li et al. (2023) introduced a drawdown constraint on consumption. In a fixed-inflation environment, Kwak and Lim (2014) analyzed a dynamic asset pricing model with life insurance, finding that optimal strategies are significantly influenced by the inflation rate. Building on this framework, other studies have broadened the analysis (e.g., Han and Hung (2017); Pan and Xiao (2017); Park et al. (2023); Chen et al. (2024); Li et al. (2025)). While all these works address inflation risk, Han and Hung (2017) and Park et al. (2023) specifically examined its effect on optimal life-insurance premiums, consumption, and investment decisions. Pan and Xiao (2017) focused on the effects of stochastic inflation and interest rates, whereas Chen et al. (2024) and Li et al. (2025) assessed the influence of labor supply flexibility on optimal financial strategies.

Motivated by these developments, we propose a unified framework that jointly incorporates inflation risk, labor supply flexibility, and life insurance within a complete market. Specifically, we introduce a class of inflation-linked index bonds whose returns are tied to the price level and adjust with inflation fluctuations. Investing in these instruments allows individuals to hedge inflation risk and maintain the real value of their wealth effectively. The financial market also includes a risk-free asset (e.g., bank deposits) with a fixed nominal interest rate and a risky asset (e.g., stocks). We assume that the price dynamics of both the stock and the inflation-linked index bond are driven by a common Brownian motion. This assumption ensures that all market uncertainty arises from a single stochastic source, thereby satisfying the conditions for market completeness in the mathematical sense.

Although the core analysis assumes costless job switching and fixed labor income, we also examine extended models that incorporate either job-switching costs or stochastic labor income. These extensions are developed with precise mathematical formulations and demonstrate how frictions in occupational mobility or income uncertainty can affect optimal decision-making. Collectively, they provide a basis for future research on the interaction between labor market frictions and individual financial strategies under risk.

Building on the recent work of Li et al. (2025), this study combines the modeling frameworks of Kwak and Lim (2014) on inflation risk, Shim and Shin (2014) on job-switching behavior, and Jeon and Oh (2023) on life insurance to analyze the joint effects of labor supply flexibility and inflation risk on optimal life-insurance premiums, consumption, and diversified investment strategies. A detailed comparative summary of the key contributions cited in this paper is presented in Table 1.

This paper makes several key contributions to the literature. First, our comparative analysis demonstrates that the availability of job-switching opportunities increases the consumption rate, life-insurance premiums, and investments in both inflation-linked index bonds and stocks, while reducing allocations to the risk-free bank account. Second, we find that a higher-leisure rate associated with job state J_1 leads individuals to lower their life-insurance premiums and bank account investments. At the same time, they raise the job-switching threshold, increase consumption, and allocate more resources to inflation-linked bonds and equities at the switching point. In contrast, when the leisure rate in job state J_2 increases, the effects are reversed. Third, our results indicate that in a low-inflation environment, individuals may borrow through the bond market to invest in higher-yielding financial assets, whereas in a high-inflation environment, the bank account effectively serves as a borrowing vehicle. As the expected inflation rate rises, the wealth threshold for job switching increases, while life-insurance premiums and stock investments decrease. Finally, we show that both the hazard rate and the strength of the bequest motive significantly influence individuals' job-switching decisions—an effect identified here for the first time.

To the best of our knowledge, the interaction between mortality risk and bequest motives has received limited attention in the context of job-switching behavior. While foundational studies—such as the pioneering work of Hurd (1989), which examined the effect of mortality on consumption and wealth trajectories; Bernheim (1991), which found that a substantial portion of aggregate savings is driven by the desire to leave bequests; Gan et al. (2015), which analyzed the influence of subjective mortality risk on saving, consumption, and bequest behavior; and Davidoff et al. (2005) and Lockwood (2012), which highlighted the importance of bequest motives in annuitization decisions—have made significant contributions, their primary focus has been on wealth accumulation and broader financial decisions, rather than on occupational choices.

Few studies have explicitly investigated the connection between mortality risk and bequest motives in the context of job-switching behavior. To the best of our knowledge, this paper is the first to formally show how both the expectation of death and the strength of bequest motives jointly affect the timing and attractiveness of switching between job states. By modeling mortality risk through the hazard rate, δ, we capture the effect of a finite lifespan on occupational transitions. Similarly, by incorporating the bequest motive parameter, k_M, into the bequest function, we describe individuals' inheritance preferences and their implications for labor supply decisions. From a theoretical perspective, our model extends

the dynamic labor supply literature by demonstrating that job-switching behavior is influenced not only by the traditional income–leisure trade-off but also by forward-looking considerations related to mortality and intergenerational wealth transfer.

The remainder of this paper is organized as follows. Section 2 presents the theoretical framework, and Section 3 derives closed-form solutions for the optimal strategies using the martingale approach. Section 4 provides a detailed numerical analysis, examining the economic implications of labor supply flexibility and the effects of the leisure rate, expected inflation rate, hazard rate, and bequest motive on individuals' optimal decisions. Section 5.1 offers an extended discussion of stochastic labor income, while Section 5.2 develops a model incorporating job-switching costs. Section 6 then concludes with final remarks.

2. The model

2.1. Financial, insurance, and labor markets

We consider a continuous-time economy that integrates the financial, insurance, and labor markets, with a particular emphasis on managing inflation risk and labor supply flexibility. In the financial market, investors have access to diversified instruments, including a bank account, S_t^0, an inflation-linked index bond, D_t, and a stock, S_t. The one-dimensional Brownian motions, B_t and Z_t, which are mutually independent, are defined on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$. Our model adopts the inflation rate framework developed by Kwak and Lim (2014), Zhang et al. (2023), and Li et al. (2025), all of which use the Consumer Price Index (CPI) as the measure of inflation. The CPI dynamics are expressed as

$$\frac{\mathrm{d}L_t}{L_t} = \mu_L \mathrm{d}t + \sigma_L \left(\rho \mathrm{d}B_t + \sqrt{1-\rho^2}\mathrm{d}Z_t\right).$$

This framework specifies the joint processes of an inflation-linked bond tied to the CPI and a stock as follows:

$$\begin{aligned}\frac{\mathrm{d}D_t}{D_t} &= r\mathrm{d}t + \frac{\mathrm{d}L_t}{L_t} = (r+\mu_L)\mathrm{d}t + \sigma_L\left(\rho \mathrm{d}B_t + \sqrt{1-\rho^2}\mathrm{d}Z_t\right),\\ \frac{\mathrm{d}S_t}{S_t} &= \mu \mathrm{d}t + \sigma \mathrm{d}B_t,\end{aligned}$$

where $\mu > 0$ and $\sigma > 0$ represent the expected return and volatility of the stock, respectively. Similarly, $\mu_L > 0$ and $\sigma_L > 0$ denote the expected inflation rate and its volatility. The parameter, $r > 0$, represents the real interest rate, and the correlation between the financial and bond markets is given by $\rho \in (-1, 1)$. In addition, the bank account process accrues at a nominal interest rate, $R > 0$, given by

$$\frac{\mathrm{d}S_t^0}{S_t^0} = R\mathrm{d}t.$$

We assume that individuals can switch freely between two job states, denoted by J_1 and J_2, and that this choice is reversible. These job-switching opportunities enable individuals to experience different levels of leisure and earn varying amounts of income. Our approach follows the job choice framework described in Shim and Shin (2014), Lee et al. (2019), Zhang et al. (2023), and Li et al. (2025). In job state J_i, an individual earns labor income Y_i and enjoys leisure rate l_i, where $i \in \{1, 2\}$, subject to conditions

$$Y_1 > Y_2 > 0 \quad \text{and} \quad 0 < l_1 < l_2.$$

As noted in Park et al. (2023) and Jeon and Oh (2023), individuals face mortality risk, which results in uncertain lifespans. Purchasing life insurance is a common strategy to mitigate this risk. We adopt the mortality rate framework from these references to define the probability density function, $f(t)$, and cumulative distribution function, $F(t)$, for the random death time, τ, as follows:

$$f(t) = \delta e^{-\delta t} \text{ and } F(t) = \int_0^t \delta e^{-\delta s} ds = 1 - e^{-\delta t},$$

where $\delta > 0$ is a fixed hazard rate. The death time, τ, is assumed to be independent of Brownian motions B_t and Z_t, implying that mortality risk is independent of financial market risks. To hedge against mortality risk, an individual pays a nominal life-insurance premium, $\tilde{p}_t$, at time t, and, in a frictionless market, receives the fair value of the nominal insurance benefit at death, τ, given by $\tilde{p}_\tau/\delta$. If $\tilde{p}_t$ is negative, the individual instead receives a nominal pension annuity at time t at a rate of $-\tilde{p}_t$ (see Dybvig and Liu (2010)). Accordingly, the total nominal value of the individual's bequest, $\tilde{M}_\tau$, at death τ is given by

$$\tilde{M}_\tau = \tilde{X}_\tau + \frac{\tilde{p}_\tau}{\delta}, \tag{2.1}$$

with the nominal wealth level, $\tilde{X}_\tau$.

2.2. Inflation-adjusted market

The control variables consist of current job state Ξ_t, nominal consumption rate $\tilde{c}_t$, nominal life-insurance premium $\tilde{p}_t$, nominal bank investment $\tilde{\pi}_{0,t}$, inflation-linked index bond $\tilde{\pi}_{1,t}$, stock $\tilde{\pi}_{2,t}$, and nominal labor income $\tilde{Y}_i$ for $i \in \{1, 2\}$. The $\mathcal{F}_t$-adapted process, Ξ_t, is assumed to take values in $\{J_1, J_2\}$. Under the nominal interest rate, the wealth process evolves according to the following dynamic:

$$\mathrm{d}\tilde{X}_t = \tilde{\pi}_{0,t}\frac{\mathrm{d}S_t^0}{S_t^0} + \tilde{\pi}_{1,t}\frac{\mathrm{d}D_t}{D_t} + \tilde{\pi}_{2,t}\frac{\mathrm{d}S_t}{S_t} - \tilde{c}_t\mathrm{d}t - \tilde{p}_t\mathrm{d}t + \left(\tilde{Y}_1\mathbf{1}_{\{\Xi_t=J_1\}} + \tilde{Y}_2\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t.$$

The terms involving $\tilde{\pi}_{2,t}$ and $\tilde{\pi}_{1,t}$ represent hedging strategies against market risk and inflation uncertainty, respectively, reflecting the role of portfolio diversification in preserving purchasing power over time. By discounting each nominal variable by the CPI, we obtain corresponding real

values: consumption c_t, life-insurance premium p_t, bank account investment $\pi_{0,t}$, inflation-linked index bond investment $\pi_{1,t}$, stock investment $\pi_{2,t}$, and labor income Y_i. This transformation yields the following real wealth process[1]:

$$\begin{aligned} \mathrm{d}X_t = &\left[rX_t - c_t - p_t + Y_1\mathbf{1}_{\{\Xi_t=J_1\}} + Y_2\mathbf{1}_{\{\Xi_t=J_2\}}\right]\mathrm{d}t + \pi_{2,t}\left\{\left(\mu - r - \mu_L + \sigma_L^2 - \rho\sigma\sigma_L\right)\mathrm{d}t + (\sigma - \rho\sigma_L)\mathrm{d}B_t - \sigma_L\sqrt{1-\rho^2}\mathrm{d}Z_t\right\} \\ &+ \pi_{0,t}\left\{\left(R - r - \mu_L + \sigma_L^2\right)\mathrm{d}t - \rho\sigma_L\mathrm{d}B_t - \sigma_L\sqrt{1-\rho^2}\mathrm{d}Z_t\right\}, \quad \text{with } X_0 = x, \end{aligned} \tag{2.2}$$

where $X_t > -\frac{Y_1}{\delta+r}$ and $\pi_{1,t} = X_t - \pi_{0,t} - \pi_{2,t}$. In addition, $c_t \geqslant 0$, $\boldsymbol{\pi}_t = (\pi_{0,t}, \pi_{1,t}, \pi_{2,t})^\top$, and p_t are $\mathcal{F}_t$-measurable processes such that

$$\int_0^\infty \left(c_t + \| \left(\pi_{0,t}, \pi_{1,t}, \pi_{2,t}\right)^\top \|^2 + p_t\right)\mathrm{d}t < \infty.$$

Note that the price dynamics of both the stock and the inflation-linked index bond are driven by a common Brownian motion, with their correlation captured by the parameter, ρ. This modeling choice ensures that all market uncertainty stems from a single source of randomness, thereby establishing market completeness in a mathematically rigorous manner. In this inflation-adjusted economic setting, individuals can construct dynamic investment portfolios by allocating wealth among a risk-free bank account, an inflation-linked index bond, and a risky stock. Such portfolios allow for effective hedging against all tradable sources of risk, reinforcing the market completeness implied by our model.

2.3. Preferences

We employ a Cobb–Douglas utility function to represent individuals' preferences over consumption and leisure:

$$u_i\left(c_t, l_i\right) \equiv \frac{1}{\alpha}\frac{\left(c_t^\alpha l_i^{1-\alpha}\right)^{1-\gamma^*}}{1-\gamma^*}, \text{ with } i \in \{1,2\}.$$

The utility function for an individual's total legacy at the time of death, τ, is defined as $u_d(M_\tau) \equiv \frac{(k_M M_\tau)^{\alpha(1-\gamma^*)}}{\alpha(1-\gamma^*)}$. Here, $\alpha \in (0,1)$ represents the elasticity of preference for consumption, $\gamma^* \in (0,1)\cup(1,\infty)$ is the coefficient of constant relative risk aversion, and k_M measures the strength of the bequest motive. Defining $\gamma \equiv 1 - \alpha(1-\gamma^*) > 0$, the utility functions can be rewritten as $u_i\left(c_t, l_i\right) = l_i^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}$ and $u_d(M_\tau) = \frac{(k_M M_\tau)^{1-\gamma}}{1-\gamma}$. These utility functions illustrate that individuals derive utility not only from consumption but also from leisure, which is determined by the job state, and from leaving a bequest. The term, $l_i^{\gamma-\gamma^*}$, modifies the marginal utility of consumption according to the level of leisure in each job, thereby linking labor supply decisions to intertemporal consumption choices. Consequently, the overall utility that individuals derive from consumption, leisure, and legacy can be expressed as follows:

$$\begin{aligned} \tilde{V}\left(x; c, p, \boldsymbol{\pi}, \Xi\right) &\equiv \mathbb{E}\left[\int_0^\tau e^{-\beta t}\left(l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_1\}} + l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t + e^{-\beta\tau}\frac{(k_M M_\tau)^{1-\gamma}}{1-\gamma}\right] \\ &= \mathbb{E}\left[\int_0^\infty f(s)\int_0^s e^{-\beta t}\left(l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_1\}} + l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t\mathrm{d}s + \int_0^\infty f(t)e^{-\beta t}\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}\mathrm{d}t\right] \\ &= \mathbb{E}\left[\int_0^\infty e^{-\beta t}\left(l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_1\}} + l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_2\}}\right)\int_t^\infty f(s)\mathrm{d}s\mathrm{d}t + \int_0^\infty f(t)e^{-\beta t}\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}\mathrm{d}t\right] \\ &= \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_1\}} + l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_2\}} + \delta\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}\right)\mathrm{d}t\right], \end{aligned}$$

where $\beta > 0$ is the subjective discount rate.

Problem 1 (*Primal Problem*)**.** The individual's objective is to maximize the expected utility derived from consumption, leisure, and total legacy, given by

$$V(x) = \sup_{(c,p,\boldsymbol{\pi},\Xi)\in\mathcal{A}(x)} \tilde{V}\left(x; c, p, \boldsymbol{\pi}, \Xi\right), \tag{2.3}$$

subject to the budget constraint, (2.2), where $\mathcal{A}(x)$ denotes the set of all admissible strategies.

3. Verification theorem and optimal strategies

In this section, we apply the martingale approach as developed in Karatzas et al. (1991), Shim and Shin (2014), Lee et al. (2019), and Li et al. (2025). We then reformulate the dynamic optimization Problem 1 in its dual form, a transformation that is instrumental in enabling an explicit characterization of the optimal policies.

Denoting the exponential martingale process as

$$\mathcal{Z}_t \equiv \exp\left\{-\left(\frac{1}{2}\theta_1^2 + \frac{1}{2}\theta_2^2\right)t - \theta_1 B_t - \theta_2 Z_t\right\},$$

we define the equivalent martingale measure $\mathbb{Q}$ by $\mathbb{Q}(A) = E\left[\mathcal{Z}_T\mathbf{1}_A\right]$, for all $A \in \mathcal{F}_T$, and introduce the stochastic discount factor as follows:

$$H_t \equiv \exp\{-(r+\delta)t\}\cdot\mathcal{Z}_t.$$

[1] $X_t \equiv \frac{\tilde{X}_t}{L_t}$, $c_t \equiv \frac{\tilde{c}_t}{L_t}$, $M_t \equiv \frac{\tilde{M}_t}{L_t}$, $p_t \equiv \frac{\tilde{p}_t}{L_t}$, $\pi_{0,t} \equiv \frac{\tilde{\pi}_{0,t}}{L_t}$, $\pi_{1,t} \equiv \frac{\tilde{\pi}_{1,t}}{L_t}$, $\pi_{2,t} \equiv \frac{\tilde{\pi}_{2,t}}{L_t}$, $Y_i \equiv \frac{\tilde{Y}_i}{L_t}$.

Following the approaches of Kwak and Lim (2014), Zhang et al. (2023), and Li et al. (2025), the market prices of risk associated with the stock and the index bond are respectively given by

$$\theta_1 = \frac{\mu - R - \rho\sigma\sigma_L}{\sigma} \quad \text{and} \quad \theta_2 = \frac{r - R + \mu_L - \sigma_L^2}{\sigma_L\sqrt{1-\rho^2}} - \frac{\rho}{\sqrt{1-\rho^2}}\theta_1.$$

Under the risk-neutral probability measure, $\mathbb{Q}$, the wealth process in (2.2) then evolves to

$$\mathrm{d}X_t = \left[(r+\delta)X_t - c_t - \delta M_t + Y_1\mathbf{1}_{\{\Xi_t=J_1\}} + Y_2\mathbf{1}_{\{\Xi_t=J_2\}}\right]\mathrm{d}t - \left[\left(\pi_{0,t}+\pi_{2,t}\right)\rho\sigma_L - \pi_{2,t}\sigma\right]\mathrm{d}\tilde{B}_t - \left(\pi_{0,t}+\pi_{2,t}\right)\sigma_L\sqrt{1-\rho^2}\mathrm{d}\tilde{Z}_t, \tag{3.1}$$

where $\tilde{B}_t \equiv B_t + \theta_1 t$ and $\tilde{Z}_t \equiv Z_t + \theta_2 t$.

By applying Itô's formula to the discounted wealth, we derive the following relation:

$$\begin{aligned}\mathrm{d}\left(e^{-(r+\delta)t}X_t\right) &= e^{-(r+\delta)t}\left[-c_t - \delta M_t + Y_1\mathbf{1}_{\{\Xi_t=J_1\}} + Y_2\mathbf{1}_{\{\Xi_t=J_2\}}\right]\mathrm{d}t \\ &\quad - e^{-(r+\delta)t}\left[\left(\pi_{0,t}+\pi_{2,t}\right)\rho\sigma_L - \pi_{2,t}\sigma\right]\mathrm{d}\tilde{B}_t - e^{-(r+\delta)t}\left(\pi_{0,t}+\pi_{2,t}\right)\sigma_L\sqrt{1-\rho^2}\mathrm{d}\tilde{Z}_t,\end{aligned} \tag{3.2}$$

for $t \in [0,T]$. The above expression can be equivalently restated in its integrated form as follows:

$$\begin{aligned}&e^{-(r+\delta)T}X_T + \int_0^T e^{-(r+\delta)t}\left(c_t + \delta M_t - Y_1\mathbf{1}_{\{\Xi_t=J_1\}} - Y_2\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t \\ &= x - \int_0^T e^{-(r+\delta)t}\left[\left(\pi_{0,t}+\pi_{2,t}\right)\rho\sigma_L - \pi_{2,t}\sigma\right]\mathrm{d}\tilde{B}_t - \int_0^T e^{-(r+\delta)t}\left(\pi_{0,t}+\pi_{2,t}\right)\sigma_L\sqrt{1-\rho^2}\mathrm{d}\tilde{Z}_t.\end{aligned} \tag{3.3}$$

For any feasible selection, $(c,p,\boldsymbol{\pi},\Xi) \in \mathcal{A}(x)$, the integrals on the right-hand side of (3.3) are continuous $\mathbb{Q}$-local martingales, while the term on the left-hand side is bounded from below. By Fatou's lemma, these integrals inherit the supermartingale property. Applying the optional sampling theorem yields

$$\tilde{\mathbb{E}}\left[e^{-(r+\delta)T}X_T + \int_0^T e^{-(r+\delta)t}\left(c_t + \delta M_t - Y_1\mathbf{1}_{\{\Xi_t=J_1\}} - Y_2\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t\right] \leqslant x,$$

where $\tilde{\mathbb{E}}[\cdot]$ denotes expectation under $\mathbb{Q}$. By applying Bayes' rule to convert back to the physical measure, $\mathbb{P}$, and taking the limit as $T \to \infty$, we obtain the following static budget constraint:

$$\mathbb{E}\left[\int_0^\infty H_t\left(c_t + \delta M_t - Y_1\mathbf{1}_{\{\Xi_t=J_1\}} - Y_2\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t\right] \leqslant x. \tag{3.4}$$

To explicitly derive the dual problem, we start from the primal optimization problem, which is subject to the static budget constraint of Eq. (3.4). Introducing a Lagrange multiplier, $\lambda > 0$, we embed the constraint into the objective function to construct the corresponding Lagrangian functional:

$$\begin{aligned}\mathcal{L} \equiv\ &\mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_1\}} + l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_2\}} + \delta\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}\right)\mathrm{d}t\right] \\ &- \lambda\cdot\mathbb{E}\left[\int_0^\infty H_t\left(c_t + \delta M_t - Y_1\mathbf{1}_{\{\Xi_t=J_1\}} - Y_2\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t\right] + \lambda x.\end{aligned} \tag{3.5}$$

Letting $y_t \equiv \lambda e^{(\beta+\delta)t}H_t$ with the differential form

$$\frac{\mathrm{d}y_t}{y_t} = (\beta - r)\mathrm{d}t - \theta_1\mathrm{d}B_t - \theta_2\mathrm{d}Z_t,$$

Eq. (3.5) simplifies to:

$$\mathcal{L} = \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left\{l_i^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma} - c_t y_t + Y_i y_t + \delta\left(\frac{(k_M M_t)^{1-\gamma}}{1-\gamma} - M_t y_t\right)\right\}\mathbf{1}_{\{\Xi_t=J_i\}}\mathrm{d}t\right] + \lambda x. \tag{3.6}$$

To derive the dual formulation, we maximize the Lagrangian pointwise with respect to (c_t, M_t) for each t, using the following Legendre–Fenchel transform:

$$\tilde{u}_i(y_t) \equiv \sup_{c_t>0}\left(l_i^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma} - c_t y_t\right) + \delta\cdot\sup_{M_t>0}\left(\frac{(k_M M_t)^{1-\gamma}}{1-\gamma} - M_t y_t\right), \ \text{for}\, i \in \{1,2\}. \tag{3.7}$$

Taking the first-order conditions of (3.7), we obtain

$$l_i^{\gamma-\gamma^*}c_t^{-\gamma} - y_t = 0, \quad k_M^{1-\gamma}M_t^{-\gamma} - y_t = 0.$$

Additionally, the above equalities allow us to derive the consumption rate and bequest as follows:

$$\hat{c}\left(y_t\right) = \begin{cases} l_1^{\frac{\gamma-\gamma^*}{\gamma}}\, y_t^{-\frac{1}{\gamma}}, \text{for}\, \Xi_t = J_1, \\ l_2^{\frac{\gamma-\gamma^*}{\gamma}}\, y_t^{-\frac{1}{\gamma}}, \text{for}\, \Xi_t = J_2, \end{cases} \tag{3.8}$$

$$\hat{M}\left(y_t\right) = k_M^{\frac{1-\gamma}{\gamma}}\, y_t^{-\frac{1}{\gamma}}. \tag{3.9}$$

Substituting (3.8) and (3.9) into (3.7), the convex conjugate utility function is rewritten as

$$\tilde{u}_i(y_t) = \frac{\gamma}{1-\gamma}\left(l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)y_t^{-\frac{1-\gamma}{\gamma}}. \tag{3.10}$$

Let us define a function $\mathcal{U}(\cdot)$ as follows:

$$\mathcal{U}_i(y_t) \equiv \frac{\gamma}{1-\gamma}\left(l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right) y_t^{-\frac{1-\gamma}{\gamma}} + Y_i y_t, \text{ for } i \in \{1,2\}.$$

Substituting $\mathcal{U}_i(y_t)$ back into the Lagrangian in (3.6), the Legendre–Fenchel transform of the utility function yields the following inequality:

$$\mathcal{L} \leqslant \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(\mathcal{U}_1(y_t)\mathbf{1}_{\{\Xi_t=J_1\}} + \mathcal{U}_2(y_t)\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t\right] + \lambda x. \tag{3.11}$$

This formulation is consistent with the convex duality framework developed by Karatzas and Wang (2000).

In addition, the functions, $\mathcal{U}_1(y_t)$ and $\mathcal{U}_2(y_t)$, satisfy the following ordering relation:

$$\begin{cases} \mathcal{U}_2(y_t) \geqslant \mathcal{U}_1(y_t), & \text{for } 0 < y_t \leqslant \hat{y}, \\ \mathcal{U}_2(y_t) < \mathcal{U}_1(y_t), & \text{for } y_t > \hat{y}, \end{cases} \tag{3.12}$$

where $\hat{y}$ is defined by the switching boundary at which the dual utilities from both job states coincide, that is, $\mathcal{U}_1(\hat{y}) = \mathcal{U}_2(\hat{y})$. This threshold is given by:

$$\hat{y} \equiv \left(\frac{\gamma}{1-\gamma}\frac{l_2^{\frac{\gamma-\gamma^*}{\gamma}} - l_1^{\frac{\gamma-\gamma^*}{\gamma}}}{Y_1 - Y_2}\right)^\gamma > 0.$$

Based on the structure in , the job state process is characterized as

$$\hat{\Xi}(y_t) = \begin{cases} J_2, & \text{for } 0 < y_t \leqslant \hat{y}, \\ J_1, & \text{for } y_t > \hat{y}. \end{cases} \tag{3.13}$$

For any admissible tuple, $(c, p, \boldsymbol{\pi}, \Xi) \in \mathcal{A}(x)$, the following inequality holds as a direct consequence of (3.11):

$$\tilde{V}(x; c, p, \boldsymbol{\pi}, \Xi) \leqslant \inf_{\lambda>0} \mathcal{L} \leqslant \inf_{\lambda>0}\left(\mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(\mathcal{U}_1(y_t)\mathbf{1}_{\{y_t>\hat{y}\}} + \mathcal{U}_2(y_t)\mathbf{1}_{\{0<y_t\leqslant\hat{y}\}}\right)\mathrm{d}t\right] + \lambda x\right),$$

which yields the weak-duality relation,

$$V(x) \leqslant \inf_{\lambda>0}\left[\phi(\lambda) + \lambda x\right], \tag{3.14}$$

and $\phi(\lambda)$ is the dual value function expressed by

$$\phi(\lambda) \equiv \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(\mathcal{U}_1(y_t)\mathbf{1}_{\{y_t>\hat{y}\}} + \mathcal{U}_2(y_t)\mathbf{1}_{\{0<y_t\leqslant\hat{y}\}}\right)\mathrm{d}t\right]. \tag{3.15}$$

Equality in (3.14) implies that the value function, $V(x)$, as defined in (2.3), is attainable. This result is formally established in Verification Theorem 1.

Assumption 1. The following inequality is assumed to hold universally, without additional justification:

$$K \equiv r + \delta + \frac{\beta - r}{\gamma} + \frac{\gamma - 1}{2\gamma^2}\left(\theta_1^2 + \theta_2^2\right) > 0.$$

Furthermore, the following assumption enables us to derive a closed-form solution to the optimization problem subject to boundary conditions:

Assumption 2. We assume that the dual value function satisfies the smooth-pasting condition at the switching boundary, $\hat{y}$; that is, it is continuously differentiable at $y = \hat{y}$.

To derive the solution to (3.15), we define a generalized version

$$\psi(t, y) \equiv \mathbb{E}^{y_t=y}\left[\int_t^\infty \mathrm{e}^{-(\beta+\delta)s}\left(\mathcal{U}_1(y_s)\mathbf{1}_{\{y_s>\hat{y}\}} + \mathcal{U}_2(y_s)\mathbf{1}_{\{0<y_s\leqslant\hat{y}\}}\right)\mathrm{d}s\right]. \tag{3.16}$$

Proposition 1. *Suppose that a function, $v(y)$, is identified by*

$$v(y) = \begin{cases} Ay^{n_1} + \frac{\gamma}{(1-\gamma)K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right) y^{-\frac{1-\gamma}{\gamma}} + \frac{Y_1}{r+\delta} y, & \text{for } y > \hat{y}, \\ By^{n_2} + \frac{\gamma}{(1-\gamma)K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right) y^{-\frac{1-\gamma}{\gamma}} + \frac{Y_2}{r+\delta} y, & \text{for } 0 < y \leqslant \hat{y}, \end{cases} \tag{3.17}$$

where

$$A = \frac{Y_1 - Y_2}{n_2 - n_1}\left(\frac{n_2 + \frac{1-\gamma}{\gamma}}{K} + \frac{1 - n_2}{r+\delta}\right)\hat{y}^{1-n_1} > 0, \tag{3.18}$$

$$B = \frac{Y_1 - Y_2}{n_2 - n_1}\left(\frac{n_1 + \frac{1-\gamma}{\gamma}}{K} + \frac{1 - n_1}{r+\delta}\right)\hat{y}^{1-n_2} > 0, \tag{3.19}$$

where $n_1 < 0$ and $n_2 > 1$ are two roots of

$$g(n) \equiv \frac{1}{2}\left(\theta_1^2 + \theta_2^2\right) n^2 + \left(\beta - r - \frac{1}{2}\left(\theta_1^2 + \theta_2^2\right)\right) n - (\beta + \delta) = 0.$$

Subsequently, $\psi(t, y) = e^{-(\beta+\delta)t} v(y)$ is obtained.

Proof. See Appendix A. □

In the region where $y > \hat{y}$, the marginal value of wealth is relatively low, indicating a preference for the high-income job, J_1. Conversely, when $y \leqslant \hat{y}$, the marginal utility of wealth rises, prompting the individual to optimally switch to the low-income, high-leisure job J_2. The function, $v(y)$, as defined in Proposition 1, represents the expected lifetime utility under optimal consumption, life insurance, and investment strategies, conditional on the individual's current job state.

The two segments of $v(y)$ are designed to satisfy both the value-matching and smooth-pasting conditions at the switching boundary $\hat{y}$, thereby ensuring continuity and differentiability. The constants, A and B, along with exponents $n_1 < 0$ and $n_2 > 1$, capture the intertemporal trade-offs shaped by individual preferences, the hazard rate, and prevailing market risks. This formulation enables a closed-form characterization of the dual value function and plays a central role in deriving the optimal strategies in Theorem 1 as well as in recovering the primal value function in Remark 1 via inverse duality.

We clarify that $\phi(\lambda) = \psi(0, \lambda) = v(\lambda)$.

Proposition 2. *The dual value function $\phi(\lambda)$ satisfies the following key properties:*

1. *The function is strictly convex for $\lambda > 0$, thereby implying that its derivative $\phi'(\lambda)$ is strictly increasing over the interval $(0, \infty)$.*
2. *The derivative of ϕ with respect to λ satisfies the following relation:*

$$\lim_{\lambda \to 0+} \phi'(\lambda) = -\infty, \quad \lim_{\lambda \to \infty} \phi'(\lambda) = \frac{Y_1}{r + \delta}.$$

Consequently, the minimization problem described in Eq. (3.21) admits a unique solution, λ^*, that satisfies the following condition:

$$x = -\phi'(\lambda^*). \tag{3.20}$$

Proof. See Appendix B. □

Theorem 1 (*Verification Theorem*).

1. *For any given $x \in \left(-\frac{Y_1}{\delta + r}, \infty\right)$, the value function, $V(x)$, satisfies the following relation with $\phi(\lambda)$:*

$$V(x) = \inf_{\lambda > 0} \left[\phi(\lambda) + \lambda x\right]. \tag{3.21}$$

2. *The optimal job state, $\left(\Xi_t^*\right)_{t=0}^{\tau}$, and consumption rate, $\left(c_t^*\right)_{t=0}^{\tau}$, are characterized as*

$$\Xi_t^* = \begin{cases} J_1, & \text{for } y_t^{\lambda_1^*} > \hat{y}, \\ J_2, & \text{for } 0 < y_t^{\lambda_2^*} \leqslant \hat{y}, \end{cases} \quad c_t^* = \begin{cases} l_1^{\frac{\gamma - \gamma^*}{\gamma}} \left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}}, & \text{for } y_t^{\lambda_1^*} > \hat{y}, \\ l_2^{\frac{\gamma - \gamma^*}{\gamma}} \left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}, & \text{for } 0 < y_t^{\lambda_2^*} \leqslant \hat{y}. \end{cases}$$

The optimal life-insurance premium, $(p_t^*)_{t=0}^{\tau}$, is given by

$$p_t^* = \begin{cases} \delta \left[k_M^{\frac{1-\gamma}{\gamma}} \left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}} - X_t^* \right], & \text{for } y_t^{\lambda_1^*} > \hat{y}, \\ \delta \left[k_M^{\frac{1-\gamma}{\gamma}} \left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}} - X_t^* \right], & \text{for } 0 < y_t^{\lambda_2^*} \leqslant \hat{y}. \end{cases}$$

The optimal portfolio processes, $\left(\pi_{0,t}^*\right)_{t=0}^{\tau}$, $\left(\pi_{1,t}^*\right)_{t=0}^{\tau}$, and $\left(\pi_{2,t}^*\right)_{t=0}^{\tau}$, are determined to be

$$\pi_{0,t}^* = \begin{cases} -\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\right) F_1\left(y_t^{\lambda_1^*}\right), & \text{for } y_t^{\lambda_1^*} > \hat{y}, \\ -\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\right) F_2\left(y_t^{\lambda_2^*}\right), & \text{for } 0 < y_t^{\lambda_2^*} \leqslant \hat{y}, \end{cases}$$

$$\pi_{1,t}^* = \begin{cases} X_t^* + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}} F_1\left(y_t^{\lambda_1^*}\right), & \text{for } y_t^{\lambda_1^*} > \hat{y}, \\ X_t^* + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}} F_2\left(y_t^{\lambda_2^*}\right), & \text{for } 0 < y_t^{\lambda_2^*} \leqslant \hat{y}, \end{cases}$$

$$\pi_{2,t}^* = \begin{cases} \left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}}\right) F_1\left(y_t^{\lambda_1^*}\right), & \text{for } y_t^{\lambda_1^*} > \hat{y}, \\ \left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}}\right) F_2\left(y_t^{\lambda_2^*}\right), & \text{for } 0 < y_t^{\lambda_2^*} \leqslant \hat{y}. \end{cases}$$

Here,

$$F_1\left(y_t^{\lambda_1^*}\right) = n_1\left(n_1-1\right)A\left(y_t^{\lambda_1^*}\right)^{n_1-1} + \frac{1}{\gamma K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}},$$

$$F_2\left(y_t^{\lambda_2^*}\right) = n_2\left(n_2-1\right)B\left(y_t^{\lambda_2^*}\right)^{n_2-1} + \frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}},$$

where each of $y_t^{\lambda_1^*}$ *and* $y_t^{\lambda_2^*}$ *satisfies the corresponding algebraic relation presented below:*

$$X_t^*\left(y_t^{\lambda_i^*}\right) = \begin{cases} -n_1 A\left(y_t^{\lambda_1^*}\right)^{n_1-1} + \frac{1}{K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}} - \frac{Y_1}{r+\delta}, & \text{for } y_t^{\lambda_1^*} > \hat{y}, \\ -n_2 B\left(y_t^{\lambda_2^*}\right)^{n_2-1} + \frac{1}{K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}} - \frac{Y_2}{r+\delta}, & \text{for } 0 < y_t^{\lambda_2^*} \leqslant \hat{y}. \end{cases} \tag{3.22}$$

In addition, the job-switching boundary of wealth level is derived as

$$\hat{x} = -n_2 B\hat{y}^{n_2-1} + \frac{1}{K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\hat{y}^{-\frac{1}{\gamma}} - \frac{Y_2}{r+\delta}.$$

Proof. We begin with proving part (1). Observe that the identities $\phi(\lambda) = \psi(0,\lambda) = \upsilon(\lambda)$ hold. Now, we define the function $\mathcal{H}(\cdot)$ by

$$\mathcal{H}(\lambda) = \mathbb{E}\left[\int_0^\infty H_t\left(\hat{c}(y_t) + \delta\hat{M}(y_t) - Y_1\mathbf{1}_{\{\hat{\Xi}(y_t)=J_1\}} - Y_2\mathbf{1}_{\{\hat{\Xi}(y_t)=J_2\}}\right)\mathrm{d}t\right],$$

where, $\hat{c}(y_t)$, $\hat{M}(y_t)$, and $\hat{\Xi}(y_t)$ are given by (3.8), (3.9), and (3.13), respectively.

When equality holds in the budget constraint (3.4), there exists $\hat{\boldsymbol{\pi}}$ such that $(\hat{c}(y_t^*), \hat{p}(y_t^*), \hat{\boldsymbol{\pi}}(y_t^*), \hat{\Xi}(y_t^*)) \in \mathcal{A}(\mathcal{H}(\lambda^*))$. The term $\hat{p}(y_t^*)$ is determined by Eqs. (2.1) and (3.9). Furthermore, the associated wealth process, X_t can be expressed as follows:

$$X_t^{x;\hat{c},\hat{p},\hat{\boldsymbol{\pi}},\hat{\Xi}} = \mathbb{E}_t\left[\int_t^\infty \frac{H_s}{H_t}\left(\hat{c}(y_s^*) + \delta\hat{M}(y_s^*) - Y_1\mathbf{1}_{\{\hat{\Xi}(y_s^*)=J_1\}} - Y_2\mathbf{1}_{\{\hat{\Xi}(y_s^*)=J_2\}}\right)\mathrm{d}s\right].$$

In particular, we have

$$\mathcal{H}(\lambda^*) = \mathbb{E}\left[\int_0^\infty H_t\left(\hat{c}(y_t^*) + \delta\hat{M}(y_t^*) - Y_1\mathbf{1}_{\{\hat{\Xi}(y_t^*)=J_1\}} - Y_2\mathbf{1}_{\{\hat{\Xi}(y_t^*)=J_2\}}\right)\mathrm{d}t\right].$$

Building on the above, we obtain

$$\begin{aligned}\lambda^*\mathcal{H}(\lambda^*) &= \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}y_t^*\left(\hat{c}(y_t^*) + \delta\hat{M}(y_t^*) - Y_1\mathbf{1}_{\{\hat{\Xi}(y_t^*)=J_1\}} - Y_2\mathbf{1}_{\{\hat{\Xi}(y_t^*)=J_2\}}\right)\mathrm{d}t\right] \\ &= \tilde{V}\left(\mathcal{H}(\lambda^*); \hat{c}(y_t^*), \hat{p}(y_t^*), \hat{\boldsymbol{\pi}}(y_t^*), \hat{\Xi}(y_t^*)\right) - \phi(\lambda^*),\end{aligned}$$

which leads to

$$\begin{aligned}\phi(\lambda^*) + \lambda^*\mathcal{H}(\lambda^*) &= \tilde{V}\left(\mathcal{H}(\lambda^*); \hat{c}(y_t^*), \hat{p}(y_t^*), \hat{\boldsymbol{\pi}}(y_t^*), \hat{\Xi}(y_t^*)\right) \\ &\leqslant \sup_{(c,p,\boldsymbol{\pi},\Xi)\in\mathcal{A}(\mathcal{H}(\lambda^*))}\tilde{V}\left(\mathcal{H}(\lambda^*); c, p, \boldsymbol{\pi}, \Xi\right) \\ &\leqslant \inf_{\lambda>0}\left[\phi(\lambda) + \lambda\mathcal{H}(\lambda^*)\right] \\ &\leqslant \phi(\lambda^*) + \lambda^*\mathcal{H}(\lambda^*).\end{aligned}$$

Hence, we have

$$\begin{aligned}\tilde{V}\left(\mathcal{H}(\lambda^*); \hat{c}(y_t^*), \hat{p}(y_t^*), \hat{\boldsymbol{\pi}}(y_t^*), \hat{\Xi}(y_t^*)\right) &= \sup_{(c,p,\boldsymbol{\pi},\Xi)\in\mathcal{A}(\mathcal{H}(\lambda^*))}\tilde{V}\left(\mathcal{H}(\lambda^*); c, \boldsymbol{\pi}, \Xi, \tau\right) \\ &= \inf_{\lambda>0}\left[\phi(\lambda) + \lambda\mathcal{H}(\lambda^*)\right] \\ &= \phi(\lambda^*) + \lambda^*\mathcal{H}(\lambda^*).\end{aligned}$$

Proposition 2 yields the identity, $x = -\phi'(\lambda^*)$. Given the convexity of $\phi(\lambda)$, it follows that

$$x = \mathcal{H}(\lambda^*) = -\phi'(\lambda^*).$$

Consequently, we arrive at

$$\tilde{V}\left(x; \hat{c}(y_t^*), \hat{p}(y_t^*), \hat{\boldsymbol{\pi}}(y_t^*), \hat{\Xi}(y_t^*)\right) = \sup_{(c,p,\boldsymbol{\pi},\Xi)\in\mathcal{A}(x)}\tilde{V}\left(x; c, p, \boldsymbol{\pi}, \Xi\right) = V(x) = \inf_{\lambda>0}\left[\phi(\lambda) + \lambda x\right],$$

which further implies that $\left(\hat{c}(y_t^*), \hat{p}(y_t^*), \hat{\boldsymbol{\pi}}(y_t^*), \hat{\Xi}(y_t^*)\right)$ constitutes the optimal control tuple. That is,

$$(c_t^*, p_t^*, \boldsymbol{\pi}_t^*, \Xi_t^*) = \left(\hat{c}(y_t^*), \hat{p}(y_t^*), \hat{\boldsymbol{\pi}}(y_t^*), \hat{\Xi}(y_t^*)\right).$$

We now proceed to prove part (2). The optimal job state is obtained directly from Eq. (3.13), while the optimal consumption strategy is derived from Eq. (3.8). The life-insurance premium is determined by Eq. (2.1), in conjunction with Eq. (3.9). Applying Itô's formula to the optimal wealth process in Eq. (3.22), for the case where $y_t^{\lambda_1^*} > \hat{y}$, we obtain the following relation:

$$\begin{aligned}
\mathrm{d}X_t^* &= \left[-n_1(n_1-1)A\left(y_t^{\lambda_1^*}\right)^{n_1-2} - \frac{1}{\gamma K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}-1}\right]\mathrm{d}y_t^{\lambda_1^*} \\
&\quad + \frac{1}{2}\left(-n_1(n_1-1)(n_1-2)A\left(y_t^{\lambda_1^*}\right)^{n_1-3} + \frac{1+\gamma}{\gamma^2 K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}-2}\right)\mathrm{d}\left\langle y^{\lambda_1^*}\right\rangle_t \\
&= \left[\left(-n_1(n_1-1)A\left(y_t^{\lambda_1^*}\right)^{n_1-1} - \frac{1}{\gamma K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}}\right)(\beta - r)\right. \\
&\quad \left. + \frac{1}{2}\left(\theta_1^2+\theta_2^2\right)\left(-n_1(n_1-1)(n_1-2)A\left(y_t^{\lambda_1^*}\right)^{n_1-1} + \frac{1+\gamma}{\gamma^2 K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}}\right)\right]\mathrm{d}t \\
&\quad + \left(-\theta_1\mathrm{d}B_t - \theta_2\mathrm{d}Z_t\right)\left(-n_1(n_1-1)A\left(y_t^{\lambda_1^*}\right)^{n_1-1} - \frac{1}{\gamma K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_1^*}\right)^{-\frac{1}{\gamma}}\right).
\end{aligned} \tag{3.23}$$

By comparing the coefficients of $\mathrm{d}B_t$ and $\mathrm{d}Z_t$ in Eq. (3.23) with those in Eq. (2.2), we obtain the following system of equations:

$$\begin{cases}
\pi_{2,t}^*(\sigma - \rho\sigma_L) - \pi_{0,t}^*\rho\sigma_L = F_1\left(y_t^{\lambda_1^*}\right)\theta_1, \\
-\pi_{2,t}^*\sigma_L\sqrt{1-\rho^2} - \pi_{0,t}\sigma_L\sqrt{1-\rho^2} = F_1\left(y_t^{\lambda_1^*}\right)\theta_2, \\
\pi_{1,t}^* = X_t^* - \pi_{0,t}^* - \pi_{2,t}^*.
\end{cases}$$

Subsequently, in the case where $y_t^{\lambda_1^*} > \hat{y}$, the optimal allocations to the bank account, the inflation-linked index bond, and the stock can be explicitly derived.

Applying the same procedure, we can derive the optimal investment strategies corresponding to the case, $0 < y_t^{\lambda_2^*} \leqslant \hat{y}$; that is,

$$\begin{aligned}
\mathrm{d}X_t^* &= \left[-n_2(n_2-1)B\left(y_t^{\lambda_2^*}\right)^{n_2-2} - \frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}-1}\right]\mathrm{d}y_t^{\lambda_2^*} \\
&\quad + \frac{1}{2}\left(-n_2(n_2-1)(n_2-2)B\left(y_t^{\lambda_2^*}\right)^{n_2-3} + \frac{1+\gamma}{\gamma^2 K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}-2}\right)\mathrm{d}\left\langle y^{\lambda_2^*}\right\rangle_t \\
&= \left[\left(-n_2(n_2-1)B\left(y_t^{\lambda_2^*}\right)^{n_2-1} - \frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}\right)(\beta - r)\right. \\
&\quad \left. + \frac{1}{2}\left(\theta_1^2+\theta_2^2\right)\left(-n_2(n_2-1)(n_2-2)B\left(y_t^{\lambda_2^*}\right)^{n_2-1} + \frac{1+\gamma}{\gamma^2 K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}\right)\right]\mathrm{d}t \\
&\quad + \left(-\theta_1\mathrm{d}B_t - \theta_2\mathrm{d}Z_t\right)\left(-n_2(n_2-1)B\left(y_t^{\lambda_2^*}\right)^{n_2-1} - \frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}\right).
\end{aligned} \tag{3.24}$$

Accordingly, matching the coefficients of $\mathrm{d}B_t$ and $\mathrm{d}Z_t$ in Eqs. (2.2) and (3.24), and solving the resulting system, yields the optimal asset allocations for the case $0 < y_t^{\lambda_2^*} \leqslant \hat{y}$.

$$\begin{cases}
\pi_{2,t}^*(\sigma - \rho\sigma_L) - \pi_{0,t}^*\rho\sigma_L = F_2\left(y_t^{\lambda_2^*}\right)\theta_1, \\
-\pi_{2,t}^*\sigma_L\sqrt{1-\rho^2} - \pi_{0,t}\sigma_L\sqrt{1-\rho^2} = F_2\left(y_t^{\lambda_2^*}\right)\theta_2, \\
\pi_{1,t}^* = X_t^* - \pi_{0,t}^* - \pi_{2,t}^*.
\end{cases}$$ □

Remark 1. From (3.21), the value function associated with the consumption, job-switching, and investment optimization problem in (2.3) takes the following form:

$$V(x) = \begin{cases}
A(1-n_1)\left(\lambda_1^*\right)^{n_1} + \frac{1}{(1-\gamma)K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(\lambda_1^*\right)^{-\frac{1-\gamma}{\gamma}}, & \text{for } \lambda_1^* > \hat{y}, \\
B(1-n_2)\left(\lambda_2^*\right)^{n_2} + \frac{1}{(1-\gamma)K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(\lambda_2^*\right)^{-\frac{1-\gamma}{\gamma}}, & \text{for } 0 < \lambda_2^* \leqslant \hat{y},
\end{cases}$$

where λ_1^* and λ_2^* are characterized by the equations:

$$x = -\phi'(\lambda_1^*) = -n_1 A\left(\lambda_1^*\right)^{n_1-1} + \frac{1}{K}\left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(\lambda_1^*\right)^{-\frac{1}{\gamma}} - \frac{Y_1}{r+\delta},$$

$$x = -\phi'(\lambda_2^*) = -n_2 B\left(\lambda_2^*\right)^{n_2-1} + \frac{1}{K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(\lambda_2^*\right)^{-\frac{1}{\gamma}} - \frac{Y_2}{r+\delta}.$$

Furthermore, for any $x \in \left(-\frac{Y_1}{r+\delta}, \hat{x}\right)$, there exists a unique $\lambda_1^* \in (\hat{y}, \infty)$ such that $x = -\phi'(\lambda_1^*)$. Likewise, for any $x \in [\hat{x}, \infty)$, the equation $x = -\phi'(\lambda_2^*)$ admits a unique solution $\lambda_2^* \in (0, \hat{y}]$.

Lemma 1. *The value function constructed in Remark 1 satisfies the following boundary conditions at $\hat{x}$:*

$$V(\hat{x}-) = V(\hat{x}+),\ V'(\hat{x}-) = V'(\hat{x}+),\ \text{and } V''(\hat{x}-) = V''(\hat{x}+).$$

Then, $V(\cdot)$ is C^2-smooth across the switching wealth boundary, $\hat{x}$.

Proof. See Appendix C. □

In job state J_2, $c_t^{J_2}$, $p_t^{J_2}$, $\pi_{0,t}^{J_2}$, $\pi_{1,t}^{J_2}$, and $\pi_{2,t}^{J_2}$ represent the optimal consumption, insurance premium, and the optimal allocations to the bank account, inflation-linked index bond, and stock, respectively. To describe the influence of job-switching opportunities on individual decision-making, we compare the optimal strategies derived in Theorem 1 with those obtained from the benchmark model in Appendix D. The results are presented in the following proposition:

Assumption 3. It is assumed that

$$\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}} > 0.$$

Proposition 3. *According to Assumption 3, for*

$$X_t^* \in \left(\max\left(-\frac{Y_2}{r+\delta}, -n_2 B\hat{y}^{n_2-1} + \frac{1}{K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\hat{y}^{-\frac{1}{\gamma}} - \frac{Y_2}{r+\delta}\right), \infty\right),$$

under the fixed job state, J_2, the availability of a switching opportunity to the alternative state, J_1, increases the optimal consumption rate, insurance premium, investment in the inflation-linked index bond, and stock investment. That is,

$$c_t^{J_2} < c_t^*, \quad p_t^{J_2} < p_t^*, \quad \pi_{1,t}^{J_2} < \pi_{1,t}^*, \quad \pi_{2,t}^{J_2} < \pi_{2,t}^*,$$

while decreasing the optimal investment in the bank account; that is,

$$\pi_{0,t}^{J_2} > \pi_{0,t}^*.$$

Proof. See Appendix E. □

Proposition 3 demonstrates that, under concurrent inflation and mortality risks, individuals with the option of reversible job-switching display markedly different optimal behaviors compared to those permanently in the low-income job state, J_2. In particular, those with the flexibility to move to the high-income job state, J_1, tend to choose higher consumption levels, larger life-insurance premiums, and greater allocations to both inflation-linked bonds and stocks. Proposition 3 provides the formal theoretical justification for this result. The following section supplements this analysis with comparative statics and numerical illustrations to further elucidate the underlying economic implications.

4. Numerical implications

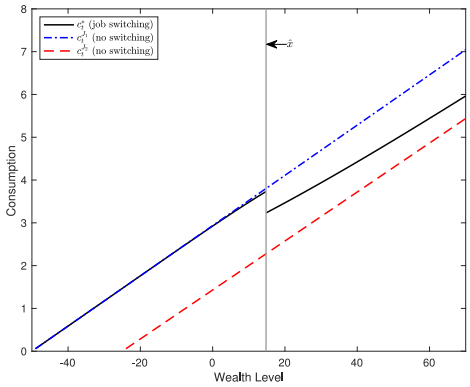
The presence of reversible job-switching opportunities substantially impacts individuals' optimal decisions regarding consumption, investment, and life-insurance premiums. In the first subsection, we compare the optimal strategies derived from our model with those from a benchmark model that excludes job-switching opportunities, as described in Appendix D. In the second subsection, we examine how variations in leisure rates, inflation risk, mortality risk, and bequest motives influence individuals' optimal behavior within the context of job-switching decisions. For illustrative purposes, we set the labor income levels to $Y_1 = 2$ and $Y_2 = 1$. Consistent with the labor–leisure trade-off, the corresponding leisure rates are set to $l_1 = 3$ and $l_2 = 4$. Our baseline parameter configuration is based on values from Kwak and Lim (2014), Park et al. (2023), and Bensoussan and Park (2024)[2]:

$$\delta = 0.02; k_M = 0.03; \beta = 0.07; \gamma^* = 3; r = 0.02; R = 0.04; \alpha = 0.5; \mu_L = 0.035; \sigma_L = 0.06; \rho = 0.5; \mu = 0.08; \sigma = 0.2.$$

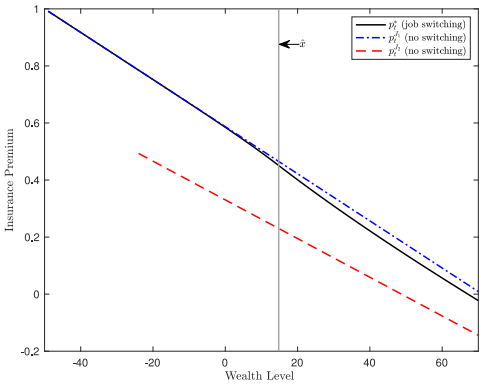
4.1. Impact of reversible job switching on optimal financial decisions

Figs. 1 and 2 illustrate the impact of job-switching opportunities on individuals' optimal strategies. Compared with those who remain fixed in job state J_2, individuals with the flexibility to switch jobs generally exhibit higher consumption rates, larger life-insurance premiums, and greater investments in the inflation-linked index bond and the stock market, while allocating less to the risk-free bank account. This pattern reflects the intuitive notion that access to a higher-income job enables individuals to increase consumption and participate more actively in both insurance and financial markets. Moreover, job-switching opportunities allow individuals to sustain comparable levels of economic performance after transitioning from J_1 to J_2, relative to those who remain permanently in the lower-income state. This underscores the economic importance of labor supply

[2] To satisfy the conditions specified in Assumptions 1 and 3, we make slight adjustments to certain parameters. These modifications are economically justified and remain consistent with realistic market conditions.

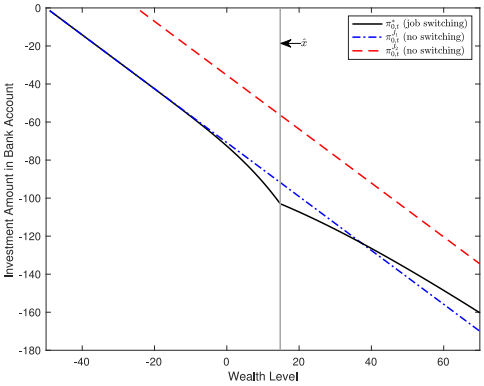


(a) Consumption

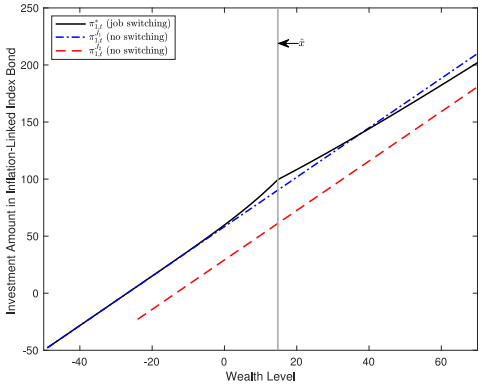


(b) Insurance premium

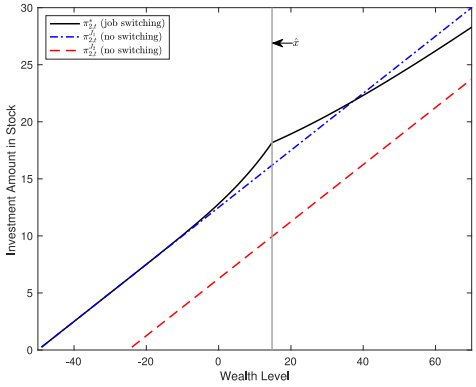
Fig. 1. Comparison analysis with benchmark model.



(a) Investment amount in bank account



(b) Investment amount in inflation-linked index bond



(c) Investment amount in stock

Fig. 2. Comparison analysis with benchmark model.

Table 2
The impacts of l_1 and l_2 on optimal strategies at the time of job switching.

Parameters	$\hat{x}$	$\hat{c}$	$\hat{p}$	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\pi}_2$
$l_1 = 2.0$	−20.6054	1.2071	0.6909	−52.8482	22.9166	9.3261
$l_1 = 2.5$	−8.7185	1.8874	0.6103	−69.6880	48.6716	12.2979
$l_1 = 3.0$	14.7755	3.2321	0.4509	−102.9711	99.5752	18.1714
$l_2 = 4.5$	−2.3823	2.2247	0.5926	−78.6642	62.4000	13.8819
$l_2 = 5.0$	−11.0097	1.7182	0.6638	−66.4421	43.7073	11.7251
$l_2 = 5.5$	−16.2192	1.4124	0.7069	−59.0620	32.4201	10.4227

Table 3
The impacts of μ_L, δ, and k_M on optimal strategies at the time of job switching.

Parameters	$\hat{x}$	$\hat{c}$	$\hat{p}$	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\pi}_2$
$\mu_L = 0.01$	3.7984	3.2321	0.6704	142.4431	−192.0609	53.4162
$\mu_L = 0.02$	13.3314	3.2321	0.4798	66.0832	−99.3987	46.6470
$\mu_L = 0.03$	16.3075	3.2321	0.4203	−47.9497	34.6412	29.6160
$\delta = 0.01$	2.5470	3.2321	0.3477	−109.7425	92.9232	19.3663
$\delta = 0.02$	14.7755	3.2321	0.4509	−102.9711	99.5752	18.1714
$\delta = 0.03$	21.4313	3.2321	0.4767	−97.8116	101.9820	17.2609
$k_M = 0.04$	13.3560	3.2321	0.3793	−100.9602	96.4998	17.8165
$k_M = 0.05$	12.3874	3.2321	0.3304	−99.5880	94.4010	17.5743
$k_M = 0.06$	11.6723	3.2321	0.2943	−98.5750	92.8517	17.3956

flexibility, which positively influences consumption, life-insurance participation, and investment behavior. Our findings on consumption and investment decisions align with those reported by Shim and Shin (2014) and Li et al. (2025), both of whom emphasize the significance of job switching. Similarly, the observed patterns in life-insurance premiums are consistent with the results of Park et al. (2023), particularly in the context of high inflation. Finally, the discontinuity observed at the switching boundary in Fig. 1(a) arises from the abrupt decline in labor income that occurs at the moment of transition between job states.

Remark 2. Additional graphs for the case of $\rho = -0.5$ are provided in Appendix F, where similar patterns are observed.

4.2. Comparative statics analysis for leisure, inflation, mortality, and bequest preferences

We conduct a comprehensive analysis of how key parameter values (i.e., l_1, l_2, μ_L, δ, and k_M) affect the optimal strategies, with particular emphasis on behavior at the point of job switching, as summarized in Tables 2 and 3. First, as shown in Table 2, an increase in leisure rate l_1 results in a decrease in both the optimal life-insurance premium, $\hat{p}$, and the investment in the bank account, $\hat{\pi}_0$, at the switching threshold. By contrast, job-switching wealth boundary $\hat{x}$, consumption rate $\hat{c}$, and investments in inflation-linked bond $\hat{\pi}_1$ and stock $\hat{\pi}_2$ all increase. This outcome reflects the inverse relationship between leisure and labor income: higher-leisure rates typically imply lower labor income due to the greater opportunity cost of time allocated to non-work activities. In response, individuals may choose to remain in the higher-income job state J_1 for a longer period to accumulate additional wealth, thereby enabling higher consumption and investment levels upon transitioning to the lower-income, higher-leisure job state, J_2. Conversely, an increase in the leisure rate, l_2, produces the opposite effects on the optimal strategies, underscoring the asymmetric influence of leisure preferences across different job states.

Furthermore, the expected inflation rate μ_L plays a pivotal role in shaping individuals' investment behavior. In a low-inflation environment, individuals may choose to borrow through the bond market to finance investments in stocks and bank deposits. However, under high inflation, bank deposits effectively function as borrowing instruments, while inflation-linked bonds and equities become the primary vehicles for investment. As the expected inflation rate, μ_L, increases, the job-switching wealth threshold, $\hat{x}$, and investment in inflation-linked bonds, $\hat{\pi}_1$, both increase at the point of job transition. By contrast, the life-insurance premium, $\hat{p}$, the allocation to bank deposits, $\hat{\pi}_0$, and investment in stocks, $\hat{\pi}_2$, all decrease, while the optimal consumption rate, $\hat{c}$, remains largely unchanged. These findings align with previous studies, including Kwak and Lim (2014), Park et al. (2023), Zhang et al. (2023), and Li et al. (2025), and further underscore the crucial role of inflation expectations in shaping optimal financial strategies under flexible labor supply.

In general, in the absence of inflation-hedging instruments such as inflation-linked bonds, a high-inflation environment heightens uncertainty about future purchasing power. This uncertainty often prompts individuals to reduce consumption and equity investments while increasing precautionary savings as a buffer against potential economic shocks or unforeseen expenses. Such savings behavior functions as a protective mechanism to shield individuals from the adverse and unpredictable effects of inflation. However, when inflation-linked bonds are incorporated into the model, individuals are incentivized to reallocate their portfolios in response to high inflation. Specifically, they increase their investment in inflation-linked bonds, whose returns are directly tied to the inflation rate, while reducing their holdings in bank deposits and equities. As inflation rises, the payouts on these bonds adjust accordingly, making them a more effective instrument for preserving purchasing power and mitigating the erosion of real wealth. This strategic shift allows individuals to hedge more effectively against inflation risk.

Moreover, as the expected inflation rate, μ_L, increases, the job-switching wealth threshold, $\hat{x}$, also rises, indicating that individuals tend to delay the transition to lower-income, high-leisure jobs in order to accumulate greater wealth beforehand. Although the optimal consumption level shown in Table 3 remains constant in absolute terms, it effectively declines in relative terms, as higher wealth levels are typically associated with increased consumption. This subtle decline underscores the impact of inflation on real consumption capacity, even when hedging mechanisms are present.

Meanwhile, as shown in Table 3, an increase in either the mortality risk rate δ or the bequest motive parameter k_M leads to higher savings, reflected in increased allocations to the bank account, $\hat{\pi}_0$. However, their effects on the job-switching threshold $\hat{x}$ and life-insurance expenditure

$\hat{p}$ differ: greater mortality risk tends to raise both $\hat{x}$ and $\hat{p}$, whereas a stronger bequest motive increases $\hat{x}$ but decreases $\hat{p}$. These findings carry important practical implications, particularly in the context of aging populations and increasingly complex life-cycle financial planning. As mortality risk rises, individuals tend to adopt more cautious financial behavior: they are more likely to postpone job transitions, accumulate greater wealth during periods of high-income employment, and expand life-insurance coverage to protect their dependents in the event of premature death. This pattern aligns with the "precautionary savings" motive described by Hubbard et al. (1995). Our results provide a theoretical basis for public policy initiatives targeting older workers, including efforts to encourage delayed retirement, promote flexible labor arrangements, and facilitate mid- or late-career transitions. Such policies may enable individuals to manage mortality risk and bequest motives more effectively within the broader context of life-cycle financial planning.

In contrast, a stronger bequest motive encourages individuals to exit high-income employment earlier and transition more quickly into lower-income positions. This behavior is driven by the desire to accumulate savings and reduce life-insurance expenditures, thereby achieving bequest objectives more efficiently. As the marginal utility of leaving a bequest rises, the marginal benefit of remaining in a high-income, low-leisure job declines. Consequently, individuals tend to prioritize increased leisure while preserving wealth for future transfers. This finding is consistent with Lockwood (2012), who demonstrated that strong bequest preferences substantially reduce the demand for annuities and insurance products. It also accords with Davidoff et al. (2005), who showed that individuals with pronounced bequest motives are less likely to annuitize their wealth, preferring instead to retain assets for inheritance purposes. Similarly, Bernheim (1991) found that bequest intentions significantly contribute to higher individual savings rates. From a policy perspective, these insights underscore the importance of accounting for heterogeneity in bequest preferences when designing retirement systems, pension schemes, and life-insurance products. Tailoring financial instruments to reflect individual legacy goals may improve both participation and effectiveness, ultimately enhancing welfare outcomes across diverse demographic groups.

Additionally, as shown in Table 3, the consumption rate at the point of job switching, $\hat{c}$, remains constant across different values of the mortality risk rate δ and the bequest motive parameter, k_M. This constancy reflects a degree of robustness in consumption behavior over the life cycle—an essential feature of consumption smoothing as emphasized in the life-cycle model framework proposed by Zimmerman and Carter (2003). However, although the optimal consumption rate $\hat{c}$ remains unchanged, an increase in the job-switching threshold $\hat{x}$—induced by a higher mortality risk δ—effectively leads to a relative decline in actual consumption levels, as individuals accumulate greater wealth before transitioning to the low-income job. This inverse relationship between mortality risk and consumption aligns with the empirical findings of Hurd (1989), who observed that individuals facing heightened mortality risk tend to reduce consumption as a precautionary measure.

5. Discussion: Towards a more realistic labor–finance framework

In this section, we briefly discuss two extensions of our baseline model that address more realistic features of labor market dynamics. First, we consider the case of stochastic labor income, which introduces additional uncertainty into individuals' financial planning and complicates the structure of market completeness. Second, we incorporate job-switching costs to capture frictions commonly observed in occupational mobility. Although closed-form solutions are not available under these extensions, we present the theoretical frameworks necessary for future analysis using partial differential equations (PDEs) and numerical methods. These discussions highlight the flexibility and adaptability of our modeling framework for addressing a broader range of economic environments.

5.1. Extended analysis of stochastic labor income

Although our primary model assumes fixed labor income, real-world earnings are often subject to significant uncertainty stemming from business cycles, wage volatility, and labor market shocks. To gain deeper insights into how individuals optimize financial decisions under inflation risk, mortality risk, and job-switching opportunities, it is worthwhile to extend the analysis to a framework that incorporates stochastic labor income.

Whereas we do not derive closed-form solutions for the optimal strategies under stochastic income in this subsection, we present the modeling framework to illustrate how wage uncertainty affects the key economic trade-offs underlying consumption, investment, and job-switching behavior. This extension offers meaningful insights and complements our baseline model by showing how the mechanisms identified in the fixed-income setting remain applicable, or are altered, when income risk is explicitly introduced.

To incorporate income uncertainty into the model, let $\tilde{I}_t$ represent the nominal wage rate at time t. Adjusting for inflation using the price index process. L_t, we define the real wage rate as $I_t = \tilde{I}_t / L_t$. Following the approach proposed by Ahn et al. (2019), we model the stochastic dynamics of the wage process as follows:

$$\frac{\mathrm{d}I_t}{I_t} = \mu_I \mathrm{d}t + \sigma_I \mathrm{d}B_t, \text{ with } I_0 = I > 0,$$

where μ_I and σ_I are positive constants representing the drift and volatility of the wage process, respectively. Under this specification, the wealth dynamics in Eq. (2.2) can be reformulated as follows:

$$\begin{aligned}\mathrm{d}X_t = &\left[rX_t - c_t - p_t + I_t\left(\left(\bar{l}-l_1\right)\mathbf{1}_{\{\Xi_t=J_1\}} + \left(\bar{l}-l_2\right)\mathbf{1}_{\{\Xi_t=J_2\}}\right)\right]\mathrm{d}t\\ &+\pi_{2,t}\left\{\left(\mu - r - \mu_L + \sigma_L^2 - \rho\sigma\sigma_L\right)dt + (\sigma - \rho\sigma_L)\mathrm{d}B_t - \sigma_L\sqrt{1-\rho^2}\mathrm{d}Z_t\right\}\\ &+\pi_{0,t}\left\{\left(R - r - \mu_L + \sigma_L^2\right)dt - \rho\sigma_L\mathrm{d}B_t - \sigma_L\sqrt{1-\rho^2}\mathrm{d}Z_t\right\}, \quad \text{with } X_0 = x,\end{aligned}$$

where $\bar{l}$ denotes the maximum leisure rate, and $\bar{l} - l_i$ corresponds to the labor supply rate. Given this, the objective function in Eq. (2.3) can now be expressed as follows:

$$V(x,I) = \sup_{(c,p,\boldsymbol{\pi},\Xi)\in\tilde{\mathcal{A}}(x,I)} \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_1\}} + l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\mathbf{1}_{\{\Xi_t=J_2\}} + \delta\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}\right)\mathrm{d}t\right].$$

By incorporating stochastic labor income, the static budget constraint in Eq. (3.4) is revised as follows:

$$\mathbb{E}\left[\int_0^\infty H_t\left\{c_t + \delta M_t - I_t\left(\left(\bar{l}-l_1\right)\mathbf{1}_{\{\Xi_t=J_1\}} + \left(\bar{l}-l_2\right)\mathbf{1}_{\{\Xi_t=J_2\}}\right)\right\}\mathrm{d}t\right] \leqslant x.$$

Similarly, the Lagrangian function in Eq. (3.6) is updated as follows:

$$\begin{aligned}\mathcal{L} \equiv\ & \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left\{l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}-c_t y_t+I_t\left(\bar{l}-l_1\right)y_t+\delta\left(\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}-M_t y_t\right)\right\}\mathbf{1}_{\{\Xi_t=J_1\}}\mathrm{d}t\right]\\ &+\mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left\{l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}-c_t y_t+I_t\left(\bar{l}-l_2\right)y_t+\delta\left(\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}-M_t y_t\right)\right\}\mathbf{1}_{\{\Xi_t=J_2\}}\mathrm{d}t\right]+\lambda x\\ \leqslant\ & \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(\mathcal{U}_1\left(y_t,I_t\right)\mathbf{1}_{\{\Xi_t=J_1\}}+\mathcal{U}_2\left(y_t,I_t\right)\mathbf{1}_{\{\Xi_t=J_2\}}\right)\mathrm{d}t\right]+\lambda x,\end{aligned}$$

where the function $\mathcal{U}_i$ in Eq. (3.10) is redefined as follows:

$$\begin{aligned}\mathcal{U}_i(y_t,I_t) &\equiv \sup_{c_t>0}\left(l_i^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}-c_t y_t\right)+\delta\sup_{M_t>0}\left(\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}-M_t y_t\right)+I_t\left(\bar{l}-l_i\right)y_t\\ &= \frac{\gamma}{1-\gamma}\left(l_i^{\frac{\gamma-\gamma^*}{\gamma}}+\delta k_M^{\frac{1-\gamma}{\gamma}}\right)y_t^{-\frac{1-\gamma}{\gamma}}+I_t\left(\bar{l}-l_i\right)y_t,\qquad \text{for} i\in\{1,2\}.\end{aligned}$$

The optimal consumption c_t^* and bequest M_t^* retain the same functional forms as in Eqs. (3.8) and (3.9), respectively. Using an approach similar to that employed in deriving Eq. (3.13), the optimal job state is determined as follows:

$$\Xi^*(y_t,I_t)=\begin{cases}J_2, & \text{for } 0<y_t I_t^\gamma\leqslant\hat{W},\\ J_1, & \text{for } y_t I_t^\gamma>\hat{W}.\end{cases}$$

Defining $W_t\equiv y_t I_t^\gamma$ as the new dual variable, we denote the switching boundary under stochastic labor income as $\hat{W}=\hat{y}$. The corresponding dual value function in Eq. (3.15) is then rewritten as follows:

$$\phi(\lambda,I)\equiv\mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(\mathcal{U}_1(y_t,I_t)\mathbf{1}_{\{y_t I_t^\gamma>\hat{W}\}}+\mathcal{U}_2(y_t,I_t)\mathbf{1}_{\{0<y_t I_t^\gamma\leqslant\hat{W}\}}\right)\mathrm{d}t\right].$$

Finally, the corresponding PDE is given as follows:

$$0=\mathcal{L}_{y,I}\phi(\lambda,I)+\mathcal{U}_1(\lambda,I)\mathbf{1}_{\{\lambda I^\gamma>\hat{W}\}}+\mathcal{U}_2(\lambda,I)\mathbf{1}_{\{0<\lambda I^\gamma\leqslant\hat{W}\}},\tag{5.1}$$

where the second-order differential operator, $\mathcal{L}_{y,I}$, is defined as:

$$\mathcal{L}_{y,I}\equiv\frac{\theta_1^2+\theta_2^2}{2}\lambda^2\frac{\partial^2}{\partial\lambda^2}+\frac{\sigma_I^2}{2}I^2\frac{\partial^2}{\partial I^2}-\theta_1\sigma_I\lambda I\frac{\partial^2}{\partial\lambda\partial I}+(\beta-r)\lambda\frac{\partial}{\partial\lambda}+\mu_I I\frac{\partial}{\partial I}-(\beta+\delta).$$

While we introduce a stochastic labor income model to investigate its impact on financial decision-making, our primary analysis remains centered on the fixed-income framework. The structure of our model effectively captures key economic dynamics, including consumption, inflation hedging, investment, life-insurance premiums, and job-switching decisions, that remain fundamental regardless of income variability. Even without explicitly modeling wage uncertainty, our framework offers valuable insights into optimal resource allocation and risk management. The results derived under the fixed-income assumption highlight robust and fundamental trade-offs among consumption, investment, insurance premiums, and job-switching behavior. Therefore, although incorporating stochastic labor income entails analytical complexity, the fixed-income model provides a strong foundation for understanding individual financial decisions in an inflationary environment. As part of future research, we plan to further investigate the PDE (5.1) using analytical or numerical methods.

5.2. Job-switching model with costs

While the main analysis in this paper assumes a frictionless labor supply environment, where individuals can switch between two job states dynamically and without incurring any cost, this simplifying assumption, though analytically convenient, overlooks the job-switching frictions inherent in realistic models of labor supply and occupational choice (see Serrano (2023) and Jeon and Shim (2023)). To address this limitation and enhance the economic relevance of our framework, we extend the model to explicitly account for non-negligible costs associated with job switching. Specifically, building on the job-switching cost structure proposed by Jeon and Shim (2023), we incorporate a fixed cost incurred each time an individual changes jobs. In this section, we retain the financial and insurance market structures introduced in Section 2, with a single modification to the labor market: we introduce a fixed cost for job transitions. All other parameters related to the two job states remain unchanged. We assign switching costs $\epsilon_2>0$ and $\epsilon_1>0$ for transitions from job J_2 to J_1 and from J_1 to J_2, respectively. When $\epsilon_2\geqslant(Y_1-Y_2)/(r+\delta)$, the transition from J_2 to J_1 becomes economically unjustifiable. Accordingly, we impose the following assumption:

Assumption 4. The following condition holds:

$$0<\epsilon_2<\frac{Y_1-Y_2}{r+\delta}.$$

Let $\omega=(\omega_t)_{t=0}^\infty$ represent the job indicator process, where $\omega_t=\mathbf{1}_{\{\Xi_t=J_1\}}$ for each $t\geqslant 0$. That is, we have

$$\omega_t=\begin{cases}1, & \text{if}\Xi_t=J_1,\\ 0, & \text{if}\Xi_t=J_2,\end{cases}\quad \text{for}t\geqslant 0.$$

Job-switching behavior is represented by the indicator process ω. We model $(\omega_t)_{t=0}^\infty$ as a càdlàg, $\mathcal{F}_t$-adapted process taking values in $\{0,1\}$. Given the infrequent nature of job switching over the life cycle, we assume that ω is of finite variation.

The dynamics of wealth in (2.2) are modified as follows:

$$\begin{aligned} \mathrm{d}X_t = &\left[rX_t - c_t - p_t + Y_1\omega_t + Y_2(1-\omega_t)\right]\mathrm{d}t + \pi_{2,t}\left\{\left(\mu - r - \mu_L + \sigma_L^2 - \rho\sigma\sigma_L\right)\mathrm{d}t + (\sigma - \rho\sigma_L)\mathrm{d}B_t - \sigma_L\sqrt{1-\rho^2}\mathrm{d}Z_t\right\} \\ &+ \pi_{0,t}\left\{\left(R - r - \mu_L + \sigma_L^2\right)\mathrm{d}t - \rho\sigma_L\mathrm{d}B_t - \sigma_L\sqrt{1-\rho^2}\mathrm{d}Z_t\right\} - \epsilon_2\left(\Delta\omega_t\right)^+ - \epsilon_1\left(\Delta\omega_t\right)^-, \quad \text{with } X_0 = x, \end{aligned} \tag{5.2}$$

where $\Delta\omega_t = \omega_{t+} - \omega_t$ and $\left(\Delta\omega_t\right)^{\pm} = \max\left\{\pm\Delta\omega_t, 0\right\}$. Specifically, $\Delta\omega_t = 1$ indicates a switch from job J_2 to J_1, while $\Delta\omega_t = -1$ corresponds to a transition from J_1 to J_2.

Because the present value of future income from job J_1 is given by $-\frac{Y_1}{r+\delta}$, and Assumption 4 ensures that $-\frac{Y_1}{r+\delta} < -\frac{Y_2}{r+\delta} - \epsilon_2$, the individual must satisfy the following wealth constraint:

$$X_t > \left(-\frac{Y_1}{r+\delta} + \epsilon_2\right)(1-\omega_t) - \frac{Y_1}{r+\delta}\omega_t.$$

Because the model incorporates job-switching costs, the optimal solution depends on both the initial job state and the initial wealth level. A strategy $(c, p, \boldsymbol{\pi}, \omega)$ that meets the required conditions, given the initial values, $\omega_0 = j \in \{0,1\}$ and $X_0 = x$, is considered admissible at (j, x). We denote the set of all such admissible strategies by $\hat{\mathcal{A}}(j,x)$.

Given the initial conditions, $\omega_0 = j \in \{0,1\}$ and $X_0 = x$, where $x > \left(-\frac{Y_1}{r+\delta} + \epsilon_2\right)(1-j) - \frac{Y_1}{r+\delta}j$, the individual's objective is to choose a strategy, $(c, p, \boldsymbol{\pi}, \omega) \in \hat{\mathcal{A}}(j,x)$, that maximizes (2.3). Accordingly, the optimization problem in (2.3) can be reformulated as follows:

$$V(j,x) = \sup_{(c,p,\boldsymbol{\pi},\omega)\in\hat{\mathcal{A}}(j,x)} \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}\omega_t + l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma}(1-\omega_t) + \delta\frac{(k_M M_t)^{1-\gamma}}{1-\gamma}\right)\mathrm{d}t\right].$$

By applying a derivation process analogous to that used for $e^{-(r+\delta)t}X_t$ in Section 3, we obtain the following static budget constraint:

$$\mathbb{E}\left[\int_0^\infty H_t\left(c_t + \delta M_t - Y_1\omega_t - Y_2(1-\omega_t)\right)\mathrm{d}t + \sum_{0\leqslant t}\left\{\epsilon_2 H_t\left(\Delta\omega_t\right)^+ + \epsilon_1 H_t\left(\Delta\omega_t\right)^-\right\}\right] \leqslant x. \tag{5.3}$$

Accounting for job-switching costs, the Lagrangian is expressed as follows:

$$\begin{aligned} \mathcal{L} \equiv &\mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left\{l_1^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma} - c_t y_t + Y_1 y_t + \delta\left(\frac{(k_M M_t)^{1-\gamma}}{1-\gamma} - M_t y_t\right)\right\}\omega_t\mathrm{d}t\right] \\ &+ \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left\{l_2^{\gamma-\gamma^*}\frac{c_t^{1-\gamma}}{1-\gamma} - c_t y_t + Y_2 y_t + \delta\left(\frac{(k_M M_t)^{1-\gamma}}{1-\gamma} - M_t y_t\right)\right\}(1-\omega_t)\mathrm{d}t\right] \\ &- \mathbb{E}\left[\sum_{0\leqslant t} e^{-(\beta+\delta)t}\left\{\epsilon_2 y_t\left(\Delta\omega_t\right)^+ + \epsilon_1 y_t\left(\Delta\omega_t\right)^-\right\}\right] + \lambda x \\ \leqslant &\mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(\left(\tilde{u}_1\left(y_t\right) + Y_1 y_t\right)\omega_t + \left(\tilde{u}_2\left(y_t\right) + Y_2 y_t\right)(1-\omega_t)\right)\mathrm{d}t\right] - \mathbb{E}\left[\sum_{0\leqslant t} e^{-(\beta+\delta)t}\left\{\epsilon_2 y_t\left(\Delta\omega_t\right)^+ + \epsilon_1 y_t\left(\Delta\omega_t\right)^-\right\}\right] + \lambda x, \end{aligned}$$

where the function, $\tilde{u}_i$, is defined in (3.10). The optimal consumption, c_t^*, and optimal bequest, M_t^*, retain the same functional forms as in (3.8) and (3.9), respectively. The corresponding dual value function is given by

$$\phi(j,\lambda) \equiv \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(\left(\tilde{u}_1\left(y_t\right) + Y_1 y_t\right)\omega_t + \left(\tilde{u}_2\left(y_t\right) + Y_2 y_t\right)(1-\omega_t)\right)\mathrm{d}t\right] - \mathbb{E}\left[\sum_{0\leqslant t} e^{-(\beta+\delta)t}\left\{\epsilon_2 y_t\left(\Delta\omega_t\right)^+ + \epsilon_1 y_t\left(\Delta\omega_t\right)^-\right\}\right].$$

If $\phi(j,\lambda)$ can be decomposed as

$$\phi(j,\lambda) = \phi_1(j,\lambda) + \phi_2(j,\lambda),$$

where

$$\begin{aligned} \phi_1(j,\lambda) &= \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(\tilde{u}_1\left(y_t\right)\omega_t + \tilde{u}_2\left(y_t\right)(1-\omega_t)\right)\mathrm{d}t\right], \\ \phi_2(j,\lambda) &= \mathbb{E}\left[\int_0^\infty e^{-(\beta+\delta)t}\left(Y_1 y_t\omega_t + Y_2 y_t(1-\omega_t)\right)\mathrm{d}t\right] - \mathbb{E}\left[\sum_{0\leqslant t} e^{-(\beta+\delta)t}\left\{\epsilon_2 y_t\left(\Delta\omega_t\right)^+ + \epsilon_1 y_t\left(\Delta\omega_t\right)^-\right\}\right], \end{aligned}$$

then applying a standard impulse control framework yields the following variational inequalities for $\phi_2(j,\lambda)$ with $j \in \{0,1\}$ and $\lambda > 0$:

$$\max\left\{\mathcal{L}_{j,\lambda}\phi_2(0,\lambda) + Y_2\lambda, \phi_2(1,\lambda) - \phi_2(0,\lambda) - \epsilon_2\lambda\right\} = 0,$$

and

$$\max\left\{\mathcal{L}_{j,\lambda}\phi_2(1,\lambda) + Y_1\lambda, \phi_2(0,\lambda) - \phi_2(1,\lambda) - \epsilon_1\lambda\right\} = 0,$$

where the differential operator $\mathcal{L}_{j,\lambda}$ is defined as

$$\mathcal{L}_{j,\lambda} = \frac{\theta_1^2 + \theta_2^2}{2}\lambda^2\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} + (\beta - r)\lambda\frac{\mathrm{d}}{\mathrm{d}\lambda} - (\beta + \delta).$$

In addition, applying Feynman–Kac formula to $\phi_1(j,\lambda)$ gives

$$0 = \mathcal{L}_{j,\lambda}\phi_1(j,\lambda) + \tilde{u}_1(\lambda)j + \tilde{u}_2(\lambda)(1-j).$$

In this subsection, we extend the original model by incorporating job-switching costs to capture the frictions inherent in real-world job transitions. This extension addresses the limitation of the unrealistic assumption of costless, instantaneous switching. By introducing such frictions, the model effectively eliminates implausibly frequent job changes and ensures that switching behavior aligns more closely with economic intuition and the standard assumptions commonly used in labor supply models. Although the added complexity precludes the derivation of closed-form solutions to the differential equations with switching costs, the dual approach and resulting variational inequalities clarify the key mechanisms through which switching costs influence the wealth threshold and optimal strategies. This framework establishes a solid foundation for future research employing numerical methods and quantitative analyses to examine how varying magnitudes of switching costs affect behavior, thereby offering valuable insights into the interaction between dynamic occupational choice and financial decision-making.

6. Concluding remarks

This study develops a unified theoretical framework that integrates inflation risk, life insurance, bequest motives, and reversible job-switching opportunities into a dynamic optimization model. Using the martingale method, we derive explicit optimal strategies that yield several novel insights. The results demonstrate that the availability of job-switching opportunities significantly increases optimal consumption, investment in risky assets, and life-insurance purchases, while inflation risk plays a crucial role in shaping portfolio composition. Furthermore, mortality risk and bequest motives emerge as key determinants of job-switching behavior—an interaction largely overlooked in previous research.

The analysis reveals that occupational choices are shaped not only by the traditional trade-off between income and leisure but also by lifespan uncertainty and bequest preferences. The proposed model offers a rigorous theoretical foundation with significant policy implications for the design of pension systems, retirement planning, and insurance pricing. Moreover, the framework is further extended to incorporate stochastic labor income, highlighting structural differences compared to the fixed-income setting. Although closed-form solutions remain intractable under income uncertainty, the model provides a foundation for analyzing optimal financial decisions in the presence of wage risk and suggests promising directions for future analytical and numerical research. Additionally, by incorporating job-switching costs, the model captures realistic frictions in occupational mobility, thereby enhancing its applicability to real-world labor market behavior. This extension establishes a crucial basis for developing solution techniques that characterize optimal strategies in the presence of such costs.

Future research should aim to refine the models of stochastic labor income and job-switching costs, derive closed-form or approximate solutions, and empirically validate the theoretical predictions. These advancements will deepen our understanding of the complex interplay between occupational choice and financial decision-making in dynamic economic environments, ultimately informing the design of more effective labor and social policies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We appreciate the insightful comments and suggestions provided by the associate editor and three anonymous referees, which have greatly improved the quality of this paper. Yong Hyun Shin was supported by the National Research Foundation of Korea (NRF), South Korea Grant [Grant No. RS-2023-00240574], and Ji-Hun Yoon was supported by the National Research Foundation of Korea (NRF) Grant [Grant No. 2022R1A5A1033624 and No. RS-2023-NR076791].

Appendix A. Proof of Proposition 1

Remark 3. For $n_1 < n < n_2$, we define the function, $\mathcal{G}_1(n)$, as follows:

$$\mathcal{G}_1(n) \equiv -\frac{g(n)}{n-n_1} = -\frac{1}{2}\left(\theta_1^2+\theta_2^2\right)\left(n-n_2\right) > 0.$$

Obviously, $\mathcal{G}_1(n)$ is a decreasing function for $n_1 < n < n_2$, and the following inequalities hold:

$$\mathcal{G}_1\left(-\frac{1-\gamma}{\gamma}\right) > \mathcal{G}_1(1) > 0 \Rightarrow -\frac{K}{\frac{1-\gamma}{\gamma}+n_1} > \frac{r+\delta}{1-n_1} > 0 \Rightarrow \frac{\frac{1-\gamma}{\gamma}+n_1}{K} + \frac{1-n_1}{r+\delta} > 0, \tag{A.1}$$

Since $g\left(-\frac{1-\gamma}{\gamma}\right) = -K$ generates that $n_1 < -\frac{1-\gamma}{\gamma} < 1 < n_2$.

Similarly, for $n_1 < n < n_2$, let us define the function $\mathcal{G}_2(n)$ as follows:

$$\mathcal{G}_2(n) \equiv -\frac{g(n)}{n-n_2} = -\frac{1}{2}\left(\theta_1^2+\theta_2^2\right)\left(n-n_1\right) < 0.$$

$\mathcal{G}_2(n)$ is also a decreasing function for $n_1 < n < n_2$, and the following inequalities hold:

$$\mathcal{G}_2(1) < \mathcal{G}_2\left(-\frac{1-\gamma}{\gamma}\right) < 0 \Rightarrow \frac{r+\delta}{1-n_2} < -\frac{K}{\frac{1-\gamma}{\gamma}+n_2} < 0 \Rightarrow \frac{\frac{1-\gamma}{\gamma}+n_2}{K} + \frac{1-n_2}{r+\delta} > 0. \tag{A.2}$$

Next, we prove Proposition 1. We redefine (3.16) as

$$\psi(t,y) \equiv \begin{cases} \psi_1(t,y), & \text{for } y > \hat{y}, \\ \psi_2(t,y), & \text{for } 0 < y \leqslant \hat{y}. \end{cases} \tag{A.3}$$

Applying the Feynman–Kac formula to (A.3) results the following PDEs

$$\begin{cases} \frac{\partial \psi_1}{\partial t} + (\beta - r) y \frac{\partial \psi_1}{\partial y} + \frac{1}{2}\left(\theta_1^2 + \theta_2^2\right) y^2 \frac{\partial^2 \psi_1}{\partial y^2} + \mathrm{e}^{-(\beta+\delta)t} \left\{ \frac{\gamma}{1-\gamma} \left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) y^{-\frac{1-\gamma}{\gamma}} + Y_1 y \right\} = 0, & \text{for } y > \hat{y}, \\ \frac{\partial \psi_2}{\partial t} + (\beta - r) y \frac{\partial \psi_2}{\partial y} + \frac{1}{2}\left(\theta_1^2 + \theta_2^2\right) y^2 \frac{\partial^2 \psi_2}{\partial y^2} + \mathrm{e}^{-(\beta+\delta)t} \left\{ \frac{\gamma}{1-\gamma} \left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) y^{-\frac{1-\gamma}{\gamma}} + Y_2 y \right\} = 0, & \text{for } 0 < y \leqslant \hat{y}. \end{cases} \tag{A.4}$$

Plugging a trial solution of the form $\psi(t,y) = e^{-(\beta+\delta)t} v_{(y)}$ into (A.4) yields the ordinary differential equations as follows:

$$\begin{cases} \frac{1}{2}\left(\theta_1^2 + \theta_2^2\right) y^2 v''(y) + (\beta - r) y v'(y) - (\beta + \delta) v(y) + \frac{\gamma}{1-\gamma} \left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) y^{-\frac{1-\gamma}{\gamma}} + Y_1 y = 0, & \text{for } y > \hat{y}, \\ \frac{1}{2}\left(\theta_1^2 + \theta_2^2\right) y^2 v''(y) + (\beta - r) y v'(y) - (\beta + \delta) v(y) + \frac{\gamma}{1-\gamma} \left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) y^{-\frac{1-\gamma}{\gamma}} + Y_2 y = 0, & \text{for } 0 < y \leqslant \hat{y}. \end{cases} \tag{A.5}$$

The solution to (A.5) is obtained as follows:

$$v(y) = \begin{cases} A y^{n_1} + \frac{\gamma}{(1-\gamma)K} \left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) y^{-\frac{1-\gamma}{\gamma}} + \frac{Y_1}{r+\delta} y, & \text{for } y > \hat{y}, \\ B y^{n_2} + \frac{\gamma}{(1-\gamma)K} \left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) y^{-\frac{1-\gamma}{\gamma}} + \frac{Y_2}{r+\delta} y, & \text{for } 0 < y \leqslant \hat{y}, \end{cases} \tag{A.6}$$

where $n_1 < 0$ and $n_2 > 1$ are two roots of

$$g(n) \equiv \frac{1}{2}\left(\theta_1^2 + \theta_2^2\right) n^2 + \left(\beta - r - \frac{1}{2}\left(\theta_1^2 + \theta_2^2\right) \right) n - (\beta + \delta) = 0.$$

The condition, $\mathcal{V}_1(\hat{y}+) = \mathcal{V}_2(\hat{y}-)$, ensures that the dual value function is continuous at the switching boundary, $\hat{y}$, i.e.,

$$v(\hat{y}-) = v(\hat{y}+).$$

From Assumption 2, we obtain

$$v'(\hat{y}-) = v'(\hat{y}+).$$

Consequently, according to (A.6), we have

$$B\hat{y}^{n_2} + \frac{\gamma}{(1-\gamma)K} \left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) \hat{y}^{-\frac{1-\gamma}{\gamma}} + \frac{Y_2}{r+\delta} \hat{y} = A\hat{y}^{n_1} + \frac{\gamma}{(1-\gamma)K} \left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) \hat{y}^{-\frac{1-\gamma}{\gamma}} + \frac{Y_1}{r+\delta} \hat{y}, \tag{A.7}$$

$$n_2 B\hat{y}^{n_2-1} - \frac{1}{K} \left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) \hat{y}^{-\frac{1}{\gamma}} + \frac{Y_2}{r+\delta} = n_1 A\hat{y}^{n_1-1} - \frac{1}{K} \left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) \hat{y}^{-\frac{1}{\gamma}} + \frac{Y_1}{r+\delta}. \tag{A.8}$$

Reorganizing (A.7) and (A.8) consequently determines the coefficients A as (3.18) and B as (3.19) in Proposition 1. Accordingly, from (A.1) and (A.2), we can show $A > 0$ and $B > 0$, respectively.

Appendix B. Proof of Proposition 2

For $\lambda > \hat{y}$,

$$\phi''(\lambda) = n_1(n_1 - 1)A\lambda^{n_1-2} + \frac{1}{\gamma K} \left(l_1^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) \lambda^{-\frac{1}{\gamma}-1} > 0,$$

since $A > 0$ and $n_1 < 0$.

For $0 < \lambda \leqslant \hat{y}$,

$$\phi''(\lambda) = n_2(n_2 - 1)B\lambda^{n_2-2} + \frac{1}{\gamma K} \left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}} \right) \lambda^{-\frac{1}{\gamma}-1} > 0,$$

since $B > 0$ and $n_2 > 1$. Hence, $\phi(\lambda)$ is strictly convex for $\lambda > 0$, implying that $\phi'(\lambda)$ is monotonically increasing in $(0, \infty)$.

By computing the first-order conditions of (3.21) with respect to λ, we obtain Eq. (3.20). Additionally,

$$\lim_{\lambda \to 0+} \phi'(\lambda) = -\infty, \quad \lim_{\lambda \to \infty} \phi'(\lambda) = \frac{Y_1}{r+\delta},$$

which generates that $-\phi'(\lambda) \in (-\frac{Y_1}{r+\delta}, \infty)$ is monotonically decreasing in $(0, \infty)$. Hence, there is a unique $\lambda^* > 0$ such that

$$x = -\phi'\left(\lambda^*\right),$$

for given $x > -\frac{Y_1}{r+\delta}$.

Appendix C. Proof of Lemma 1

Since $\mathcal{U}_1(\hat{y}+) = \mathcal{U}_2(\hat{y}-)$, the dual value function is continuous at the dual boundary $\hat{y}$, i.e.,

$$v(\hat{y}-) = v(\hat{y}+).$$

The duality relationship between the dual value function and value function is proven in Theorem 1 as (3.21). Consequently, from (3.20) and Assumption 2, the value function satisfies the value-matching condition at $\hat{x}$ is

$$V(\hat{x}-) = v(\hat{y}+) + (\hat{y}+)(\hat{x}-) = v(\hat{y}+) - (\hat{y}+)v'(\hat{y}+) = v(\hat{y}-) - (\hat{y}-)v'(\hat{y}-) = v(\hat{y}-) + (\hat{y}-)(\hat{x}+) = V(\hat{x}+).$$

Taking the first derivative of $V(\hat{x}-)$ with respect to $\hat{x}-$, we obtain

$$\begin{aligned} V'(\hat{x}-) &= v'(\hat{y}+)\frac{\mathrm{d}\hat{y}+}{\mathrm{d}\hat{x}-} + (\hat{y}+) + \frac{\mathrm{d}\hat{y}+}{\mathrm{d}\hat{x}-}(\hat{x}-) \\ &= -\frac{v'(\hat{y}+)}{v''(\hat{y}+)} + (\hat{y}+) + \frac{v'(\hat{y}+)}{v''(\hat{y}+)} \\ &= \hat{y}+. \end{aligned} \tag{C.1}$$

Similarly,

$$\begin{aligned} V'(\hat{x}+) &= v'(\hat{y}-)\frac{\mathrm{d}\hat{y}-}{\mathrm{d}\hat{x}+} + (\hat{y}-) + \frac{\mathrm{d}\hat{y}-}{\mathrm{d}\hat{x}+}(\hat{x}+) \\ &= -\frac{v'(\hat{y}-)}{v''(\hat{y}-)} + (\hat{y}-) + \frac{v'(\hat{y}-)}{v''(\hat{y}-)} \\ &= \hat{y}-. \end{aligned} \tag{C.2}$$

Then, Eqs. (C.1) and (C.2) yield

$$V'(\hat{x}-) = V'(\hat{x}+).$$

Additionally, from (3.20), we have

$$\hat{x} = -v'(\hat{y}). \tag{C.3}$$

Taking the first derivative of (C.3) with respect to $\hat{x}-$ and $\hat{x}+$ respectively, we obtain

$$1 = -v''(\hat{y}+)\frac{\mathrm{d}\hat{y}+}{\mathrm{d}\hat{x}-},\ 1 = -v''(\hat{y}-)\frac{\mathrm{d}\hat{y}-}{\mathrm{d}\hat{x}+}. \tag{C.4}$$

From Assumption 2 and (A.5), it is easy to derive

$$v''(\hat{y}-) = v''(\hat{y}+). \tag{C.5}$$

Then, from (C.4) and (C.5), we can obtain

$$\frac{\mathrm{d}\hat{y}+}{\mathrm{d}\hat{x}-} = \frac{\mathrm{d}\hat{y}-}{\mathrm{d}\hat{x}+},$$

which implies that

$$V''(\hat{x}-) = V''(\hat{x}+).$$

Appendix D. Benchmark model

For $i \in \{1,2\}$, the optimal strategies of an optimal consumption and investment model with a single job state are

$$\begin{aligned} c_t^{J_i} &= l_i^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{J_i}\right)^{-\frac{1}{\gamma}} \\ &= \frac{K l_i^{\frac{\gamma-\gamma^*}{\gamma}}}{l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}}\left(X_t^{J_i} + \frac{Y_i}{r+\delta}\right), \\ p_t^{J_i} &= \delta\left(k_M^{\frac{1-\gamma}{\gamma}}\left(y_t^{J_i}\right)^{-\frac{1}{\gamma}} - X_t^{J_i}\right) \\ &= \delta\left(\frac{K k_M^{\frac{1-\gamma}{\gamma}} - l_i^{\frac{\gamma-\gamma^*}{\gamma}} - \delta k_M^{\frac{1-\gamma}{\gamma}}}{l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}} X_t^{J_i} + \frac{K k_M^{\frac{1-\gamma}{\gamma}}}{l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}}\frac{Y_i}{r+\delta}\right), \\ \pi_{0,t}^{J_i} &= -\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\right)\left\{\frac{1}{\gamma K}\left(l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{J_i}\right)^{-\frac{1}{\gamma}}\right\} \\ &= -\frac{1}{\gamma}\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\right)\left(X_t^{J_i} + \frac{Y_i}{r+\delta}\right), \end{aligned}$$

$$\begin{aligned}
\pi_{1,t}^{J_i} &= X_t^{J_i} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\left\{\frac{1}{\gamma K}\left(l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{J_i}\right)^{-\frac{1}{\gamma}}\right\} \\
&= \frac{\theta_2+\gamma\sigma_L\sqrt{1-\rho^2}}{\gamma\sigma_L\sqrt{1-\rho^2}}X_t^{J_i} + \frac{\theta_2}{\gamma\sigma_L\sqrt{1-\rho^2}}\frac{Y_i}{r+\delta}, \\
\pi_{2,t}^{J_i} &= \left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}}\right)\left\{\frac{1}{\gamma K}\left(l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{J_i}\right)^{-\frac{1}{\gamma}}\right\} \\
&= \frac{1}{\gamma}\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}}\right)\left(X_t^{J_i} + \frac{Y_i}{r+\delta}\right),
\end{aligned}$$

where

$$y_t^{J_i} = \left[\frac{l_i^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}}{K\left(X_t^{J_i} + \frac{Y_i}{r+\delta}\right)}\right]^{\gamma}. \tag{D.1}$$

Appendix E. Detailed proof of Proposition 3

We claim from the notation clearly that $X_t^{J_i} = X_t^*$. For $X_t^* \geqslant \max\left(-\frac{Y_2}{r+\delta}, \hat{x}\right)$, we can substitute $y_t^{J_2}$ into (3.22) to generate the following function:

$$X_t^*|_{y_t^{\lambda_2^*}=y_t^{J_2}} = -n_2 B\left(y_t^{J_2}\right)^{n_2-1} + \frac{1}{K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{J_2}\right)^{-\frac{1}{\gamma}} - \frac{Y_2}{r+\delta}. \tag{E.1}$$

Substituting (D.1) into (E.1) yields

$$\begin{aligned}
X_t^*|_{y_t^{\lambda_2^*}=y_t^{J_2}} &= -n_2 B\left(\frac{l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}}{K(X_t^* + \frac{Y_2}{r+\delta})}\right)^{\gamma(n_2-1)} + X_t^* \\
&< X_t^*.
\end{aligned} \tag{E.2}$$

Inequality (E.2) holds because $B > 0$. Thus, we can conclude that

$$y_t^{\lambda_2^*} < y_t^{J_2},$$

because X_t^* is a monotonic decreasing function. Hence, we obtain

$$\begin{aligned}
c_t^{J_2} &= l_2^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{J_2}\right)^{-\frac{1}{\gamma}} < l_2^{\frac{\gamma-\gamma^*}{\gamma}}\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}} = c_t^*, \\
p_t^{J_2} &= \delta\left(k_M^{\frac{1-\gamma}{\gamma}}\left(y_t^{J_2}\right)^{-\frac{1}{\gamma}} - X_t^*\right) < \delta\left(k_M^{\frac{1-\gamma}{\gamma}}\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}} - X_t^*\right) = p_t^*, \\
\pi_{0,t}^* &= -\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\right)\left[n_2\left(n_2-1\right)B\left(y_t^{\lambda_2^*}\right)^{n_2-1} + \frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}\right] \\
&< -\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\right)\frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}} \\
&< -\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\right)\frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{J_2^*}\right)^{-\frac{1}{\gamma}} = \pi_{0,t}^{J_2}, \\
\pi_{1,t}^* &= X_t^* + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\left[n_2\left(n_2-1\right)B\left(y_t^{\lambda_2^*}\right)^{n_2-1} + \frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}\right] \\
&> X_t^* + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}} \\
&> X_t^* + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\left(\frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{J_2^*}\right)^{-\frac{1}{\gamma}}\right) = \pi_{1,t}^{J_2}, \\
\pi_{2,t}^* &= \left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}}\right)\left[n_2\left(n_2-1\right)B\left(y_t^{\lambda_2^*}\right)^{n_2-1} + \frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}\right] \\
&> \left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}}\right)\frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{\lambda_2^*}\right)^{-\frac{1}{\gamma}}
\end{aligned}$$

$$> \left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}}\right)\left(\frac{1}{\gamma K}\left(l_2^{\frac{\gamma-\gamma^*}{\gamma}} + \delta k_M^{\frac{1-\gamma}{\gamma}}\right)\left(y_t^{J_2^*}\right)^{-\frac{1}{\gamma}}\right) = \pi_{2,t}^{J_2},$$

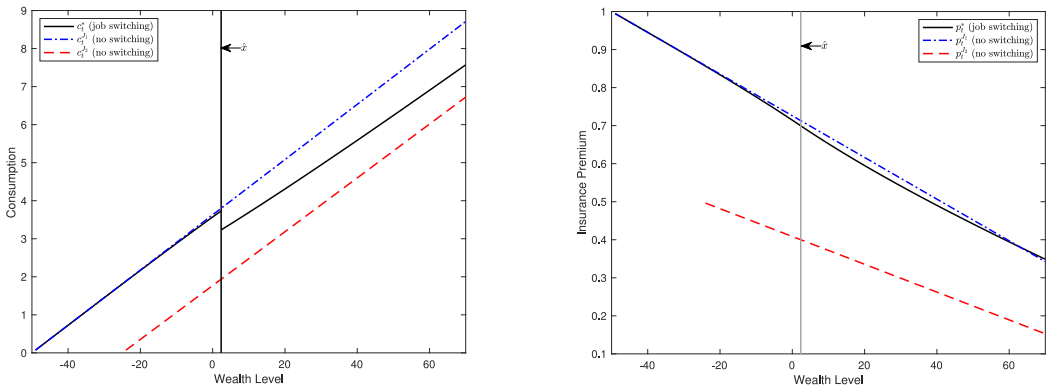
because

$$\left(\frac{\theta_1}{\sigma} - \frac{\rho\theta_2}{\sigma\sqrt{1-\rho^2}} + \frac{\theta_2}{\sigma_L\sqrt{1-\rho^2}}\right) > 0$$

holds from Assumption 3.

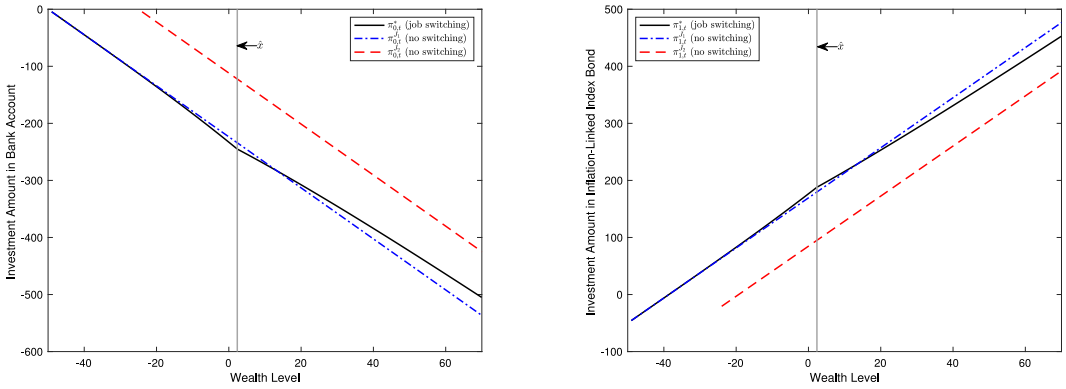
Appendix F. Negative correlation between stock and index markets

See Figs. F.3 and F.4.

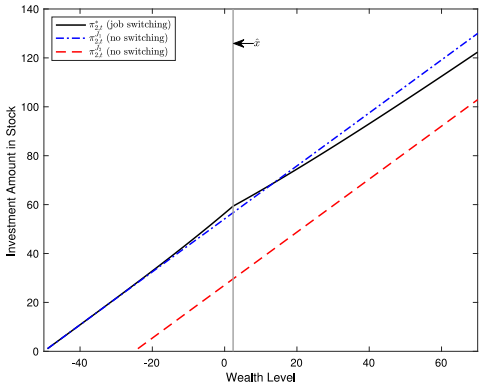


(a) Consumption (b) Insurance premium

Fig. F.3. Comparison analysis with benchmark model.



(a) Investment amount in bank account (b) Investment amount in inflation-linked index bond



(c) Investment amount in stock

Fig. F.4. Comparison analysis with benchmark model.

Data availability

No data was used for the research described in the article.

Replication Package for "Job Switching and Bequest Motives in an Optimal Consumption--Investment Model under Inflation and Mortality Risks" (Original data) (Mendeley Data)

References

Ahn, S., Choi, K.J., Lim, B.H., 2019. Optimal consumption and investment under time-varying liquidity constraints. J. Financ. Quant. Anal. 54 (4), 1643–1681. http://dx.doi.org/10.1017/S0022109018001047.

Bensoussan, A., Park, S., 2024. Optimal consumption and investment with independent stochastic labor income. Math. Oper. Res. http://dx.doi.org/10.1287/moor.2023.0119, forthcoming.

Bernheim, B.D., 1991. How strong are bequest motives? Evidence based on estimates of the demand for life insurance and annuities. J. Political Econ. 99 (5), 899–927. http://dx.doi.org/10.1086/261783.

Chen, C.C., Chang, C.C., Sun, E.W., Yu, M.T., 2022. Optimal decision of dynamic wealth allocation with life insurance for mitigating health risk under market incompleteness. European J. Oper. Res. 300 (2), 727–742. http://dx.doi.org/10.1016/j.ejor.2021.10.016.

Chen, X., Chen, P., He, Y., Wang, X., 2024. The optimal investment problem with inflation and liquidity risk. J. Comput. Appl. Math. 438, 115580. http://dx.doi.org/10.1016/j.cam.2023.115580.

Cocco, J.F., Gomes, F.J., Maenhout, P.J., 2005. Consumption and portfolio choice over the life cycle. Rev. Financ. Stud. 18 (2), 491–533. http://dx.doi.org/10.1093/rfs/hhi017.

Davidoff, T., Brown, J.R., Diamond, P.A., 2005. Annuities and individual welfare. Am. Econ. Rev. 95 (5), 1573–1590. http://dx.doi.org/10.1257/000282805775014281.

Duffie, D., Fleming, W., Soner, H.M., Zariphopoulou, T., 1997. Hedging in incomplete markets with HARA utility. J. Econom. Dynam. Control 21 (4–5), 753–782. http://dx.doi.org/10.1016/S0165-1889(97)00002-X.

Dybvig, P.H., Liu, H., 2010. Lifetime consumption and investment: retirement and constrained borrowing. J. Econom. Theory 145 (3), 885–907. http://dx.doi.org/10.1016/j.jet.2009.08.003.

Gan, L., Gong, G., Hurd, M., McFadden, D., 2015. Subjective mortality risk and bequests. J. Econometrics 188 (2), 514–525. http://dx.doi.org/10.1016/j.jeconom.2015.03.015.

Han, N.W., Hung, M.W., 2017. Optimal consumption, portfolio, and life-insurance policies under interest rate and inflation risks. Insurance Math. Econom. 73, 54–67. http://dx.doi.org/10.1016/j.insmatheco.2017.01.004.

Hubbard, R.G., Skinner, J., Zeldes, S.P., 1995. Precautionary saving and social insurance. J. Political Econ. 103 (2), 360–399. http://dx.doi.org/10.1086/261987.

Hurd, M.D., 1989. Mortality risk and bequests. Econometrica 57 (4), 779–813. http://dx.doi.org/10.2307/1913772.

Jeon, J., Oh, J., 2023. Labor supply flexibility and portfolio selection with early retirement option. Appl. Math. Optim. 88 (3), 88. http://dx.doi.org/10.1007/s00245-023-10066-6.

Jeon, J., Shim, G., 2023. Optimal consumption and investment with a costly reversible job-switching option. http://dx.doi.org/10.2139/ssrn.4553244, Working paper, Available at SSRN 4553244.

Karatzas, I., Lehoczky, J.P., Shreve, S.E., Xu, G.L., 1991. Martingale and duality methods for utility maximization in an incomplete market. SIAM J. Control Optim. 29, 702–730. http://dx.doi.org/10.1137/0329039.

Karatzas, I., Wang, H., 2000. Utility maximization with discretionary stopping. SIAM J. Control Optim. 39, 306–329. http://dx.doi.org/10.1137/S0363012998346323.

Kwak, M., Lim, B.H., 2014. Optimal portfolio selection with life insurance under inflation risk. J. Bank. Financ. 46, 59–71. http://dx.doi.org/10.1016/j.jbankfin.2014.04.019.

Lee, H.S., Shim, G., Shin, Y.H., 2019. Borrowing constraints, effective flexibility in labor supply, and portfolio selection. Math. Financ. Econ. 13, 173–208. http://dx.doi.org/10.1007/s11579-018-0224-5.

Li, Q., Shin, Y.H., Yoon, J.H., 2025. The effects of inflation risk on voluntary retirement and job switching by a martingale approach. Math. Control Relat. Fields http://dx.doi.org/10.3934/mcrf.2025006, in press.

Li, X., Yu, X., Zhang, Q., 2023. Optimal consumption and life insurance under shortfall aversion and a drawdown constraint. Insurance Math. Econom. 108, 25–45. http://dx.doi.org/10.1016/j.insmatheco.2022.11.001.

Lockwood, L.M., 2012. Bequest motives and the annuity puzzle. Rev. Econ. Dyn. 15 (2), 226–243. http://dx.doi.org/10.1016/j.red.2011.03.001.

Merton, R.C., 1969. Lifetime portfolio selection under uncertainty: The continuous-time case. Rev. Econ. Stat. 51 (3), 247–257.

Pan, J., Xiao, Q., 2017. Optimal dynamic asset–liability management with stochastic interest rates and inflation risks. Chaos Solitons Fractals 103, 460–469. http://dx.doi.org/10.1016/j.chaos.2017.07.001.

Park, K., Wong, H.Y., Yan, T., 2023. Robust retirement and life insurance with inflation risk and model ambiguity. Insurance Math. Econom. 110, 1–30. http://dx.doi.org/10.1016/j.insmatheco.2023.01.003.

Richard, S.F., 1975. Optimal consumption, portfolio, and life-insurance rules for an uncertain lived individual in a continuous-time model. J. Financ. Econ. 2 (2), 187–203. http://dx.doi.org/10.1016/0304-405X(75)90004-5.

Serrano, R., 2023. Climbing the income ladder: Search and investment in a regime-switching affine income model. Financ. Res. Lett. 58, 104330. http://dx.doi.org/10.1016/j.frl.2023.104330.

Shim, G., Shin, Y.H., 2014. An optimal job, consumption/leisure, and investment policy. Oper. Res. Lett. 42 (2), 145–149. http://dx.doi.org/10.1016/j.orl.2014.01.009.

Wang, C., Wang, N., Yang, J., 2016. Optimal consumption and savings with stochastic income and recursive utility. J. Econom. Theory 165, 292–331. http://dx.doi.org/10.1016/j.jet.2016.04.002.

Wei, J., Cheng, X., Jin, Z., Wang, H., 2020. Optimal consumption–investment and life-insurance purchase strategy for couples with correlated lifetimes. Insurance Math. Econom. 91, 244–256. http://dx.doi.org/10.1016/j.insmatheco.2020.02.006.

Zhang, Y.S., Fei, C., Pan, H.F., Huang, J., 2023. Optimal consumption, leisure, and job choice under inflationary environment. J. Oper. Res. Soc. China 11 (1), 83–107. http://dx.doi.org/10.1007/s40305-021-00369-2.

Zimmerman, F.J., Carter, M.R., 2003. Asset smoothing, consumption smoothing and the reproduction of inequality under risk and subsistence constraints. J. Dev. Econ. 71 (2), 233–260. http://dx.doi.org/10.1016/S0304-3878(03)00028-2.

Finance Research Letters 86 (2025) 108573

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Work hour flexibility and job mobility

Yong Hyun Shin [a], Ho-Seok Lee [b,*]

[a] Department of Mathematics & Research Institute of Natural Sciences, Sookmyung Women's University, Seoul 04310, Republic of Korea
[b] Department of Mathematics, Kwangwoon University, Seoul 01897, Republic of Korea

ARTICLE INFO

JEL classification:
C61
G11
J62

Keywords:
Work hour
Job mobility
Cobb–Douglas utility
Martingale and duality method

ABSTRACT

The right to arrange flexible work hours provides workers with increased utility and productivity. A job that allows employees to schedule their own work hour provides job satisfaction, and employees tend to have less mobility towards another job. We present a model that is compatible with these effects of flexible work hours. Our worker is assumed to have a Cobb–Douglas utility from consumption and leisure and maximizes expected discounted lifetime utility.

1. Introduction

Flexible work hour programs allow workers to deviate from the traditional 9 to 5 routine but to arrange their own work hour. Work flexibility legislation was enacted in many countries. For examples, German Act on Part-Time Work and Fixed-Term Contracts (Germany), Working Hours Adjustment Act (Netherlands), EU Working Time Directive (EU).[1] Flexible work hour arrangement improvement can improve work-life balance (Russell et al., 2009), raise job satisfaction and productivity (Scandura and Lankau, 1997). Angelici and Profeta (2023) provide evidence that "smart-working" enhances worker productivity and contributes to improved well-being and work-life balance.

Flexibility in work hour is indeed closely tied to job mobility. If workers were eligible for flexible work hour they are more likely to stay in their current job because they can better balance work with personal responsibilities, reducing stress and enhancing job satisfaction. Lack of flexibility in work hour may cause workers to change job or to exit job market (Gielen, 2008). Therefore, job mobility has significant implications for both employees and firms, particularly concerning costs and the loss of human capital.

In this paper we assume that the worker engages in the stochastic financial market and the labor market in which the worker can switch job. A job offers a fixed income and fixed work hour, while the other allow the worker to schedule their own work hour and receive income in accordance with labor force engagement. The first job (we call fixed time job) has income benefit comparing to the second job, and the second job (we call flexitime job) has flexitime benefit comparing to the first one. Job switch model considering utility maximization problem with optimal stochastic control problem originates from Shim and Shin (2014). With similar models of Shim and Shin (2014), job switch problem considering more realistic conditions were investigated by Jeon and Shim (2025) (job switch costs) and Lee et al. (2019) (borrowing constraints). All of Shim and Shin (2014), Jeon and Shim (2025), and Lee et al. (2019) model workers' choice between two job types differing in income and leisure (or disutility), with both assumed exogenously

* Corresponding author.
E-mail addresses: yhshin@sookmyung.ac.kr (Y.H. Shin), hoseoklee@kw.ac.kr (H.-S. Lee).

[1] In the United States, flexible work arrangements are typically driven by individual company policies rather than federal regulations. For example, Fair Labor Standards Act (FLSA) does not address flexible work hour.

https://doi.org/10.1016/j.frl.2025.108573
Received 16 May 2025; Received in revised form 26 September 2025; Accepted 26 September 2025
Available online 10 October 2025
1544-6123/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

fixed. Therefore, these studies are not suited to analyzing the effects of greater work hour flexibility on labor market outcomes. In contrast, this paper develops a job switching model between flexitime and fixed-time jobs and provides a theoretical framework that illustrates how flexible working arrangements can improve worker well-being and reduce job turnover, and promote personal wealth accumulation.

Quantitative results from our closed-form solution for the optimization problem derives meaningful explanations for effects of work hour flexibility. Firstly work hour flexibility may stimulate personal wealth accumulation and leads to economic growth of the society. Secondly the expected time for staying flexitime job increases with work hour flexibility therefore our theoretical study supports results of empirical investigations (for example, Russell et al. (2009) and Gielen (2008)) that improvement of hour flexibility reduces job mobility.

2. Financial and labor markets

We consider an infinitely-lived worker who can choose between two different types of jobs, denoted by ϕ_1 (fixed-time job) and ϕ_0 (flexitime job). Under the fixed-time job ϕ_1, the worker receives a constant income rate Y_1 and enjoys a fixed leisure share L_1 thereby working a fixed number of hours. In contrast, under the flexitime job ϕ_0, the worker can freely arrange working hours, choose the leisure share and accordingly can control the income rate. Let L denote the total time endowment, which is divided at time t into a labor share $(L - l_t)$ and a leisure share l_t, where $w > 0$ is the constant wage rate. The instantaneous income rate under ϕ_0 is then given by $w(L - l_t)$.

Assumption 2.1. We assume that $L_1 \le l_t \le L$ and $Y_1 > wL$.

From Assumption 2.1, the fixed-time job ϕ_1 provides a higher income rate and a lower leisure share compared to the flexitime job ϕ_0.

The financial market consists of a riskless asset, yielding a constant instantaneous rate of return $r > 0$, and a risky asset whose price process S_t at time t evolves according to the stochastic differential equation (SDE)

$$\frac{dS_t}{S_t} = \mu dt + \sigma dB_t,$$

where $\mu > r$, $\sigma > 0$, and $(B_t)_{t\ge0}$ is a standard Brownian motion defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathbb{F} := (\mathcal{F}_t)_{t\ge0}$ denote the $\mathbb{P}$-augmentation of the filtration generated by the standard Brownian motion $(B_t)_{t\ge0}$. We denote by $c_t \ge 0$ the consumption rate at time t, by π_t the amount of money invested in the risky asset at time t, and by λ_t the job at time t. The worker's wealth process X_t at time t evolves according to the following SDE

$$dX_t = \left[rX_t + (\mu - r)\pi_t - c_t + I(\lambda_t, l_t)\right] dt + \sigma\pi_t dB_t, \ X_0 = x, \tag{2.1}$$

where the income process $I(\lambda_t, l_t)$ is given by

$$I(\lambda_t, l_t) := Y_1 \mathbf{1}_{\{\lambda_t = \phi_1\}} + w(L - l_t)\mathbf{1}_{\{\lambda_t = \phi_0\}}.$$

A nonnegative process $\mathbf{c} := (c_t)_{t\ge0}$ is called a consumption process if it is $\mathbb{F}$-adapted and satisfies $\int_0^t c_s ds < \infty$, for all $t \ge 0$ a.s. Similarly, a nonnegative process $\mathbf{l} := (l_t)_{t\ge0}$ is called a leisure process if it is $\mathbb{F}$-adapted and satisfies $\int_0^t l_s ds < \infty$, for all $t \ge 0$ a.s. An $\mathbb{F}$-adapted process $\boldsymbol{\pi} := (\pi_t)_{t\ge0}$ is called a portfolio process if $\int_0^t \pi_s^2 ds < \infty$ for all $t \ge 0$ a.s. A job process $\lambda := (\lambda_t)_{t\ge0}$ is also assumed to be $\mathbb{F}$-adapted. We say that a strategy $(\mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda)$ is admissible at initial wealth $x > -Y_1/r$ if the corresponding wealth process X_t in (2.1) satisfies

$$X_t > -\frac{Y_1}{r}, \ t \ge 0 \text{ a.s.}$$

Let $\mathcal{A}(x)$ denote the set of all admissible strategies at x. Then, for any triple $(\mathbf{c}, \mathbf{l}, \lambda)$ of consumption, leisure, and job processes, there exists a portfolio process $\boldsymbol{\pi}$ such that $(\mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda) \in \mathcal{A}(x)$, and

$$\mathbb{E}\left[\int_0^\infty H_t \left\{c_t - I(\lambda_t, l_t)\right\} dt\right] \le x, \tag{2.2}$$

where

$$H_t := e^{-\left(r + \frac{1}{2}\theta^2\right)t - \theta B_t}, \ \theta := \frac{\mu - r}{\sigma}.$$

The worker's instantaneous utility from consumption c and leisure j is represented by the Cobb–Douglas form

$$u(c, j) := \frac{1}{\alpha} \cdot \frac{(c^\alpha j^{1-\alpha})^{1-\gamma^*}}{1 - \gamma^*}, \ 0 < \alpha < 1, \ \gamma^* > 0 \ (\gamma^* \ne 1),$$

where γ^* denotes the coefficient of relative risk aversion with respect to the joint bundle of consumption and leisure, and α (the consumption share parameter) measures contribution to the worker's utility. If we define $\gamma := 1 - \alpha(1 - \gamma^*)$, the utility function can be equivalently expressed as

$$u(c, j) = \frac{c^{1-\gamma} j^{\gamma - \gamma^*}}{1 - \gamma}.$$

3. The optimization problem

The worker aims to choose an admissible strategy $(\mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda)$ that maximizes the expected discounted lifetime utility

$$\boldsymbol{\Phi}(x; \mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda) := \mathbb{E}\left[\int_0^\infty e^{-\beta t} u(c_t, J(\lambda_t, l_t)) dt\right] = \mathbb{E}\left[\int_0^\infty e^{-\beta t} \frac{c_t^{1-\gamma} J(\lambda_t, l_t)^{\gamma-\gamma^*}}{1-\gamma} dt\right], \ x > -\frac{Y_1}{r},$$

where the leisure process $J(\lambda_t, l_t)$ is given by

$$J(\lambda_t, l_t) := L_1 \mathbf{1}_{\{\lambda_t = \phi_1\}} + l_t \mathbf{1}_{\{\lambda_t = \phi_0\}},$$

and $\beta > 0$ denotes the worker's subjective time preference rate. The associated value function is now defined as

$$V(x) := \max_{(\mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda) \in \mathcal{A}(x)} \boldsymbol{\Phi}(x; \mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda), \tag{3.1}$$

and the optimal policy is denoted by $(\mathbf{c}^*, \mathbf{l}^*, \boldsymbol{\pi}^*, \lambda^*) \in \mathcal{A}(x)$.

Assumption 3.1. We assume that

$$K^* := r + \frac{\beta - r}{\gamma^*} + \frac{\gamma^* - 1}{2(\gamma^*)^2}\theta^2 > 0, \quad K := r + \frac{\beta - r}{\gamma} + \frac{\gamma - 1}{2\gamma^2}\theta^2 > 0.$$

Assumption 3.2. We further assume that

$$\beta - r > \frac{1}{2}\theta^2.$$

We employ the martingale/duality method. From inequality (2.2), it follows that for any $y > 0$

$$\begin{aligned}
&\boldsymbol{\Phi}(x; \mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda) \\
&\le \mathbb{E}\left[\int_0^\infty e^{-\beta t} u(c_t, J(\lambda_t, l_t)) dt\right] - y\,\mathbb{E}\left[\int_0^\infty H_t \left\{c_t - I(\lambda_t, l_t)\right\} dt\right] + yx \\
&= \mathbb{E}\left[\int_0^\infty e^{-\beta t} \left\{u(c_t, J(\lambda_t, l_t)) - \left\{c_t - I(\lambda_t, l_t)\right\} y_t\right\} dt\right] + yx,
\end{aligned}$$

where $y_t := y e^{\beta t} H_t$, $y_0 = y$. We define the following convex dual utility functions:

$$\tilde{u}(\lambda, y) := \max_{c \ge 0,\ L_1 \le l \le L} \left[u(c, J(\lambda, l)) - \{c - I(\lambda, l)\} y\right]$$

and

$$\begin{aligned}
\tilde{u}_{\phi_1}(y) &= \max_{c \ge 0}\left[\frac{c^{1-\gamma} L_1^{\gamma-\gamma^*}}{1-\gamma} - cy + Y_1 y\right] = L_1^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma} y^{\frac{\gamma-1}{\gamma}} + Y_1 y, \ y > 0, \\
\tilde{u}_{\phi_0}(y) &= \max_{c \ge 0,\ L_1 \le l \le L}\left[\frac{c^{1-\gamma} l^{\gamma-\gamma^*}}{1-\gamma} - cy + w(L-l)y\right] \\
&= \begin{cases}
\tilde{u}_{\phi_{0,1}}(y) := L^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma} y^{\frac{\gamma-1}{\gamma}}, \ 0 < y \le y_a, \\
\tilde{u}_{\phi_{0,2}}(y) := \left(\frac{w\alpha}{1-\alpha}\right)^{\frac{\gamma^*-\gamma}{\gamma^*}} \frac{\gamma^*}{\alpha(1-\gamma^*)} y^{\frac{\gamma^*-1}{\gamma^*}} + wLy, \ y_a < y < y_b, \\
\tilde{u}_{\phi_{0,3}}(y) := L_1^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma} y^{\frac{\gamma-1}{\gamma}} + w(L - L_1)y, \ y \ge y_b,
\end{cases}
\end{aligned}$$

where $y_a := \left(\frac{1-\alpha}{w\alpha}\right)^\gamma L^{-\gamma^*}$, $y_b := \left(\frac{1-\alpha}{w\alpha}\right)^\gamma L_1^{-\gamma^*}$, $y_a < y_b$. Thus we have

$$\tilde{u}(\lambda, y) = \tilde{u}_{\phi_1}(y)\mathbf{1}_{\{\lambda = \phi_1\}} + \tilde{u}_{\phi_0}(y)\mathbf{1}_{\{\lambda = \phi_0\}}, \ y > 0,$$

and

$$\begin{aligned}
&\boldsymbol{\Phi}(x; \mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda) \\
&\le \mathbb{E}\left[\int_0^\infty e^{-\beta t} \tilde{u}(\lambda_t, y_t) dt\right] + yx \\
&= \mathbb{E}\left[\int_0^\infty e^{-\beta t} \left\{\tilde{u}_{\phi_1}(y_t)\mathbf{1}_{\{\lambda_t = \phi_1\}} + \tilde{u}_{\phi_0}(y_t)\mathbf{1}_{\{\lambda_t = \phi_0\}}\right\} dt\right] + yx.
\end{aligned} \tag{3.2}$$

Lemma 3.1. *Define the function*

$$H(y) := L_1^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma} y^{-\frac{1}{\gamma}} - \left(\frac{w\alpha}{1-\alpha}\right)^{\frac{\gamma^*-\gamma}{\gamma^*}} \frac{\gamma^*}{1-\gamma} y^{-\frac{1}{\gamma^*}} + (Y_1 - wL)$$

and the constant

$$\Lambda := \frac{\alpha\gamma}{(1-\alpha)(1-\gamma)}\left\{1-\left(\frac{L_1}{L}\right)^{\frac{\gamma-\gamma^*}{\gamma}}\right\} \propto \frac{L}{L_1}.$$

If $Y_1/(wL) < \Lambda$, then the equation $H(y) = 0$ admits a unique solution in (y_a, y_b), which we denote by $\bar{y}$.

Proof. From Assumption 2.1, we have

$$H(y_b) = Y_1 - w(L - L_1) > 0, \quad \lim_{y\to\infty} H(y) = Y_1 - wL > 0.$$

Moreover,

$$\lim_{y\to 0^+} H(y) = -\infty$$

and

$$H'(y) = -\frac{1}{1-\gamma}y^{-\frac{1}{\gamma^*}-1}\left\{L_1^{\frac{\gamma-\gamma^*}{\gamma}}y^{\frac{\gamma-\gamma^*}{\gamma\gamma^*}} - \left(\frac{w\alpha}{1-\alpha}\right)^{\frac{\gamma^*-\gamma}{\gamma^*}}\right\},$$

which satisfies $H'(y_b) = 0$. Furthermore,

$$H'(y) > 0 \text{ for } 0 < y < y_b, \quad H'(y) < 0 \text{ for } y > y_b.$$

Therefore, H is strictly increasing in $(0, y_b)$ and strictly decreasing in (y_b, ∞), which implies that the equation $H(y) = 0$ admits a unique solution in $(0, y_b)$, denoted by $\bar{y}$. In addition,

$$H(y_a) = Y_1 - wL\Lambda.$$

Hence, if

$$H(y_a) < 0 \quad \Leftrightarrow \quad \frac{Y_1}{wL} < \Lambda,$$

then the equation $H(y) = 0$ has a unique solution $\bar{y} \in (y_a, y_b)$. □

Assumption 3.3. We assume that

$$\frac{Y_1}{wL} < \Lambda.$$

Remark 3.1. If Assumption 3.3 does not hold, the solution to the optimization problem (3.1) coincides with that in Shim and Shin (2014).

From Assumption 3.3 and Lemma 3.1, it follows that

$$\tilde{u}_{\phi_1}(y) < \tilde{u}_{\phi_0}(y) \text{ for } 0 < y \le \bar{y}, \text{ and } \tilde{u}_{\phi_1}(y) \ge \tilde{u}_{\phi_0}(y) \text{ for } \bar{y} < y. \tag{3.3}$$

Combining (3.2) with (3.3) yields

$$\begin{aligned}&\boldsymbol{\Phi}(x; \mathbf{c}, \mathbf{l}, \boldsymbol{\pi}, \lambda)\\ &\le \mathbb{E}\left[\int_0^\infty e^{-\beta t}\left\{\tilde{u}_{\phi_0}(y_t)\mathbf{1}_{\{0<y_t\le\bar{y}\}} + \tilde{u}_{\phi_1}(y_t)\mathbf{1}_{\{\bar{y}<y_t\}}\right\}dt\right] + yx.\end{aligned}$$

Theorem 3.1. *The value function admits the representation*

$$V(x) = \min_{y>0} J(x; y),$$

where

$$\begin{aligned}J(x; y) &= \mathbb{E}\left[\int_0^\infty e^{-\beta t}\left\{\tilde{u}_{\phi_0}(y_t)\mathbf{1}_{\{0<y_t\le\bar{y}\}} + \tilde{u}_{\phi_1}(y_t)\mathbf{1}_{\{\bar{y}<y_t\}}\right\}dt\right] + yx\\ &= v(y) + yx,\end{aligned}$$

with

$$v(y) = \mathbb{E}\left[\int_0^\infty e^{-\beta t}\left\{\tilde{u}_{\phi_0}(y_t)\mathbf{1}_{\{0<y_t\le\bar{y}\}} + \tilde{u}_{\phi_1}(y_t)\mathbf{1}_{\{\bar{y}<y_t\}}\right\}dt\right], \ y_0 = y.$$

Proof. The result follows by an argument analogous to that of Theorem 1 in Lim and Lee (2021); hence, the proof is omitted. □

From Theorem 3.1, the dual value function v satisfies the following ordinary differential equation (ODE)

$$\begin{cases} \mathcal{L}v(y) + \tilde{u}_{\phi_0}(y) = 0, \ 0 < y \le \bar{y}, \\ \mathcal{L}v(y) + \tilde{u}_{\phi_1}(y) = 0, \ \bar{y} < y, \end{cases} \tag{3.4}$$

where the differential operator $\mathcal{L}$ is given by

$$\mathcal{L}v(y) = \frac{1}{2}\theta^2 y^2 \frac{d^2}{dy^2}v(y) + (\beta - r)y\frac{d}{dy}v(y) - \beta v(y).$$

Substituting the explicit forms of $\tilde{u}_{\phi_0}(\cdot)$ and $\tilde{u}_{\phi_1}(\cdot)$ into (3.4) yields

$$\begin{cases} \mathcal{L}v(y) + L^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma} y^{\frac{\gamma-1}{\gamma}} = 0, \ 0 < y \le y_a, \\ \mathcal{L}v(y) + \left(\frac{w\alpha}{1-\alpha}\right)^{\frac{\gamma^*-\gamma}{\gamma}} \frac{\gamma^*}{\alpha(1-\gamma^*)} y^{\frac{\gamma^*-1}{\gamma^*}} + wLy = 0, \ y_a < y \le \bar{y}, \\ \mathcal{L}v(y) + L_1^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma} y^{\frac{\gamma-1}{\gamma}} + Y_1 y = 0, \ \bar{y} < y. \end{cases}$$

Let $m_+ > 1$ and $m_- < 0$ be the positive and negative roots of the following quadratic equation

$$\frac{1}{2}\theta^2 m^2 + \left(\beta - r - \frac{1}{2}\theta^2\right)m - \beta = 0,$$

respectively. Then the solution to the ODE (3.4) can be expressed in the form

$$\begin{cases} v(y) = A_1 y^{m_+} + B y^{\frac{\gamma-1}{\gamma}}, \ 0 < y \le y_a, \\ v(y) = A_2 y^{m_+} + A_3 y^{m_-} + C y^{\frac{\gamma^*-1}{\gamma^*}} + Dy, \ y_a < y \le \bar{y}, \\ v(y) = A_4 y^{m_-} + E y^{\frac{\gamma-1}{\gamma}} + Fy, \ \bar{y} < y. \end{cases}$$

The explicit forms of these coefficients are given by

$$A_1 = \frac{A_2 y_a^{m_+} + A_3 y_a^{m_-} + C y_a^{\frac{\gamma^*-1}{\gamma^*}} + Dy_a - B y_a^{\frac{\gamma-1}{\gamma}}}{y_a^{m_+}},$$

$$A_2 = \frac{\left(m_- - \frac{\gamma-1}{\gamma}\right) E\bar{y}^{\frac{\gamma-1}{\gamma}} - \left(m_- - \frac{\gamma^*-1}{\gamma^*}\right) C\bar{y}^{\frac{\gamma^*-1}{\gamma^*}} + (m_- - 1)(F - D)\bar{y}}{\left(m_- - m_+\right)\bar{y}^{m_+}},$$

$$A_3 = \frac{\left(m_+ - \frac{\gamma-1}{\gamma}\right) B y_a^{\frac{\gamma-1}{\gamma}} - \left(m_+ - \frac{\gamma^*-1}{\gamma^*}\right) C y_a^{\frac{\gamma^*-1}{\gamma^*}} - (m_+ - 1)Dy_a}{\left(m_+ - m_-\right) y_a^{m_-}},$$

$$A_4 = \frac{A_2\bar{y}^{m_+} + A_3\bar{y}^{m_-} + C\bar{y}^{\frac{\gamma^*-1}{\gamma^*}} - E\bar{y}^{\frac{\gamma-1}{\gamma}} + (D - F)\bar{y}}{\bar{y}^{m_-}},$$

$$B = \frac{L^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma}}{K},$$

$$C = \frac{\left(\frac{w\alpha}{1-\alpha}\right)^{\frac{\gamma^*-\gamma}{\gamma^*}} \frac{\gamma^*}{1-\gamma}}{K^*},$$

$$D = \frac{wL}{r},$$

$$E = \frac{L_1^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma}}{K},$$

$$F = \frac{Y_1}{r}.$$

4. Applications

Define

$$\mathcal{X}(y) := -v'(y), \ y > 0,$$

and let $\mathcal{Y}$ denote the inverse function of $\mathcal{X}$. We also define $y_t^* := \mathcal{Y}(X_t)$. By arguments analogous to those employed in the proof of Theorem 2 in Lim and Lee (2021), together with the optimal strategies characterized in Theorem A.1. (see Online Appendix), it follows that the process y_t^* evolves according to the SDE

$$dy_t^* = (\beta - r)y_t^* dt - \theta y_t^* dB_t.$$

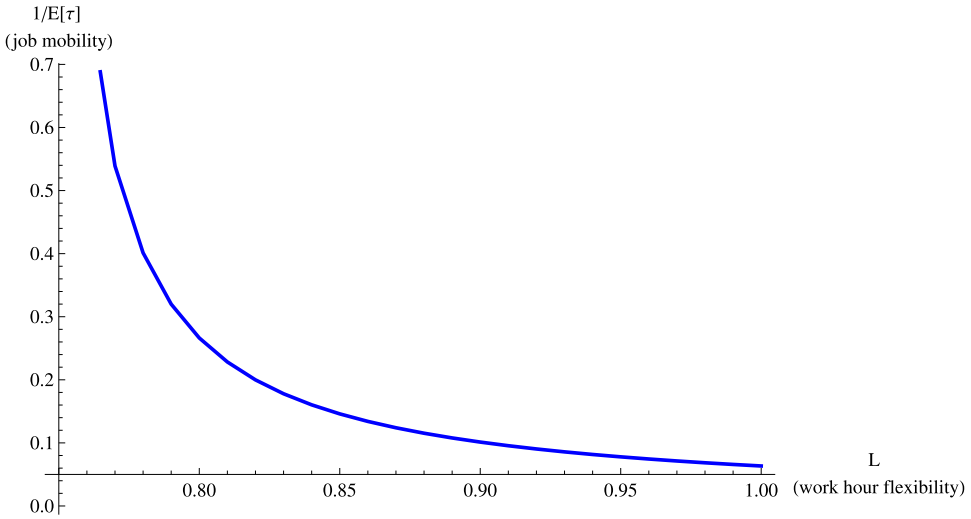


Fig. 1. Work hour flexibility and job mobility ($\gamma^* = 1.2,\ \alpha = 0.39,\ \mu = 7.89\%,\ \sigma = 18\%,\ \beta = 10.54\%,\ r = 2.38\%,\ w = 30,\ Y_1 = 50,\ L_1 = 0.1$).

By applying results from (Cox and Miller, 1965), we obtain the following proposition.

Proposition 4.1. *Let* $\tau := \inf\{t \geq s; X_t < \mathcal{X}(\bar{y}),\ X_s = \mathcal{X}(y_a)\}$. *Then, the expected hitting time is given by*

$$\mathbb{E}[\tau] = \begin{cases} \dfrac{\ln\left(\frac{\bar{y}}{y_a}\right)}{\beta - r - \frac{1}{2}\theta^2}, & \text{if } \beta - r > \frac{1}{2}\theta^2, \\ \infty, & \text{if } \beta - r \leq \frac{1}{2}\theta^2. \end{cases}$$

Moreover, under Assumption 3.2, It follows that $\mathbb{E}[\tau]$ *increases with* L.

Lemma 4.1. *The threshold* $\bar{y}$ *is strictly increasing with* L.

Proof. The constant $\bar{y}$ is characterized as the unique solution to $H(\bar{y}) = 0$, which can be equivalently written as

$$-\left(\frac{w\alpha}{1-\alpha}\right)^{\frac{\gamma^*-\gamma}{\gamma^*}} \frac{\gamma^*}{1-\gamma}\bar{y}^{\frac{\gamma^*-\gamma}{\gamma^*\gamma}} + (Y_1 - wL)\bar{y}^{\frac{1}{\gamma}} + L_1^{\frac{\gamma-\gamma^*}{\gamma}} \frac{\gamma}{1-\gamma} = 0.$$

Differentiating both sides with respect to L yields

$$\frac{\partial \bar{y}}{\partial L} = \frac{w\bar{y}^{\frac{1}{\gamma}}}{\frac{1-\alpha}{\alpha\gamma}\left(\frac{w\alpha}{1-\alpha}\right)^{\frac{\gamma^*-\gamma}{\gamma^*}}\bar{y}^{\frac{\gamma^*-\gamma}{\gamma^*\gamma}-1} + \frac{1}{\gamma}(Y_1 - wL)\bar{y}^{\frac{1}{\gamma}-1}} > 0,$$

which establishes the claim. □

Proposition 4.2. *The difference* $\mathcal{X}(y_a) - \mathcal{X}(\bar{y})$ *is strictly increasing with* L.

Proof. Since $\mathcal{X}(y)$ is strictly decreasing with y, y_a is strictly decreasing with L, and $\bar{y}$ is strictly increasing with L by Lemma 4.1, it follows that $\mathcal{X}(y_a) - \mathcal{X}(\bar{y})$ increases with L. □

Effects of working hour flexibility on productivity can be implied from the results. We can consider $1/\mathbb{E}[\tau]$ as job mobility, which decreases with work hour flexibility from Proposition 4.1. Improved work-life balance due to work hour flexibility may increase job satisfaction so workers are likely to stay with their current job (see Fig. 1).

For the worker holding job ϕ_0, $\mathcal{X}(y_a)$ is the wealth threshold for taking a break from work and enjoying full leisure, which means that the worker breaks from the job when the wealth level hits $\mathcal{X}(y_a)$ from below and goes back to work when the wealth level touches $\mathcal{X}(y_a)$ from above. We consider $\mathcal{X}(y_a) - \mathcal{X}(\bar{y})$ as the wealth accumulation while working for the job ϕ_0. With other parameters fixed, we regard L as a measurement of work hour flexibility of job ϕ_0. Therefore, increased work hour flexibility may stimulate personal wealth accumulation and improve productivity (see Fig. 2).

To improve the empirical grounding of our model, we base the parameters used in Figs. 1 and 2 on estimates from the literature and financial market data. The coefficient of relative risk aversion is set to $\gamma^* = 1.2$, in line with microeconomic estimates by Chetty

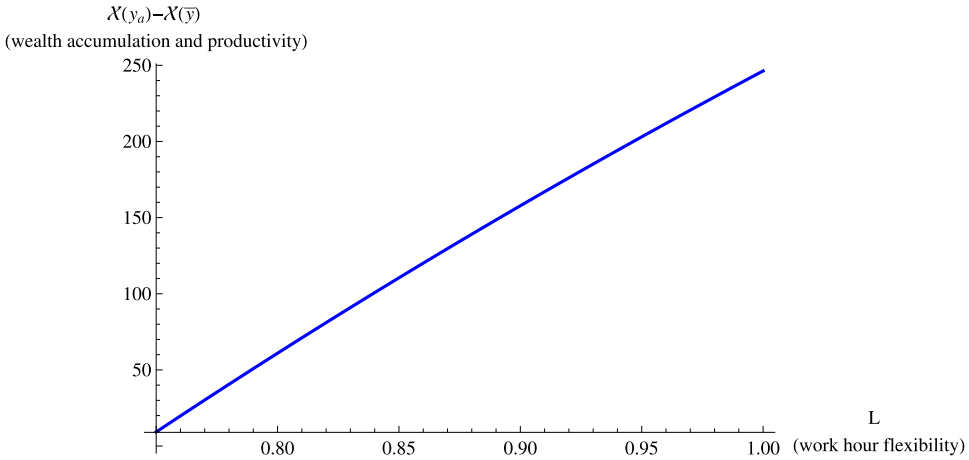


Fig. 2. Work hour flexibility and productivity ($\gamma^* = 1.2,\ \alpha = 0.39,\ \mu = 7.89\%,\ \sigma = 18\%,\ \beta = 10.54\%,\ r = 2.38\%,\ w = 30,\ Y_1 = 50,\ L_1 = 0.1$).

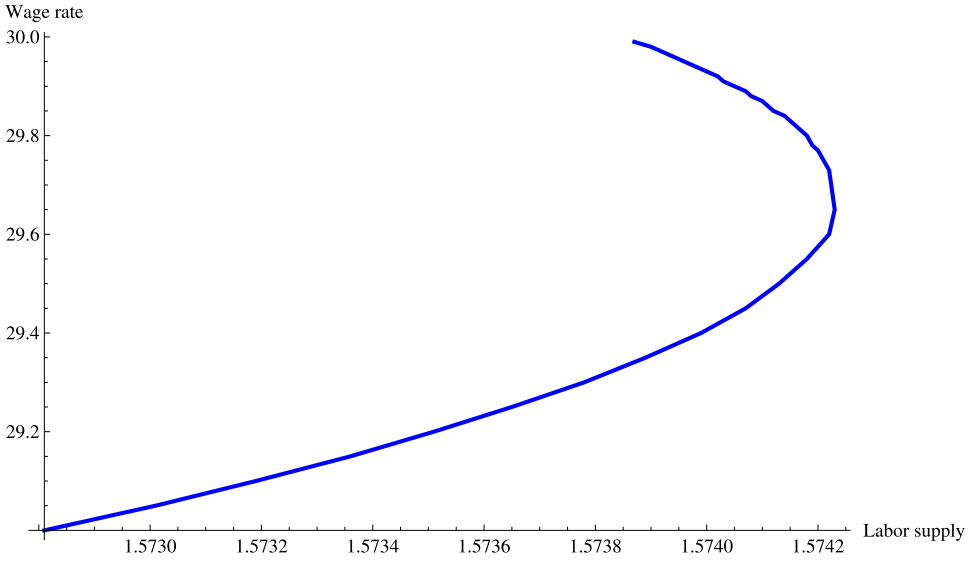


Fig. 3. Backward bending labor supply curve ($\gamma^* = 2,\ \alpha = 0.39,\ \mu = 7.89\%,\ \sigma = 18\%,\ \beta = 10.54\%,\ r = 2.38\%,\ Y_1 = 60,\ L_1 = 0.1,\ L = 2$).

(2006). The consumption share parameter $\alpha = 0.39$ follows the calibration in Domeij and Flodèn (2006). The time preference rate is set to $\beta = 10.54\%$,[2] consistent with the life-cycle consumption patterns documented in DellaVigna and Paserman (2005). For financial market calibration, we set the average return to $\mu = 7.89\%$ and the volatility to $\sigma = 18\%$, based on long-term historical data from the STOXX Europe 600 index (1979–2024). These values are supported by Curvo (2025) and the STOXX factsheet (2023). The real interest rate is set to $r = 2.38\%$, based on recent equilibrium estimates reported by the ECB (Brand et al., 2025). The remaining parameters (wage rate w, fixed-job income Y_1, and fixed-job leisure time L_1) are normalized for interpretability while preserving the qualitative features of the model.

Fig. 2 presents a partial equilibrium analysis where wages are held constant to isolate the effects of work flexibility on productivity and job switching behavior. In a full general equilibrium context, however, productivity improvements resulting from increased flexibility would typically be transmitted into higher wages, assuming competitive labor market conditions. Nonetheless, deviations from this pattern may arise due to non-competitive factors in the labor market. Furthermore, according to the theory of compensating wage differentials, for example (Hwang et al., 1998), fixed time jobs may offer higher wages to offset the disutility from reduced flexibility. This explains why fixed time jobs may pay higher wages even when marginal productivity is lower. This highlights that the patterns in Fig. 2 should not be interpreted as equilibrium wage outcomes, but rather as a stylized representation

[2] This is the continuously compounded rate converted from short-run discount factor $\beta_{imp} = 0.9$ of DellaVigna and Paserman (2005).

of workers' behavior holding wage rate fixed. A full equilibrium analysis incorporating endogenous wage setting by firms is beyond the scope of this paper but represents an important direction for future research.

In addition to these findings, our model produces a backward-bending labor supply curve under specific parameter values, a stylized feature that is widely discussed in labor supply theory (see Fig. 3). This implies that under flexible work hour arrangements, labor supply responses to wage changes can be non-monotonic: higher wages may reduce labor supply once the income effect dominates the substitution effect. The emergence the backward bending nature of labor supply in our framework, driven by the Cobb–Douglas utility function and the endogenous labor-leisure choice, confirms the model's consistency with standard theoretical results (for example, Heckman (1974), MaCurdy (1981)).

5. Conclusion and future research

Our theoretical framework formulates an optimal stochastic problem to investigate effects of flexible work hour. Quantitative results provide analytical support for empirical findings on the benefits of flexible work hours: greater wealth accumulation of the society and reduced job mobility. While our current framework focuses on a binary set of job types, it can be naturally extended to incorporate additional categories of employment that characterize modern labor markets. Non-regular jobs may offer lower or uncertain income with fixed or partly flexible leisure. Freelance work can be modeled with stochastic income risk. Remote jobs can be represented by slightly lower wages (compensating differentials) or added utility from autonomy. We leave the full implementation and estimation of such a model to future research.

CRediT authorship contribution statement

Yong Hyun Shin: Writing – original draft, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Ho-Seok Lee:** Writing – original draft, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Acknowledgments

The present research of Ho-Seok Lee has been conducted by the Research Grant of Kwangwoon University in 2023. Yong Hyun Shin was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF), South Korea [Grant No. NRF-2022S1A3A2A02089950].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.frl.2025.108573.

Data availability

No data was used for the research described in the article.

References

Angelici, M., Profeta, P., 2023. Smart working: Work flexibility without constraints. Manag. Sci. 70, 1680–1705.
Brand, C., Lisack, N., Mazelis, F., 2025. Natural rate estimates for the Euro area: insights, uncertainties and shortcomings. ECB Econ. Bull. (1/2025).
Chetty, R., 2006. A new method of estimating risk aversion. Am. Econ. Rev. 96, 1821–1834.
Cox, D.R., Miller, H.D., 1965. The Theory of Stochastic Processes. Chapman and Hall, London.
DellaVigna, S., Paserman, M.D., 2005. Job search and impatience. J. Labor Econ. 23, 437–466.
Domeij, D., Flodèn, M., 2006. The labor-supply elasticity and borrowing constraints: Why estimates are biased. Rev. Econ. Dyn. 9, 242–262.
Gielen, A.C., 2008. Working hours flexibility and older workers' labor supply. Oxf. Econ. Pap. 21, 240–274.
Heckman, J.J., 1974. Life cycle consumption and labor supply: An explanation of the relationship between income and consumption over the life cycle. Am. Econ. Rev. 64, 188–194.
Hwang, H., Reed, W.R., Hubbard, C., 1998. Compensating wage differentials and unobserved productivity. J. Political Econ. 100, 835–858.
Jeon, J., Shim, G., 2025. Optimal consumption and investment with a costly reversible job-switching option. Math. Methods Oper. Res. 101, 459–506.
Lee, H.S., Shim, G., Shin, Y.H., 2019. Borrowing constraints, effective flexibility in labor supply, and portfolio selection. Math. Financ. Econ. 13, 173–208.
Lim, B.H., Lee, H.S., 2021. Household utility maximization with life insurance: a CES utility case. Jpn. J. Ind. Appl. Math. 38, 271–295.
MaCurdy, T.E., 1981. An empirical model of labor supply in a life-cycle setting. J. Political Econ. 89, 1059–1085.
Russell, H., O'Connell, P.J., McGinnity, F., 2009. The impact of flexible work arrangements on work-life conflict and work pressure in Ireland. Gend. Work. Organ. 16, 73–97.
Scandura, T.A., Lankau, M.J., 1997. Relationships of gender, family responsibility and flexible work hours to organizational commitment and job satisfaction. J. Organ. Behav. 18, 377–391.
Shim, G., Shin, Y.H., 2014. An optimal job, consumption/leisure, and investment policy. Oper. Res. Lett. 42, 145–149.

Journal of Computational and Applied Mathematics 472 (2026) 116810

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Impact of income and leisure on optimal portfolio, consumption, and retirement decisions under exponential utility☆

Tae Ung Gang [a], Yong Hyun Shin [b,*]

[a] Stochastic Analysis and Application Research Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Republic of Korea

[b] Department of Mathematics & Research Institute of Natural Sciences, Sookmyung Women's University, Seoul, 04310, Republic of Korea

ARTICLE INFO

MSC:
91G10
93E20

Keywords:
Optimization problems
CARA utility
Retirement decision
Portfolio
Consumption-leisure

ABSTRACT

We study an optimal control problem encompassing investment, consumption, and retirement decisions under exponential (CARA-type) utility. The financial market comprises a bond with constant drift and a stock following geometric Brownian motion. The agent receives continuous income, consumes over time, and has the option to retire irreversibly, gaining increased leisure post-retirement compared to pre-retirement. The objective is to maximize the expected exponential utility of weighted consumption and leisure over an infinite horizon. Using a martingale approach and dual value function, we derive implicit solutions for the optimal portfolio, consumption, and retirement time. The analysis highlights key contributions: first, the equivalent condition for no retirement is characterized by a *retirement income threshold*; second, the influence of income and leisure levels on optimal portfolio, consumption, and retirement decisions is thoroughly examined. These results provide valuable insights into the interplay between financial and lifestyle choices in retirement planning.

1. Introduction

After the pioneering work of Merton [1,2], continuous-time optimal consumption and investment problems have become one of the central areas of mathematical finance, extensively studied under various economic conditions. One of the most significant themes in this field is the voluntary retirement problem, coupled with optimal portfolio selection (Choi and Shim [3]; Farhi and Panageas [4]; Choi et al. [5]; Dybvig and Liu [6] etc.). These studies typically address the problem as an irreversible decision, formulated mathematically as a free boundary problem. Notably, most of the existing research has focused on power-type utility functions, including the two goods utility like Cobb–Douglas utility.

In this paper, we aim to advance the field by analyzing the optimization problem using a combination of exponential-type (CARA) utility and two goods utility. To the best of our knowledge, no prior literature has examined voluntary retirement decisions involving CARA utility, making this study a novel contribution to the area.

When addressing continuous-time optimal consumption and investment problems, researchers commonly rely on two mathematical approaches. The first approach involves deriving the Hamilton–Jacobi-Bellman (HJB) equation and obtaining the solution, known as the dynamic programming method (Karatzas et al. [7]). The second approach employs the Lagrangian method to obtain a dual solution, referred to as the dual/martingale method (Karatzas et al. [8] and Cox and Huang [9]). In this study, we adopt the

☆ Tae Ung Gang was supported by the National Research Foundation of Korea (NRF) under contract No. RS-2019-NR040050, No. RS-2023-00237770, and Yong Hyun Shin was supported by the NRF under contract No. RS-2023-00240574.

* Corresponding author.
E-mail addresses: gangtaeung@kaist.ac.kr (T.U. Gang), yhshin@sookmyung.ac.kr (Y.H. Shin).

https://doi.org/10.1016/j.cam.2025.116810
Received 30 December 2024; Received in revised form 18 May 2025
Available online 6 June 2025
0377-0427/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

dual approach, integrating it with the optimal stopping problem framework (Karatzas and Wang [10]). This combination allows us to effectively model and analyze the decision-making process.

Our benchmark model builds upon the foundational work of Farhi and Panageas [4]. In their study, the authors considered an optimization problem involving power-type Cobb–Douglas utility, where the optimal leisure was represented by different constants before and after retirement. Inspired by their approach, we extend the analysis to include exponential-type two goods utility, where optimal leisure is characterized by two constants, $\underline{L} > 0$ before retirement and $\overline{L} > 0$ after retirement, with $\underline{L} < \overline{L}$. This setup enables us to examine the optimal solutions, particularly focusing on how labor income $Y > 0$ influences the voluntary retirement wealth threshold $\bar{x}$.

This paper makes two main contributions. First, we derive an equivalent condition under which individuals never retire from labor. This condition is expressed as an inequality, requiring that income Y should exceed a specific threshold (called "*retirement income threshold*") determined by the consumption weight parameter α and the difference in leisure, $\overline{L} - \underline{L}$, while Farhi and Panageas [4] theoretically analyze the threshold for retirement wealth, our study uniquely identifies the retirement income threshold. This result provides a clear economic interpretation and offers insights into how income and leisure preferences influence retirement decisions.

Second, we analytically demonstrate how optimal consumption, portfolio choices, and retirement timing are influenced by labor income Y, pre-retirement leisure $\underline{L}$, and post-retirement leisure $\overline{L}$. Specifically, we show that the retirement wealth threshold $\bar{x}$ increases with respect to Y and $\underline{L}$ but decreases with respect to $\overline{L}$. Furthermore, we find that post-retirement optimal consumption and portfolio choices are independent of $(Y, \underline{L}, \overline{L})$, and pre-retirement optimal consumption and portfolio choices show opposite tendency of monotonicity in $\underline{L}, \overline{L}$. To be more specific, pre-retirement optimal consumption is non-decreasing in $\underline{L}$ and non-increasing in $\overline{L}$ whereas pre-retirement optimal portfolio is non-increasing in $\underline{L}$ and non-decreasing in $\overline{L}$. These findings deepen our understanding of the interplay between labor income, leisure preferences, and financial decisions in the context of voluntary retirement.

Our model assumes a constant labor income stream; however, in reality, earnings are subject to various uncertainties, such as automation, climate change, and macroeconomic fluctuations. The impact of stochastic labor income on retirement planning has been studied in works like Duffie et al. [11], Cocco et al. [12], Bensoussan et al. [13], and Park [14], among others. Analyzing optimal retirement and investment decisions under income uncertainty remains an important avenue for future research. Extending our framework to incorporate stochastic income processes such as jump or diffusion models would provide a richer and more realistic understanding of financial and retirement strategies. Additionally, integrating stochastic income into the dual/martingale approach could offer new insights into how individuals adjust consumption and portfolio allocation in response to unpredictable earnings volatility.

Our study also contributes to the broader literature on retirement and investment under uncertainty. Recent researches, including Barucci and Marazzina [15], Park and Wong [16], and Ding and Marazzina [17], have explored optimal retirement and consumption decisions in settings with income and market uncertainty. These studies provide complementary perspectives, particularly in the context of incomplete markets and ambiguity aversion. Incorporating these insights into our framework could further enhance the applicability of our results, allowing for a more comprehensive analysis of retirement decision-making in dynamic economic environments.

The rest of this paper is organized as follows: We introduce our model in Section 2. Section 3 provides an optimization problem. We analyze the value function for post-retirement in Section 4 and then return to the optimization problem in Section 5. In Section 6, we derive a solution to free boundary value problem. We recall a benchmark case (Merton problem) in Section 7, and provide verification and summary in Section 8. We analyze some properties of the optimal plan in Section 9 and reach a conclusion in Section 10.

2. Model

Under the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, which satisfies the usual conditions, we define a standard Brownian motion $(B_t)_{t\geq 0}$. We consider a financial market comprising one bond and one stock, whose prices evolve according to the following ordinary differential equation (ODE) and stochastic differential equation (SDE), respectively:

$$dS_t^{(0)} = r S_t^{(0)} dt, \qquad dS_t = S_t(\mu dt + \sigma dB_t),$$

where r, μ, and σ are constants. Here, $r > 0$ represents the risk-free rate, $\mu > r$ is the stock's drift rate, and $\sigma > 0$ denotes the stock's volatility.

An agent in this market chooses a consumption rate $(c_t)_{t\geq 0}$, which is non-negative and progressively measurable with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$ and allocates $(\pi_t)_{t\geq 0}$ to the stock at time t. The agent also earns a constant labor income $Y > 0$ per unit of time until retirement. The retirement decision is modeled as a one-time irreversible choice, represented by the stopping time τ, which is $\mathcal{F}_t$-measurable. Once retired, the agent cannot return to work.

Beyond financial considerations, the agent derives utility from both consumption and leisure, where the leisure differs before and after retirement. Specifically, the leisure $(l_t)_{t\geq 0}$ is defined as

$$l_t = \begin{cases} \underline{L} & \text{if } t < \tau, \\ \overline{L} & \text{if } t \geq \tau, \end{cases}$$

where $0 < \underline{L} < \overline{L}$. This reflects the assumption that the agent experiences greater enjoyment from leisure after retirement.

The agent's consumption-leisure-portfolio plan is denoted by (c, l, π). Let $X_t = X_t^{(c,l,\pi,x)}$ denote the agent's total wealth at time t under the plan (c, l, π), with initial wealth $X_0 = x$. The wealth dynamics of the agent is given by the SDE:

$$dX_t = \left\{(\mu - r)\pi_t + rX_t - c_t + Y \cdot \mathbf{1}_{\{t\in[0,\tau)\}}\right\} dt + \sigma\pi_t dB_t, \tag{2.1}$$

where the indicator function $\mathbf{1}_{\{t\in[0,\tau)\}}$ ensures that labor income Y ceases upon retirement.

A consumption-leisure-portfolio plan (c, l, π) is admissible if it satisfies the following conditions:

$$\begin{aligned} &\int_0^t c_s ds + \int_0^t \pi_s^2 ds < \infty \quad \forall t \geq 0, \\ &X_t \begin{cases} > -\frac{Y}{r} & \forall t \in [0, \tau), \\ \geq 0 & \forall t \geq \tau. \end{cases} \end{aligned} \tag{2.2}$$

The first condition ensures that consumption and the portfolio's squared allocation are integrable over any time horizon, guaranteeing finite wealth dynamics. The second condition imposes a borrowing constraint before retirement, restricting the agent's wealth to be greater than $-Y/r$, and requires non-negative wealth after retirement. These constraints reflect the agent's inability to sustain excessive borrowing or negative wealth in retirement.

3. Optimization problem

The agent seeks to determine an admissible consumption-leisure-portfolio plan (c, l, π) that makes the following value function:

$$V(x) = \sup_{(c,\pi,\tau)} \mathbb{E}^x \left[\int_0^\infty e^{-\beta s} U(c_s, l_s) ds\right], \tag{3.1}$$

where $\beta > 0$ is the subjective discount rate, $\mathbb{E}^x[\cdot]$ represents the conditional expectation generated by the initial condition $X_0 = x$, and the utility function is defined as

$$U(c, l) := -\frac{e^{-\gamma^*\{\alpha c + (1-\alpha)l\}}}{\alpha\gamma^*} = -\frac{e^{-\gamma c}}{\gamma} \cdot A(l).$$

Here $\alpha \in (0, 1)$ is the weight parameter for consumption, $\gamma^* > 0$ is the agent's risk aversion, $\gamma := \alpha\gamma^*$, and $A(l) := e^{-\gamma^*(1-\alpha)l}$. The objective reflects the agent's trade-off between consumption and leisure over time, discounted at the rate β.

Remark 3.1 (*Allowing Leisure to Vary Continuously Before Retirement*)**.** It is not applicable in the case where leisure can vary in $[0, \underline{L}]$ before retirement, as in Choi et al. [5]. To be more specific, let $l_t \in [0, \underline{L}]$ denote leisure in the pre-retirement period, and let $Y = w(\overline{L} - l_t)$ for some fixed $w > 0$. Then, under this setup, the wealth dynamics is given by

$$dX_t = \left\{(\mu - r)\pi_t + rX_t - c_t + w(\overline{L} - l_t) \cdot \mathbf{1}_{\{t\in[0,\tau)\}}\right\} dt + \sigma\pi_t dB_t$$

and the objective function is

$$V(x) = \sup_{(c,\pi,l,\tau)} \mathbb{E}^x \left[\int_0^\infty e^{-\beta s} U(c_s, l_s) ds\right]. \tag{3.2}$$

We first observe that the budget constraint for pre-retirement is

$$\mathbb{E}^x \left[\int_0^\tau H_s(c_s + wl_s) ds + H_\tau \left(X_\tau + \frac{w\overline{L}}{r}\right)\right] \leq x + \frac{w\overline{L}}{r}.$$

Using this, we can derive $\sup_{(c,l)}[U(c, l) - (c + wl)z]$ for fixed $z > 0$ in order to find the optimal consumption-leisure (c^*, l^*) during pre-retirement. Solving the supremum yields the following first-order conditions:

$$U_c(c^*, l^*) = z, \quad U_l(c^*, l^*) = wz$$

for the optimal consumption-leisure (c^*, l^*). Since $U_c(c^*, l^*) = -\gamma U(c^*, l^*)$ and $U_l(c^*, l^*) = -\gamma^*(1 - \alpha)U(c^*, l^*)$, the above equations imply that

$$U(c^*, l^*) = -\frac{z}{\gamma} = -\frac{wz}{\gamma^*(1 - \alpha)},$$

which implies that the optimal consumption-leisure (c^*, l^*) either has no solution or infinitely many solutions.

4. Value function for post-retirement

For the post-retirement case, the wealth dynamics in (2.1) is reduced to

$$dX_t = \left\{(\mu - r)\pi_t + rX_t - c_t\right\} dt + \sigma\pi_t dB_t, \tag{4.1}$$

where the condition $X_t \geq 0$ for all $t \geq 0$ should be satisfied. Let $\theta := (\mu - r)/\sigma > 0$ denote the Sharpe ratio, and define $H_t := e^{-\left(r + \frac{1}{2}\theta^2\right)t - \theta B_t} > 0$. Applying Itô's lemma to $H_t X_t$, we obtain

$$\int_0^t H_s c_s ds + H_t X_t = x + \int_0^t H_s(\sigma\pi_s - \theta X_s) dB_s. \tag{4.2}$$

The right-hand side of (4.2) is a bounded (from below) local martingale because the left-hand side is non-negative. By Fatou's lemma, the left-hand side in (4.2) is a supermartingale. Thus,

$$\mathbb{E}^x\left[\int_0^t H_s c_s ds + H_t X_t\right] \le x.$$

If the condition $\lim_{t\to\infty}\mathbb{E}\left[H_t X_t\right] = 0$ holds (to be verified later), then the monotone convergence theorem implies

$$\mathbb{E}^x\left[\int_0^\infty H_s c_s ds\right] \le x. \tag{4.3}$$

Let us define the post-retirement value function as

$$V_{\text{post}}(x) := \sup_{(c,\pi)} \mathbb{E}^x\left[\int_0^\infty e^{-\beta s}\left(-\frac{e^{-\gamma c_s}}{\gamma}\cdot A(\overline{L})\right) ds\right]. \tag{4.4}$$

To solve (4.4), we apply the Lagrange multiplier method. For a fixed $y > 0$, the following inequality holds:

$$V_{\text{post}}(x) - yx \le \mathbb{E}^y\left[\int_0^\infty e^{-\beta s}\overline{U}\left(ye^{\beta s}H_s\right) ds\right], \tag{4.5}$$

where $\mathbb{E}^y[\cdot]$ denotes the conditional expectation given $y_0 = y$. The inequality is derived using the budget constraint in (4.3). The dual utility function $\overline{U}(y)$ is defined as

$$\overline{U}(y) := \sup_{c\ge 0}\left(-\frac{e^{-\gamma c}}{\gamma}\cdot A(\overline{L}) - cy\right) = \frac{-y\ln\left(\frac{A(\overline{L})}{y}\right) - y}{\gamma}\cdot \mathbf{1}_{\{y\le A(\overline{L})\}} - \frac{A(\overline{L})}{\gamma}\cdot \mathbf{1}_{\{y>A(\overline{L})\}}$$

with the maximizer $c = \frac{1}{\gamma}\left(\ln\left(\frac{A(\overline{L})}{y}\right)\right)^+$, where the notation $(x)^+ := \max(x, 0)$ ensures non-negativity. This result exploits the strict concavity of the map $c \mapsto -\frac{e^{-\gamma c}}{\gamma}\cdot A(\overline{L}) - cy$. Also we see that $\overline{U}$ is decreasing and convex.

Let us define the process

$$y_t := \lambda e^{\beta t} H_t, \quad \text{which evolves according to} \quad dy_t = y_t\left[(\beta - r)dt - \theta dB_t\right]. \tag{4.6}$$

The dual function for the post-retirement case is given by

$$\tilde{V}_{\text{post}}(y) := \mathbb{E}^y\left[\int_0^\infty e^{-\beta s}\overline{U}(y_s) ds\right] < 0. \tag{4.7}$$

By the Feynman–Kac formula, the function $\tilde{V}_{\text{post}}(y)$ satisfies the ODE:

$$\frac{1}{2}\theta^2 y^2 \tilde{V}''_{\text{post}}(y) + (\beta - r)y\tilde{V}'_{\text{post}}(y) - \beta\tilde{V}_{\text{post}}(y) + \overline{U}(y) = 0 \tag{4.8}$$

provided the conditions

$$\lim_{t\to\infty}\mathbb{E}\left[e^{-\beta t}\tilde{V}_{\text{post}}(y_t)\right] = 0 \quad \text{and} \quad y_t\tilde{V}'_{\text{post}}(y_t) \text{ is square integrable}$$

are satisfied. To derive this, we apply Itô's lemma to $e^{-\beta t}\tilde{V}_{\text{post}}(y_t)$:

$$\begin{aligned} d\left(e^{-\beta t}\tilde{V}_{\text{post}}(y_t)\right) &= e^{-\beta t}\left(\frac{1}{2}\theta^2 y_t^2\tilde{V}''_{\text{post}}(y_t) + (\beta - r)y_t\tilde{V}'_{\text{post}}(y_t) - \beta\tilde{V}_{\text{post}}(y_t)\right)dt - e^{-\beta t}\theta y_t\tilde{V}'_{\text{post}}(y_t)dB_t \\ &= -e^{-\beta t}\overline{U}(y_t)dt - e^{-\beta t}\theta y_t\tilde{V}'_{\text{post}}(y_t)dB_t \\ \Rightarrow \tilde{V}_{\text{post}}(y) &= \mathbb{E}^y\left[\int_0^t e^{-\beta s}\overline{U}(y_s)ds\right] + e^{-\beta t}\mathbb{E}^y[\tilde{V}_{\text{post}}(y_t)] + \theta\mathbb{E}^y\left[\int_0^t e^{-\beta s}y_s\tilde{V}'(y_s)dB_s\right] \\ &\xrightarrow{t\to\infty} \mathbb{E}^y\left[\int_0^\infty e^{-\beta s}\overline{U}(y_s)ds\right]. \end{aligned}$$

The quadratic equation $f(m) := \theta^2 m^2 + (2(\beta - r) - \theta^2)m - 2\beta = 0$ has two real roots m_- and m_+. Since $m \mapsto f(m)$ is strictly convex, $f(0) = -2\beta < 0$, and $f(1) = -2r < 0$, we get $m_- < 0 < 1 < m_+$. Furthermore, we can rewrite β and r using these roots:

$$\begin{cases} \beta = -\frac{\theta^2}{2}\cdot m_+ m_-, \\ r = -\frac{\theta^2}{2}\cdot (m_+ - 1)(m_- - 1). \end{cases} \tag{4.9}$$

Now we solve the ODE (4.8) separately for the two regions $y \le A(\overline{L})$ and $y > A(\overline{L})$ and use the above relation, we derive $C^2((0,\infty))$ solution

$$\tilde{V}_{\text{post}}(y) = \begin{cases} \frac{m_- - 1}{\gamma r m_+ (m_+ - 1)(m_+ - m_-) A(\overline{L})^{m_+ - 1}} y^{m_+} + \left(\frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln(A(\overline{L})) + 1}{\gamma r} \right) y + \frac{y \ln(y)}{\gamma r}, & \text{if } y \le A(\overline{L}), \\ \frac{m_+ - 1}{\gamma r m_- (m_- - 1)(m_+ - m_-) A(\overline{L})^{m_- - 1}} y^{m_-} - \frac{A(\overline{L})}{\gamma \beta}, & \text{if } y > A(\overline{L}), \end{cases}$$

$$\tilde{V}'_{\text{post}}(y) = \begin{cases} \frac{m_- - 1}{\gamma r (m_+ - 1)(m_+ - m_-)} \left(\frac{A(\overline{L})}{y} \right)^{1 - m_+} + \frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln\left(\frac{A(\overline{L})}{y} \right)}{\gamma r}, & \text{if } y \le A(\overline{L}), \\ \frac{m_+ - 1}{\gamma r (m_- - 1)(m_+ - m_-)} \left(\frac{A(\overline{L})}{y} \right)^{1 - m_-}, & \text{if } y > A(\overline{L}), \end{cases} \tag{4.10}$$

$$y\tilde{V}''_{\text{post}}(y) = \begin{cases} \frac{m_- - 1}{\gamma r (m_+ - m_-)} \left(\frac{A(\overline{L})}{y} \right)^{1 - m_+} + \frac{1}{\gamma r}, & \text{if } y \le A(\overline{L}), \\ \frac{m_+ - 1}{\gamma r (m_+ - m_-)} \left(\frac{A(\overline{L})}{y} \right)^{1 - m_-}, & \text{if } y > A(\overline{L}). \end{cases}$$

From the duality relation

$$V_{\text{post}}(x) = \inf_{y>0} \left(\tilde{V}_{\text{post}}(y) + yx \right),$$

we derive the first-order condition $X_t = -\tilde{V}'_{\text{post}}(y_t)$. The smoothness property $\tilde{V}_{\text{post}} \in C^3((0,\infty))$ (see Lemma A.1 for details) allows us to apply Itô's lemma to obtain the dynamics of X_t:

$$dX_t = -\left((\beta - r) y_t \tilde{V}''_{\text{post}}(y_t) + \frac{1}{2} \theta^2 y_t^2 \tilde{V}'''_{\text{post}}(y_t) \right) dt + \theta y_t \tilde{V}''_{\text{post}}(y_t) dB_t.$$

Comparing this with the dynamics of X_t in (4.1), we identify the portfolio as

$$\pi_t = \frac{\theta y_t \tilde{V}''_{\text{post}}(y_t)}{\sigma} > 0.$$

To summarize the post-retirement case, we present the following theorem:

Theorem 4.1. *In the post-retirement case, the value function V_{post}, its dual function $\tilde{V}_{\text{post}}$, the total wealth X^{post}, the optimal consumption rate c^{post}, and the optimal portfolio π^{post} are given by*

$$\begin{aligned} \tilde{V}_{\text{post}}(y) &= -\frac{1}{\gamma} \mathbb{E}^y \left[\int_0^\infty e^{-\beta s} \left\{ y_s \left(\ln\left(\frac{A(\overline{L})}{y_s} \right) + 1 \right) \mathbf{1}_{\{y_s \le A(\overline{L})\}} + A(\overline{L}) \cdot \mathbf{1}_{\{y_s > A(\overline{L})\}} \right\} ds \right], \\ V_{\text{post}}(x) &= \inf_{y>0} \left(\tilde{V}_{\text{post}}(y) + yx \right), \\ (X_t^{\text{post}}, \pi_t^{\text{post}}, c_t^{\text{post}}) &= \left(-\tilde{V}'_{\text{post}}(y_t), \frac{\theta y_t \tilde{V}''_{\text{post}}(y_t)}{\sigma}, \frac{1}{\gamma} \left(\ln\left(\frac{A(\overline{L})}{y_t} \right) \right)^+ \right), \end{aligned} \tag{4.11}$$

where $y_t := y_0 e^{\left(\beta - r - \frac{\theta^2}{2} \right) t - \theta B_t}$.

5. Return to the original problem

Let $t \ge 0$, an admissible consumption rate $(c_t)_{t \ge 0}$, and any $\mathcal{F}_t$-stopping time τ be given. Jeanblanc et al. [18] states that

$$\mathbb{E}^x \left[e^{-\beta \tau} V_{\text{post}}(X_\tau) \right] = \mathbb{E}^x \left[\int_\tau^\infty e^{-\beta s} U(c_s, \overline{L}) ds \right].$$

This equivalence allows us to reformulate (3.1) as:

$$V(x) = \sup_{(c,\pi,\tau)} \mathbb{E}^x \left[\int_0^\tau e^{-\beta s} \left(-\frac{e^{-\gamma c_s}}{\gamma} \cdot A(\underline{L}) \right) ds + e^{-\beta \tau} V_{\text{post}}(X_\tau) \right]. \tag{5.1}$$

Next, applying Itô's lemma to $H_t X_t$ (on $t < \tau$), we obtain:

$$\int_0^t H_s c_s ds + H_t \left(X_t + \frac{Y}{r} \right) = x + \int_0^t H_s (\sigma \pi_s - \theta X_s) dB_s + Y \left(\frac{H_t}{r} + \int_0^t H_s ds \right). \tag{5.2}$$

The right-hand side of (5.2) is a non-negative local martingale because (2.2) holds and $\frac{H_t}{r} + \int_0^t H_s ds$ is a martingale (shown using the definition and the independent increments of Brownian motion). Thus, by Fatou's lemma, the left-hand side of (5.2) is a supermartingale. Applying the optional sampling theorem, we derive:

$$\mathbb{E}^x \left[\int_0^\tau H_s c_s ds + H_\tau \left(X_\tau + \frac{Y}{r} \right) \right] \le x + \frac{Y}{r} \tag{5.3}$$

(see Koo et al. [19] for another approach to deriving budget constraints). Using a Lagrange multiplier approach, as in the post-retirement case, we combine (5.1) with (5.3). Together with (4.6) and the dual value for the post-retirement case, this leads to:

$$\begin{aligned} &V(x) - yx \\ &\le \sup_{\tau} \mathbb{E}^y \left[\int_0^\tau e^{-\beta s} \left(-\frac{e^{-\gamma c_s}}{\gamma} \cdot A(\underline{L}) - c_s y e^{\beta s} H_s \right) ds + e^{-\beta\tau} \left(\tilde{V}_{\text{post}}(y_\tau) - \frac{Y y_\tau}{r} \right) \right] + \frac{Y}{r} y \\ &=: \tilde{V}(y). \end{aligned} \tag{5.4}$$

Let us define the pre-retirement value function as

$$V_{\text{pre}}(x) := \sup_{(c,\pi,\tau)} \mathbb{E}^x \left[\int_0^\tau e^{-\beta s} \left(-\frac{e^{-\gamma c_s}}{\gamma} \cdot A(\underline{L}) \right) ds \right]. \tag{5.5}$$

To solve (5.5), we consider the dual function for pre-retirement

$$\tilde{V}_{\text{pre}}(y) := \sup_{\tau} \mathbb{E}^y \left[\int_0^\tau e^{-\beta s} \underline{U}(y_s) ds \right] + \frac{Y}{r} y, \tag{5.6}$$

where $\underline{U}(y) := \sup_{c \ge 0} \left(-\frac{e^{-\gamma c}}{\gamma} \cdot A(\underline{L}) - cy \right) = \frac{-y \ln\left(\frac{A(\underline{L})}{y}\right) - y}{\gamma} \cdot \mathbf{1}_{\{y \le A(\underline{L})\}} - \frac{A(\underline{L})}{\gamma} \cdot \mathbf{1}_{\{y > A(\underline{L})\}}$ with the maximizer $c = \frac{1}{\gamma} \left(\ln\left(\frac{A(\underline{L})}{y} \right) \right)^+$ since the map $c \mapsto -\frac{e^{-\gamma c}}{\gamma} \cdot A(\underline{L}) - cy$ is strictly concave. Also we see that $\underline{U}$ is decreasing and convex.

Similarly to the post-retirement case, we can establish the following results if the optimal retirement time τ^* is identified:

Theorem 5.1. *Let τ^* be the optimal retirement time. In the pre-retirement case, the value function V_{pre}, its dual function $\tilde{V}_{\text{pre}}$, the total wealth X^{pre}, the optimal consumption rate c^{pre}, and the optimal portfolio π^{pre} are given by*

$$\begin{aligned} &\tilde{V}_{\text{pre}}(y) = -\frac{1}{\gamma} \mathbb{E}^y \left[\int_0^{\tau^*} e^{-\beta s} \left\{ y_s \left(\ln\left(\frac{A(\underline{L})}{y_s} \right) + 1 \right) \mathbf{1}_{\{y_s \le A(\underline{L})\}} + A(\underline{L}) \cdot \mathbf{1}_{\{y_s > A(\underline{L})\}} \right\} ds \right] + \frac{Yy}{r}, \\ &V_{\text{pre}}(x) = \inf_{y>0} \left(\tilde{V}_{\text{pre}}(y) + yx \right), \\ &(X_t^{\text{pre}}, \pi_t^{\text{pre}}, c_t^{\text{pre}}) = \left(-\tilde{V}'_{\text{pre}}(y_t), \frac{\theta y_t \tilde{V}''_{\text{pre}}(y_t)}{\sigma}, \frac{1}{\gamma} \left(\ln\left(\frac{A(\underline{L})}{y_t} \right) \right)^+ \right). \end{aligned} \tag{5.7}$$

This formulation mirrors the structure used in the post-retirement scenario, with adjustments to account for the pre-retirement context. The components are defined in terms of the dual function and the optimal stopping time, ensuring consistency with the economic framework.

6. Solution to free boundary value problem

Under considering (5.4), let us define

$$\tilde{v}(y) := \sup_{\tau} \mathbb{E}^y \left[\int_0^\tau e^{-\beta s} \underline{U}(y_s) ds + e^{-\beta\tau} \left(\tilde{V}_{\text{post}}(y_\tau) - \frac{Y}{r} y_\tau \right) \right].$$

From (5.1) and (5.7), it follows that $\tilde{v}(y) = \tilde{V}(y) - \frac{Y}{r} y$. Solving $\tilde{v}$ is equivalent to finding $\bar{y} \ge 0$ and a function $\tilde{w} : [0, \infty) \to \mathbb{R}$ satisfying the following variational inequality (see p. 232 of Oksendal [20]):

$$\begin{cases} \frac{1}{2}\theta^2 y^2 \tilde{w}''(y) + (\beta - r) y \tilde{w}'(y) - \beta \tilde{w}(y) + \underline{U}(y) - \overline{U}(y) + Yy = 0, & \text{for } y > \bar{y}, \\ \frac{1}{2}\theta^2 y^2 \tilde{w}''(y) + (\beta - r) y \tilde{w}'(y) - \beta \tilde{w}(y) + \underline{U}(y) - \overline{U}(y) + Yy \le 0, & \text{for } 0 < y \le \bar{y}, \\ \tilde{w}(y) \ge 0, & \text{for } y > \bar{y}, \\ \tilde{w}(y) = 0, & \text{for } 0 < y \le \bar{y}, \end{cases} \tag{6.1}$$

where $\tilde{w}(y) = \tilde{v}(y) - \tilde{V}_{\text{post}}(y) + \frac{Y}{r} y$, as derived from (4.8). The stochastic representation of (6.1) is given by

$$\tilde{w}(y) = \sup_{\tau} \mathbb{E}^y \left[\int_0^\tau e^{-\beta s} \left(\underline{U}(y_s) - \overline{U}(y_s) + Y y_s \right) ds \right] \tag{6.2}$$

with $\tau = \inf\{s \ge 0 : y_s \le \bar{y}\}$ and $\bar{y} \ge 0$ is to be determined.

Note that $A(\overline{L}) < A(\underline{L})$ and

$$0 > \underline{U}(y) - \overline{U}(y) = \begin{cases} -\frac{1-\alpha}{\alpha} \cdot (\overline{L} - \underline{L}) y, & \text{if } y \le A(\overline{L}), \\ \frac{y}{\gamma} \cdot \ln\left(\frac{y}{A(\underline{L})} \right) + \frac{A(\overline{L}) - y}{\gamma}, & \text{if } A(\overline{L}) < y \le A(\underline{L}), \\ \frac{A(\overline{L}) - A(\underline{L})}{\gamma}, & \text{if } y > A(\underline{L}). \end{cases} \tag{6.3}$$

By observing this function (6.3), the map $y \mapsto \underline{U}(y) - \overline{U}(y)$ is continuously differentiable, and its derivative $y \mapsto \underline{U}'(y) - \overline{U}'(y)$ is bounded and non-decreasing. This implies that $y \mapsto \underline{U}(y) - \overline{U}(y)$ is C^1, Lipschitz continuous, and convex.

To solve (6.1) for $y > \overline{y}$, we use the variation of parameters method, yielding

$$\tilde{w}(y) = \begin{cases} Cy^{m_-} + \frac{2}{\theta^2(m_+ - m_-)} \cdot \left[y^{m_+} \int_y^\infty z^{-1-m_+} \left\{ \underline{U}(z) - \overline{U}(z) + Yz \right\} dz \right. \\ \qquad \left. + y^{m_-} \int_0^y z^{-1-m_-} \left\{ \underline{U}(z) - \overline{U}(z) + Yz \right\} dz \right], & \text{if } y > \overline{y}, \\ 0, & \text{if } 0 < y \le \overline{y}, \end{cases} \tag{6.4}$$

where C and $\overline{y}$ are to be determined. Assuming $\tilde{w} \in C^1((0,\infty))$ so that we apply the boundary conditions $\tilde{w}(\overline{y}) = \tilde{w}'(\overline{y}) = 0$, we derive the following system:

$$\begin{aligned} 0 &= C\overline{y}^{m_-} + \frac{2}{\theta^2(m_+ - m_-)} \left[\overline{y}^{m_+} \int_{\overline{y}}^\infty z^{-1-m_+} \left\{ \underline{U}(z) - \overline{U}(z) + Yz \right\} dz \right. \\ &\quad \left. + \overline{y}^{m_-} \int_0^{\overline{y}} z^{-1-m_-} \left\{ \underline{U}(z) - \overline{U}(z) + Yz \right\} dz \right], \\ 0 &= m_- C\overline{y}^{m_- - 1} + \frac{2}{\theta^2(m_+ - m_-)} \left[m_+ \overline{y}^{m_+ - 1} \int_{\overline{y}}^\infty z^{-1-m_+} \left\{ \underline{U}(z) - \overline{U}(z) + Yz \right\} dz \right. \\ &\quad \left. + m_- \overline{y}^{m_- - 1} \int_0^{\overline{y}} z^{-1-m_-} \left\{ \underline{U}(z) - \overline{U}(z) + Yz \right\} dz \right]. \end{aligned}$$

From these equations, we deduce the following:

$$\begin{aligned} 0 &= \int_{\overline{y}}^\infty z^{-1-m_+} \left\{ \underline{U}(z) - \overline{U}(z) + Yz \right\} dz =: R(\overline{y}; Y), \\ C &= -\frac{2}{\theta^2(m_+ - m_-)} \cdot \int_0^{\overline{y}} z^{-1-m_-} \left\{ \underline{U}(z) - \overline{U}(z) + Yz \right\} dz. \end{aligned} \tag{6.5}$$

The system implies that if income Y is sufficiently large, the agent will not retire. The following lemma provides a threshold for Y beyond which retirement is not optimal.

Proposition 6.1. *$\overline{y} < \infty$. In particular, $\overline{y} = 0$ if $Y \ge \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L})$.*

Proof. If $\tau = 0$ (equivalently, $\overline{y} = \infty$), it is a candidate maximizer for (6.2). To prove that $\tau = 0$ is not optimal, it suffices to demonstrate that the expectation in (6.2) is strictly positive for some stopping time $\tau > 0$.

If $Y \ge \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L})$, then $\underline{U}(y) - \overline{U}(y) + Yy \ge 0$ for all $y \le A(\overline{L})$. Since

$$\underline{U}'(y) - \overline{U}'(y) + Y = \begin{cases} Y - \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L}), & \text{if } y \le A(\overline{L}), \\ Y + \frac{1}{\gamma} \cdot \ln\left(\frac{y}{A(\underline{L})} \right), & \text{if } A(\overline{L}) < y \le A(\underline{L}), \\ Y, & \text{if } y > A(\underline{L}) \end{cases}$$

and $y \mapsto \frac{1}{\gamma} \cdot \ln\left(\frac{y}{A(\underline{L})} \right)$ is strictly increasing, we have $\underline{U}'(y) - \overline{U}'(y) + Y > 0$ for all $y > A(\overline{L})$. Therefore, $\underline{U}(y) - \overline{U}(y) + Yy > 0$ for all $y > A(\overline{L})$, and the integral (6.2) is positive. This leads to $\tau = \infty$, and to satisfy this condition, $\overline{y} = 0$.

If $Y < \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L})$, then the convexity of $y \mapsto \underline{U}(y) - \overline{U}(y)$ ensures the existence of a unique $j \in (A(\overline{L}), \infty)$ such that

$$\begin{cases} \underline{U}(y) - \overline{U}(y) + Yy < 0, & \text{if } y < j, \\ \underline{U}(y) - \overline{U}(y) + Yy = 0, & \text{if } y = j, \\ \underline{U}(y) - \overline{U}(y) + Yy > 0, & \text{if } y > j, \end{cases} \tag{6.6}$$

(see Fig. 1 for an illustration of this behavior). Define the stopping time $\tau = \inf\{s \ge 0 : y_s \le j\}$. For this choice of τ, $\tilde{w}(y) > 0$, which implies that $\overline{y} < \infty$. □

The following proposition establishes the existence and uniqueness of $\overline{y}$.

Proposition 6.2. *If $Y < \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L})$, the solution $\overline{y}$ of (6.5) exists and is unique satisfying $\overline{y} \in (0, j)$ where j is defined in (6.6).*

Proof. If $Y < \frac{1-\alpha}{\alpha} \cdot (\overline{L} - \underline{L})$, we observe that $\lim_{y \downarrow 0} R(y; Y) < 0 < R(j; Y)$. By the intermediate value theorem, there exists $\overline{y} \in (0, j)$ such that $\overline{y}$ solves (6.5). The function $y \mapsto R(y; Y)$ is strictly increasing in $(0, j)$, ensuring that $\overline{y}$ is unique. Additionally, for $y \ge j$, $R(y; Y) > 0$, which implies that $\overline{y} \in [j, \infty)$ is impossible. □

Using Proposition 6.1 and Proposition 6.2, we deduce the following corollary:

Corollary 6.1. *$\overline{y} = 0$ (i.e the agent never retires) if and only if $Y \ge \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L})$.*

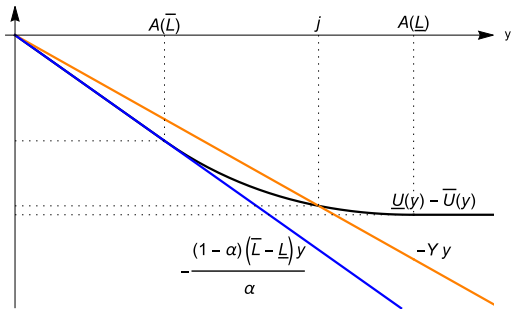


Fig. 1. The graph shows $\underline{U}(y)-\overline{U}(y)$ (black), $-Yy$ (orange), $-\frac{1-\alpha}{\alpha}\cdot(\overline{L}-\underline{L})y$ (blue), and the intersection of black and orange (denoted by j) as functions of y. In this case, $Y<\frac{1-\alpha}{\alpha}\cdot(\overline{L}-\underline{L})$.

We observe that $y\mapsto R(y;Y)$, $Y\mapsto R(y;Y)$ are continuous in $(0,\infty)$, $\underline{U}(y)-\overline{U}(y)$ is independent of Y, and $Y\mapsto R(y;Y)$ is continuous and strictly increasing with

$$\lim_{Y\downarrow 0} R(y;Y) < 0 < \lim_{Y\to\infty} R(y;Y)$$

for fixed $y\in(0,\infty)$ (these limits are valid because the monotone convergence theorem allows us to exchange the limits with the integral). As a result, we conclude that "*$Y\mapsto \bar{y}$ is a strictly decreasing function for* $Y\in\left(0,\frac{1-\alpha}{\alpha}(\overline{L}-\underline{L})\right)$". In particular, the intermediate value theorem guarantees the existence of unique values $0<Y_2<Y_1<\infty$ such that

$$0=R(A(\overline{L});Y_1)=R(A(\underline{L});Y_2).$$

The location of $\bar{y}$ depends on the value of Y. Specifically,

$$\bar{y}=\begin{cases}0, & \text{if}\quad Y\geq\frac{1-\alpha}{\alpha}(\overline{L}-\underline{L}),\\ y_1\in(0,A(\overline{L})), & \text{if}\quad Y_1<Y<\frac{1-\alpha}{\alpha}(\overline{L}-\underline{L}),\\ A(\overline{L}), & \text{if}\quad Y=Y_1,\\ y_2\in(A(\overline{L}),A(\underline{L})), & \text{if}\quad Y_2<Y<Y_1,\\ A(\underline{L}), & \text{if}\quad Y=Y_2,\\ \frac{(m_+-1)[A(\underline{L})-A(\overline{L})]}{\gamma m_+ Y}, & \text{if}\quad 0<Y<Y_2,\end{cases} \tag{6.7}$$

where y_1 and y_2 satisfy the equations

$$\begin{aligned} y_1 &= \left[\frac{A(\underline{L})^{1-m_+}-A(\overline{L})^{1-m_+}}{\gamma m_+(m_+-1)\left\{Y-\frac{1-\alpha}{\alpha}(\overline{L}-\underline{L})\right\}}\right]^{\frac{1}{1-m_+}}\\ 0 &= \left[\frac{Y}{m_+-1}+\frac{2-m_+}{\gamma(m_+-1)^2}-\frac{\ln\left(\frac{A(\underline{L})}{y_2}\right)}{\gamma(m_+-1)}\right]\left(\frac{A(\underline{L})}{y_2}\right)^{m_+-1}+\frac{A(\overline{L})}{\gamma m_+ A(\underline{L})}\left(\frac{A(\underline{L})}{y_2}\right)^{m_+}-\frac{1}{\gamma m_+(m_+-1)^2}. \end{aligned} \tag{6.8}$$

Finally, it remains to verify that $\tilde{w}(y)$ satisfies (6.1). This step completes the analysis of the problem.

Proposition 6.3. *The function $\tilde{w}$, introduced in (6.4), belongs to $C^1((0,\infty))\cap C^2((0,\infty)\setminus\{\bar{y}\})$ and solves (6.1).*

Proof. Since (6.4) satisfies both equalities in (6.1), it suffices to verify the following conditions:

$$\tfrac{1}{2}\theta^2y^2\tilde{w}''(y)+(\beta-r)y\tilde{w}'(y)-\beta\tilde{w}(y)+\underline{U}(y)-\overline{U}(y)+Yy\leq 0, \qquad \text{for } 0<y\leq\bar{y}, \tag{6.9}$$

$$\tilde{w}(y)\geq 0, \qquad \text{for } y>\bar{y}. \tag{6.10}$$

To verify (6.9), first observe that $\tilde{w}(y)=0$ for $y\leq\bar{y}$. Substituting this into (6.9), the inequality reduces to $\underline{U}(y)-\overline{U}(y)+Yy\leq 0$ for $y\leq\bar{y}$. This holds true by combining (6.6) with Proposition 6.2.

Next, to verify (6.10), restrict $y>\bar{y}$ and rewrite (6.4) using the value of C in (6.5) as follows:

$$\begin{aligned} \frac{\theta^2(m_+-m_-)\tilde{w}(y)}{2y^{m_-}} &= y^{m_+-m_-}\int_y^\infty z^{-1-m_+}\left\{\underline{U}(z)-\overline{U}(z)+Yz\right\}dz\\ &\quad+\int_{\bar{y}}^y z^{-1-m_-}\left\{\underline{U}(z)-\overline{U}(z)+Yz\right\}dz\\ &\Rightarrow \frac{d}{dy}\left(\frac{\theta^2\tilde{w}(y)}{2y^{m_-}}\right)=y^{m_+-m_--1}\int_y^\infty z^{-1-m_+}\left\{\underline{U}(z)-\overline{U}(z)+Yz\right\}dz>0, \end{aligned}$$

where the positivity follows from (6.5) and (6.6). Since $\tilde{w}(\bar{y}) = 0$, the above implies that $\tilde{w}(y) > 0$ for $y > \bar{y}$. □

7. Benchmark case: Merton problem

We consider the Merton problem to compare this with our model. In this setup, the agent has leisure $\underline{L}$ and income Y throughout her lifetime (i.e., there is no retirement). The total wealth process X_t and the value function V_{Mer} are defined as follows:

$$dX_t = \{(\mu - r)\pi_t + rX_t - c_t + Y\}\,dt + \sigma\pi_t dB_t,$$
$$V_{\text{Mer}}(x) = \sup_{(c,\pi)} \mathbb{E}^x\left[\int_0^\infty e^{-\beta s}\left(-\frac{e^{-\gamma c_s}}{\gamma}\cdot A(\underline{L})\right)ds\right].$$

Using a similar approach to the pre-retirement and post-retirement cases, we derive the following:

For the benchmark case, the value function V_{Mer}, its dual $\tilde{V}_{\text{Mer}}$, the total wealth X_t^{Mer}, optimal consumption c_t^{Mer}, and optimal portfolio π_t^{Mer} are given by

$$\tilde{V}_{\text{Mer}}(y) = -\tfrac{1}{\gamma}\,\mathbb{E}^y\left[\int_0^\infty e^{-\beta s}\left\{y_s\left(\ln\left(\tfrac{A(\underline{L})}{y_s}\right)+1\right)\mathbf{1}_{\{y_s\le A(\underline{L})\}} + A(\underline{L})\cdot\mathbf{1}_{\{y_s>A(\underline{L})\}}\right\}ds\right] + \tfrac{Y}{r}y,$$
$$V_{\text{Mer}}(x) = \inf_{y>0}(\tilde{V}_{\text{Mer}}(y) + yx),$$
$$(X_t^{\text{Mer}}, \pi_t^{\text{Mer}}, c_t^{\text{Mer}}) = \left(-\tilde{V}'_{\text{Mer}}(y_t), \tfrac{\theta y_t \tilde{V}''_{\text{Mer}}(y_t)}{\sigma}, \tfrac{1}{\gamma}\left(\ln\left(\tfrac{A(\underline{L})}{y_t}\right)\right)^+\right).$$

The closed-form expressions for $\tilde{V}_{\text{Mer}}$, $\tilde{V}'_{\text{Mer}}$, $\tilde{V}''_{\text{Mer}}$ are

$$\tilde{V}_{\text{Mer}}(y) = \tfrac{Y}{r}y + \begin{cases} \frac{m_- - 1}{\gamma r m_+(m_+-1)(m_+-m_-)A(\underline{L})^{m_+-1}}\,y^{m_+} + \left(\frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln(A(\underline{L}))+1}{\gamma r}\right)y + \frac{y\ln(y)}{\gamma r} & \text{if } y \le A(\underline{L}), \\ \frac{m_+ - 1}{\gamma r m_-(m_- - 1)(m_+-m_-)A(\underline{L})^{m_- - 1}}\,y^{m_-} - \frac{A(\underline{L})}{\gamma\beta} & \text{if } y > A(\underline{L}), \end{cases}$$
$$\tilde{V}'_{\text{Mer}}(y) = \tfrac{Y}{r} + \begin{cases} \frac{m_- - 1}{\gamma r(m_+-1)(m_+-m_-)}\left(\frac{A(\underline{L})}{y}\right)^{1-m_+} + \frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln\left(\frac{A(\underline{L})}{y}\right)}{\gamma r} & \text{if } y \le A(\underline{L}), \\ \frac{m_+ - 1}{\gamma r(m_- - 1)(m_+-m_-)}\left(\frac{A(\underline{L})}{y}\right)^{1-m_-} & \text{if } y > A(\underline{L}), \end{cases} \tag{7.1}$$
$$y\tilde{V}''_{\text{Mer}}(y) = \begin{cases} \frac{m_- - 1}{\gamma r(m_+-m_-)}\left(\frac{A(\underline{L})}{y}\right)^{1-m_+} + \frac{1}{\gamma r} & \text{if } y \le A(\underline{L}), \\ \frac{m_+ - 1}{\gamma r(m_+-m_-)}\left(\frac{A(\underline{L})}{y}\right)^{1-m_-} & \text{if } y > A(\underline{L}), \end{cases}$$

and $\tilde{V}_{\text{Mer}} \in C^3((0,\infty))$ with $\tilde{V}_{\text{Mer}}(y) < \frac{Y}{r}y$, $\tilde{V}'_{\text{Mer}}(y) < \frac{Y}{r}$, $\tilde{V}''_{\text{Mer}}(y) > 0$ for all $y > 0$.

8. Verification and summary

Theorem 8.1 (*Verification of the Value Function*). *The functions V_{post} in (4.11) and V_{pre} in (5.7) are indeed the value functions for post-retirement and pre-retirement satisfying (4.4) and (5.5), respectively.*

Proof. The basic idea follows Lemma 6.3 of Karatzas and Wang [10]. We first consider the post-retirement case. Let $y = (-\tilde{V}'_{\text{post}})^{-1}(x)$. The inequality in (4.5) becomes an equality when $\mathbb{E}^x\left[\int_0^\infty H_s c_s^{\text{post}} ds\right] = x$. Hence, it suffices to check that $\int_0^\infty \mathbb{E}^x[H_s^2\{(\pi_s^{\text{post}})^2 + (X_s^{\text{post}})^2\}]ds < \infty$ ensuring that $\int_0^t H_s(\sigma\pi_s^{\text{post}} - \theta X_s^{\text{post}})dB_s$ is a martingale (e.g see (4.2)). This is shown by

$$\int_0^\infty \mathbb{E}^x\left[H_s^2\left\{(\pi_s^{\text{post}})^2 + (X_s^{\text{post}})^2\right\}\right]ds = \int_0^\infty \mathbb{E}^x\left[H_s^2\left\{\left(\frac{\theta y_s \tilde{V}''_{\text{post}}(y_s)}{\sigma}\right)^2 + (\tilde{V}'_{\text{post}}(y_s))^2\right\}\right]ds$$
$$\le \left(\left(\frac{\theta}{\sigma}\right)^2 \sup_{z>0}|z^2\tilde{V}''_{\text{post}}(z)|^2 + \sup_{z>0}|z\tilde{V}'_{\text{post}}(z)|^2\right)\int_0^\infty \mathbb{E}^x\left[\left(\frac{H_s}{y_s}\right)^2\right]ds$$
$$= \left(\left(\frac{\theta}{\sigma}\right)^2 \sup_{z>0}|z^2\tilde{V}''_{\text{post}}(z)|^2 + \sup_{z>0}|z\tilde{V}'_{\text{post}}(z)|^2\right) y^{-2}\int_0^\infty e^{-2\beta s}ds < \infty,$$

where Lemma A.1 ensures finiteness. Combining this with Lemma A.2, we conclude that V_{post} is indeed the value function for post-retirement. For the pre-retirement case, the equality in (5.3) can be shown using a similar argument. Lemma 6.3 of Karatzas and Wang [10] ensures the existence of a portfolio π satisfying (2.2) and the equality in (5.3). □

The optimal control problem can be summarized as follows:

Theorem 8.2. *Let $\tau^* = \inf\{s > 0 : y_s \le \overline{y}\}$ where $\overline{y}$ is defined in (6.7). Then the value function V, its dual $\tilde{V}$, the optimal consumption rate c^*, and the optimal portfolio π^* are given by*

$$\begin{aligned} V(x) &= \inf_{y>0}\left(\tilde{V}(y) + yx\right), \\ \tilde{V}(y) &= \tilde{V}_{\text{pre}}(y) + \mathbb{E}\left[e^{-\beta\tau^*}\tilde{V}_{\text{post}}\left(ye^{\beta\tau^*}H_{\tau^*}\right)\right], \\ c_t^* &= c_t^{\text{pre}}\cdot\mathbf{1}_{\{t<\tau^*\}} + c_t^{\text{post}}\cdot\mathbf{1}_{\{t\ge\tau^*\}}, \\ \pi_t^* &= \pi_t^{\text{pre}}\cdot\mathbf{1}_{\{t<\tau^*\}} + \pi_t^{\text{post}}\cdot\mathbf{1}_{\{t\ge\tau^*\}}, \end{aligned} \tag{8.1}$$

and τ^* is the optimal retirement time. Furthermore, the agent never retires if and only if $Y \ge \frac{1-\alpha}{\alpha}\cdot(\overline{L}-\underline{L})$.

9. Properties of the optimal plan

We now provide a comparison between the benchmark and the optimal plan, and then analyze how income Y and the two leisure levels $(\underline{L}, \overline{L})$ affect the optimal portfolio, consumption, and retirement decision.

Proposition 9.1 (*Comparison at Same Wealth Level*). *Let us define $I_{(\text{text})} = \left(-\tilde{V}'_{(\text{text})}\right)^{-1}$ for $(\text{text}) \in \{\text{Mer}, \text{post}, \text{pre}\}$ and*

$$\begin{cases} c_*^{\text{Mer}}(x) := \frac{1}{\gamma}\left(\ln\left(\frac{A(\underline{L})}{I_{\text{Mer}}(x)}\right)\right)^+, \\ c_*^{\text{pre}}(x) := \frac{1}{\gamma}\left(\ln\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)\right)^+, \\ c_*^{\text{post}}(x) := \frac{1}{\gamma}\left(\ln\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right)\right)^+, \end{cases} \qquad \begin{cases} \pi_*^{\text{Mer}}(x) := \frac{\theta I_{\text{Mer}}(x)\tilde{V}''_{\text{Mer}}(I_{\text{Mer}}(x))}{\sigma}, \\ \pi_*^{\text{pre}}(x) := \frac{\theta I_{\text{pre}}(x)\tilde{V}''_{\text{pre}}(I_{\text{pre}}(x))}{\sigma}, \\ \pi_*^{\text{post}}(x) := \frac{\theta I_{\text{post}}(x)\tilde{V}''_{\text{post}}(I_{\text{post}}(x))}{\sigma}, \end{cases}$$

where the optimal consumption and portfolio are observed at each wealth level x. Specifically, $c_*^{\text{Mer}}(x)$ and $\pi_*^{\text{Mer}}(x)$ are defined in $(-\frac{Y}{r}, \infty)$, $c_*^{\text{post}}(x)$ and $\pi_*^{\text{post}}(x)$ are defined in $(0, \infty)$, and $c_*^{\text{pre}}(x)$ are $\pi_*^{\text{pre}}(x)$ are defined in $(-\frac{Y}{r}, -\tilde{V}'_{\text{pre}}(\overline{y})]$. Then for each x,
(i)

$$c_*^{\text{post}}(x) \le c_*^{\text{Mer}}(x) \quad \textit{and} \quad c_*^{\text{pre}}(x) \le c_*^{\text{Mer}}(x). \tag{9.1}$$

(ii)

$$\begin{cases} x \le -\tilde{V}'_{\text{Mer}}(A(\underline{L})) & \Rightarrow \quad \pi_*^{\text{post}}(x) < \pi_*^{\text{Mer}}(x) = \pi_*^{\text{pre}}(x) = (1-m_-)(x+\frac{Y}{r}) \\ x > -\tilde{V}'_{\text{Mer}}(A(\underline{L})) & \Rightarrow \quad \pi_*^{\text{post}}(x) < \pi_*^{\text{Mer}}(x) < \pi_*^{\text{pre}}(x). \end{cases} \tag{9.2}$$

Proof. (i) It suffices to verify the following two inequalities:

$$I_{\text{Mer}}(x) < I_{\text{pre}}(x), \tag{9.3}$$

and

$$\frac{A(\overline{L})}{I_{\text{post}}(x)} < \frac{A(\overline{L})}{I_{\text{Mer}}(x)}. \tag{9.4}$$

Since $\tilde{V}''_{\text{Mer}}$, $\tilde{V}''_{\text{pre}}$, $\tilde{V}''_{\text{post}}$ are positive (see Appendix for the details), (9.3) is equivalent to

$$\tilde{V}'_{\text{pre}} < \tilde{V}'_{\text{Mer}}. \tag{9.5}$$

This equivalence follows from (A.2) and (7.1) due to $C_2 > 0$. Thus (9.3) is shown.

To verify (9.4), define

$$F(X) := \begin{cases} \frac{m_- - 1}{\gamma r(m_+ - 1)(m_+ - m_-)}X^{1-m_+} + \frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln(X)}{\gamma r} & \text{for } X \ge 1, \\ \frac{m_+ - 1}{\gamma r(m_- - 1)(m_+ - m_-)}X^{1-m_-} & \text{for } X < 1, \end{cases}$$

implying

$$F'(X) = \begin{cases} -\frac{1}{X\gamma r}\left(\frac{m_- - 1}{m_+ - m_-}X^{1-m_+} + 1\right) & \text{for } X \ge 1, \\ -\frac{m_+ - 1}{\gamma r(m_+ - m_-)}X^{-m_-} & \text{for } X < 1, \end{cases} \quad < 0. \tag{9.6}$$

From the condition $0 = x + \tilde{V}'_{\text{post}}(I_{\text{post}}(x)) = x + \tilde{V}'_{\text{Mer}}(I_{\text{Mer}}(x))$, we obtain $0 = x + F\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right) = x + \frac{Y}{r} + F\left(\frac{A(\underline{L})}{I_{\text{Mer}}(x)}\right)$. Using (9.6) and $Y > 0$, inequality (9.4) follows. Thus, we have proven (9.1).

(ii) It is not hard to verify the following inequality (see Appendix for the details)

$$\tilde{V}''_{\text{post}} < \tilde{V}''_{\text{Mer}} < \tilde{V}''_{\text{pre}} \tag{9.7}$$

due to (4.10), (A.2), (7.1), $A(\overline{L}) < A(\underline{L})$, and $C_2 > 0$. Additionally, since $\tilde{V}_{\text{Mer}} \in C^3((0,\infty))$, we can compute

$$\frac{d}{dy}\left(y\tilde{V}''_{\text{Mer}}(y)\right) = \begin{cases} \frac{(m_--1)(m_+-1)}{\gamma r(m_+-m_-)A(\underline{L})^{m_+-1}} y^{m_+-2} & \text{if } y \le A(\underline{L}) \\ \frac{(m_--1)(m_+-1)}{\gamma r(m_+-m_-)A(\underline{L})^{m_--1}} y^{m_--2} & \text{if } y > A(\underline{L}) \end{cases} < 0. \tag{9.8}$$

This yields the conclusion

$$I_{\text{post}}(x)\tilde{V}''_{\text{post}}\left(I_{\text{post}}(x)\right) < I_{\text{post}}(x)\tilde{V}''_{\text{Mer}}\left(I_{\text{post}}(x)\right) < I_{\text{Mer}}(x)\tilde{V}''_{\text{Mer}}\left(I_{\text{Mer}}(x)\right), \tag{9.9}$$

where the first inequality in (9.9) uses (9.7), and the second inequality uses (9.3) and (9.8). Therefore, we conclude that $\pi_*^{\text{post}}(x) < \pi_*^{\text{Mer}}(x)$.

Next, let us consider (4.10) and (7.1). For $x \le -\tilde{V}'_{\text{Mer}}(A(\underline{L}))$, we get $\pi_*^{\text{Mer}}(x) = \pi_*^{\text{pre}}(x) = (1-m_-)(x+\frac{Y}{r})$, which follows from $0 = x + \tilde{V}'_{\text{pre}}(I_{\text{pre}}(x)) = x + \tilde{V}'_{\text{Mer}}(I_{\text{Mer}}(x))$. Thus, we restrict our attention to $x > -\tilde{V}'_{\text{Mer}}(A(\underline{L}))$ and define the following function:

$$\tilde{W}_{(p)}(I) := \frac{YI}{r} + \begin{cases} \frac{m_--1}{\gamma r m_+(m_+-1)(m_+-m_-)A(\underline{L})^{m_+-1}} I^{m_+} + \tilde{C}(p)I^{m_-} + \left(\frac{\beta-r+\frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln(A(\underline{L}))+1}{\gamma r}\right) I + \frac{I\ln(I)}{\gamma r} & \text{if } p < I \le A(\underline{L}), \\ C_3 I^{m_-} - \frac{A(\underline{L})}{\gamma\beta} & \text{if } I > A(\underline{L}), \end{cases}$$

where

$$\tilde{C}(p) := \frac{r[\ln(A(\underline{L})) - \ln(p)] - \left(\beta - 2r + \frac{\theta^2}{2}\right)}{\gamma r^2 p^{m_--1}} - C_1 p^{m_+-m_-}.$$

The above form is considered as the ODE

$$\begin{cases} \frac{\theta^2 I^2}{2} \cdot \tilde{W}''_{(p)}(I) + (\beta - r) I \tilde{W}'_{(p)}(I) - \beta \tilde{W}_{(p)}(I) + \underline{U}(I) = 0 & \forall I > p, \\ \tilde{W}_{(p)}(I) = 0 & \forall\, 0 < I \le p, \end{cases}$$

and the stochastic representation is

$$\tilde{W}_{(p)}(I) = \frac{YI}{r} + \mathbb{E}^x\left[\int_0^{\inf\{u>0\,:\,y_u \le p\}} e^{-\beta s}\underline{U}(y_s)ds\right]. \tag{9.10}$$

From this, we can find that $\tilde{W}_{(p)} \in C^2((p,\infty)) \cap C([p,\infty))$ for fixed p, and compute

$$\tilde{W}'_{(p)}(I) = \frac{Y}{r} + \begin{cases} \frac{m_--1}{\gamma r(m_+-1)(m_+-m_-)A(\underline{L})^{m_+-1}} I^{m_+-1} + \tilde{C}(p)m_- I^{m_--1} + \frac{\beta-r+\frac{\theta^2}{2}}{\gamma r^2} + \frac{\ln(I)-\ln(A(\underline{L}))}{\gamma r} & \text{if } p < I \le A(\underline{L}), \\ \left(m_-\tilde{C}(p) + \frac{m_+-1}{\gamma r(m_--1)(m_+-m_-)A(\underline{L})^{m_--1}}\right) I^{m_--1} & \text{if } I > A(\underline{L}), \end{cases}$$

$$\tilde{W}''_{(p)}(I) = \begin{cases} \frac{1}{I}\left(\frac{m_--1}{\gamma r(m_+-m_-)A(\underline{L})^{m_+-1}} I^{m_+-1} + \tilde{C}(p)m_-(m_- - 1)I^{m_--1} + \frac{1}{\gamma r}\right) & \text{if } p < I \le A(\underline{L}), \\ \left(m_-(m_- - 1)\tilde{C}(p) + \frac{m_+-1}{\gamma r(m_+-m_-)A(\underline{L})^{m_--1}}\right) I^{m_--2} & \text{if } I > A(\underline{L}). \end{cases}$$

Let a function $x \mapsto I_{(p)}(x)$ be the inverse function of $-\tilde{W}'_{(p)}$ for fixed $p \ge 0$, which is well-defined because $\tilde{W}''_{(p)} > 0$ (as in the pre-retirement case). One can then check that $\tilde{V}'_{\text{Mer}}(I_{(0)}(x)) = \tilde{W}'_{(0)}(I_{(0)}(x))$ and $\tilde{V}'_{\text{pre}}(I_{(\overline{y})}(x)) = \tilde{W}'_{(\overline{y})}(I_{(\overline{y})}(x))$. Also one can obtain $\tilde{C}'(p) \ge 0 = \tilde{C}(0)$ with positive measure of the set $\{q \ge 0 : \tilde{C}'(q) > 0\}$ by using the strictly decreasing function $p \mapsto \inf\{u > 0 : y_u \le p\}$ and (9.10).

To show $\pi_*^{\text{Mer}}(x) < \pi_*^{\text{pre}}(x)$, it is enough to show that $I_{(\overline{y})}(x)\tilde{W}''_{(\overline{y})}(I_{(\overline{y})}(x)) > I_{(0)}(x)\tilde{W}''_{(0)}(I_{(0)}(x))$. Differentiate the equation $x + \tilde{W}'_{(p)}(I_{(p)}(x)) = 0$ with respect to p to get

$$0 = \tilde{W}''_{(p)}(I_{(p)}(x))\frac{\partial I_{(p)}(x)}{\partial p} + \tilde{C}'(p)m_- I_{(p)}(x)^{m_--1}. \tag{9.11}$$

On the other hand, we have

$$\begin{aligned} \frac{\partial}{\partial p}\left(I_{(p)}(x)\tilde{W}''_{(p)}(I_{(p)}(x))\right) &= \frac{\partial}{\partial I}\left(I\tilde{W}''_{(p)}(I)\right)\Big|_{I=I_{(p)}(x)} \cdot \frac{\partial I_{(p)}(x)}{\partial p} + \tilde{C}'(p)m_-(m_- - 1)I_{(p)}(x)^{m_--1} \\ &= m_- I_{(p)}(x)^{m_--1}\tilde{C}'(p) \cdot \frac{(m_--1)\tilde{W}''_{(p)}(I) - \frac{\partial}{\partial I}\left(I\tilde{W}''_{(p)}(I)\right)}{\tilde{W}''_{(p)}(I)}\Big|_{I=I_{(p)}(x)}, \end{aligned}$$

where (9.11) is used in the second equality. Given that $I_{(p)}(x) < A(\underline{L})$ (since we restricted $x > -\tilde{V}'_{\text{Mer}}(A(\underline{L}))$), the positive measure of the set $\{q \ge 0 : m_-\tilde{C}'(q) < 0\}$, $\tilde{W}''_{(p)}(I) > 0$, and

$$(m_- - 1)\tilde{W}''_{(p)}(I)\Big|_{I=I_{(p)}(x)} - \frac{\partial}{\partial I}\left(I\tilde{W}''_{(p)}(I)\right)\Big|_{I=I_{(p)}(x)} = \frac{m_--1}{\gamma r I_{(p)}(x)}\left(1 - \left(\frac{I_{(p)}(x)}{A(\underline{L})}\right)^{m_+-1}\right) < 0,$$

we conclude

$$I_{(\overline{y})}(x)\tilde{W}''_{(\overline{y})}(I_{(\overline{y})}(x)) - I_{(0)}(x)\tilde{W}''_{(0)}(I_{(0)}(x)) = \int_0^{\overline{y}} \frac{\partial}{\partial p}\left(I_{(p)}(x)\tilde{W}''_{(p)}(I_{(p)}(x))\right) dp > 0,$$

which gives the desired result (9.2). □

Remark 9.1. When Y and $\overline{L} - \underline{L}$ are small, $c_*^{\text{pre}}(x) < c_*^{\text{post}}(x)$, contrary to expectations, as illustrated in Fig. 2, suggesting that income stability plays a crucial role in pre-retirement financial strategies. This arises due to reduced resources and uncertainty in the pre-retirement phase.

Proposition 9.2 (*Effect of Income and Leisure on Optimal Behavior*). *Let x be fixed. Then*
(i) $\overline{y}$ is strictly decreasing in $Y, \underline{L}$, and is strictly increasing in $\overline{L}$.
(ii) $c_*^{\text{post}}(x)$ and $\pi_*^{\text{post}}(x)$ are independent of $(Y, \underline{L}, \overline{L})$.
(iii) If $(-\tilde{V}'_{\text{pre}})^{-1}(x) \geq A(\underline{L})$, then $c_*^{\text{pre}}(x) = 0$ and $\pi_*^{\text{pre}}(x) = \frac{\theta(1-m_-)}{\sigma}\left(x + \frac{Y}{r}\right)$. If $(-\tilde{V}'_{\text{pre}})^{-1}(x) < A(\underline{L})$, then $c_*^{\text{pre}}(x)$ is strictly increasing in Y, $\underline{L}$, and is non-increasing in $\overline{L}$. $\pi_*^{\text{pre}}(x)$ is strictly decreasing in $\underline{L}$, and is non-decreasing in $\overline{L}$.

Proof.
(i) The integrand of the first equality in (6.5) is strictly increasing in Y, $\underline{L}$ and strictly decreasing in $\overline{L}$. Hence, we get the desired results.
(ii) From (4.10), we have $x + F\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right) = 0$. Since (m_+, m_-, γ, r) are independent of $(Y, \underline{L}, \overline{L})$ and due to Eq. (9.6) , differentiating $x + F\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right) = 0$ with respect to $P \in \{Y, \underline{L}, \overline{L}\}$ implies $F'\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right)\frac{\partial}{\partial P}\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right) = 0$. Thus, $\frac{\partial}{\partial P}\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right) = 0$. Also since $\frac{\partial}{\partial P}\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right) = 0$ and

$$I_{\text{post}}(x)\tilde{V}''_{\text{post}}(I_{\text{post}}(x)) = \begin{cases} \frac{m_- - 1}{\gamma r(m_+ - m_-)}\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right)^{1-m_+} + \frac{1}{\gamma r} & \text{if } \frac{A(\overline{L})}{I_{\text{post}}(x)} \geq 1, \\ \frac{m_+ - 1}{\gamma r(m_+ - m_-)}\left(\frac{A(\overline{L})}{I_{\text{post}}(x)}\right)^{1-m_-} & \text{if } \frac{A(\overline{L})}{I_{\text{post}}(x)} < 1, \end{cases} \tag{9.12}$$

we conclude that $\frac{\partial \pi_*^{\text{post}}}{\partial P} = 0$.
(iii) The first statement follows directly from $x + \tilde{V}'_{\text{pre}}(I_{\text{pre}}(x)) = 0$. Now, assume $(-\tilde{V}'_{\text{pre}})^{-1}(x) < A(\underline{L})$. We aim to show that c_*^{pre} is strictly increasing in Y. Differentiating $x + \tilde{V}'_{\text{pre}}(I_{\text{pre}}(x)) = 0$, and using (A.2), with respect to Y and $\overline{L}$ respectively gives

$$\frac{\partial I_{\text{pre}}(x)}{\partial Y} = -\frac{\frac{1}{r} + \frac{\partial C_2}{\partial Y} m_-[I_{\text{pre}}(x)]^{m_- - 1}}{\tilde{V}''_{\text{pre}}(I_{\text{pre}}(x))} < 0, \quad \frac{\partial I_{\text{pre}}(x)}{\partial \overline{L}} = -\frac{\frac{\partial C_2}{\partial \overline{L}} m_-[I_{\text{pre}}(x)]^{m_- - 1}}{\tilde{V}''_{\text{pre}}(I_{\text{pre}}(x))} \geq 0,$$

where $\frac{\partial C_2}{\partial Y} = \frac{\partial C_2}{\partial \overline{y}} \cdot \frac{\partial \overline{y}}{\partial Y} \leq 0$, $\frac{\partial C_2}{\partial \overline{L}} = \frac{\partial C_2}{\partial \overline{y}} \cdot \frac{\partial \overline{y}}{\partial \overline{L}} \geq 0$ (due to (i) and $\tilde{C}'(\overline{y}) \geq 0$), and $\tilde{V}''_{\text{pre}} > 0$ are used in the first and second sign, respectively. This implies $\frac{\partial c_*^{\text{pre}}(x)}{\partial Y} > 0$, $\frac{\partial c_*^{\text{pre}}(x)}{\partial \overline{L}} \leq 0$.

Combining (A.2) with $x + \tilde{V}'_{\text{pre}}(I_{\text{pre}}(x)) = 0$, we have

$$I_{\text{pre}}(x)\tilde{V}''_{\text{pre}}(I_{\text{pre}}(x)) = \frac{m_- - 1}{\gamma r}\left(\frac{1}{m_+ - 1}\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)^{1-m_+} + \ln\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)\right) + (1 - m_-)\left(x + \frac{Y}{r} + \frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2}\right) + \frac{1}{\gamma r}. \tag{9.13}$$

Differentiating (9.13) with respect to $\overline{L}$, we get $sgn\left(\frac{\partial \pi_*^{\text{pre}}(x)}{\partial \overline{L}}\right) = -sgn\left(\frac{\partial c_*^{\text{pre}}(x)}{\partial \overline{L}}\right)$ due to $I_{pre}(x) < A(\underline{L})$ and $m_+ > 1$. Combining this with $\frac{\partial c_*^{\text{pre}}(x)}{\partial \overline{L}} \leq 0$, we conclude $\frac{\partial \pi_*^{\text{pre}}(x)}{\partial \overline{L}} \geq 0$.

To find sign of $\frac{\partial c_*^{\text{pre}}(x)}{\partial \underline{L}}$ and $\frac{\partial \pi_*^{\text{pre}}(x)}{\partial \underline{L}}$, we evaluate (by the same way as (9.13))

$$sgn\left(\frac{\partial \pi_*^{\text{pre}}(x)}{\partial \underline{L}}\right) = -sgn\left(\frac{\partial c_*^{\text{pre}}(x)}{\partial \underline{L}}\right). \tag{9.14}$$

From (A.2), together with $\tilde{V}_{\text{pre}}(\overline{y}) - \frac{Y\overline{y}}{r} = 0 = x + \tilde{V}'_{\text{pre}}(I_{\text{pre}}(x))$, we derive

$$\begin{aligned} -C_2 m_- A(\underline{L})^{m_- - 1} &= m_-\left(\frac{A(\underline{L})}{\overline{y}}\right)^{m_- - 1}\left(\frac{m_- - 1}{\gamma r m_+(m_+ - 1)(m_+ - m_-)}\left(\frac{A(\underline{L})}{\overline{y}}\right)^{1-m_+} + \frac{\beta - 2r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln\left(\frac{A(\underline{L})}{\overline{y}}\right)}{\gamma r}\right) \\ &= \left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)^{m_- - 1}\left(x + \frac{Y}{r} + \frac{m_- - 1}{\gamma r(m_+ - 1)(m_+ - m_-)}\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)^{1-m_+} + \frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)}{\gamma r}\right). \end{aligned} \tag{9.15}$$

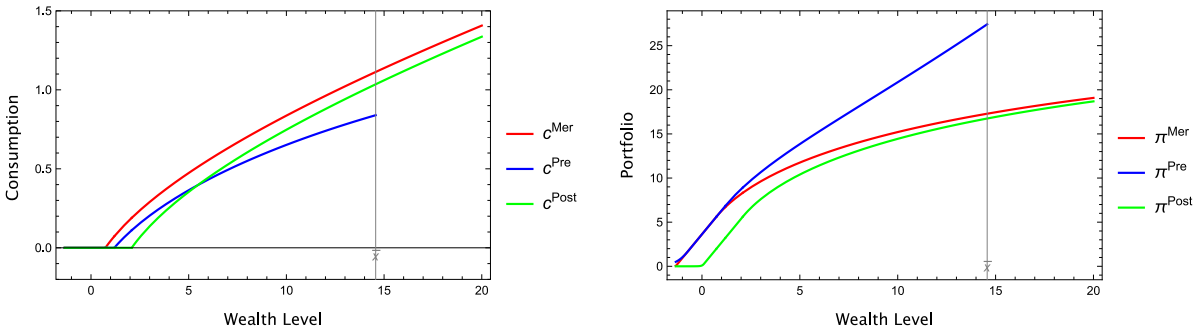


Fig. 2. The two graphs describe optimal consumption and portfolio as a function of total wealth level. The parameters are $\beta = r = 0.0371$, $\mu = 0.1123$, $\sigma = 0.1954$, $\gamma = 2$, $\underline{L} = 0.05$, $\overline{L} = 0.1$, $Y = 0.05$, $\alpha = 0.4$.

For any $X \geq 1$, combining (9.15) with (4.9) gives

$$\begin{aligned}
&\frac{\partial}{\partial X}\left(m_- X^{m_- -1}\left(\frac{m_- -1}{\gamma r m_+(m_+-1)(m_+-m_-)}X^{1-m_+} + \frac{\beta - 2r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln(X)}{\gamma r}\right)\right)\\
&= \frac{m_- X^{m_- -2}}{\gamma r}\left((1-m_-)\left(\frac{X^{1-m_+}}{m_+(m_+-1)} + \ln(X) + \frac{m_+ m_- - 2m_+ - 2m_- + 3}{(m_+-1)(m_- -1)}\right) - 1\right) < 0,\\
&\frac{\partial}{\partial X}\left(X^{m_- -1}\left(x + \frac{Y}{r} + \frac{m_- -1}{\gamma r(m_+-1)(m_+-m_-)}X^{1-m_+} + \frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln(X)}{\gamma r}\right)\right)\\
&= (m_- - 1)X^{m_- -2}\left(-C_2 m_- A(\underline{L})^{m_- -1}\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)^{1-m_-}\right) + \frac{X^{m_- -2}}{\gamma r}\left(\frac{1-m_-}{m_+-m_-}X^{1-m_+} - 1\right) < 0.
\end{aligned}$$

Hence, differentiating (9.15) with respect to $\underline{L}$ implies $sgn\left(\frac{\partial}{\partial \underline{L}}\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)\right) = sgn\left(\frac{\partial}{\partial \underline{L}}\left(\frac{A(\underline{L})}{\bar{y}}\right)\right)$.

If $\bar{y} \in [A(\overline{L}), A(\underline{L}))$, then differentiating (6.8) with respect to $\underline{L}$ implies

$$0 = \frac{\left(\frac{A(\underline{L})}{\bar{y}}\right)^{-1}}{\gamma m_+(m_+-1)}\left(\frac{A(\overline{L})}{A(\underline{L})}\left(\frac{A(\underline{L})}{\bar{y}}\right)^{m_+}(m_+-1) + 1 - m_+\left(\frac{A(\underline{L})}{\bar{y}}\right)^{m_+-1}\right)\frac{\partial}{\partial \underline{L}}\left(\frac{A(\underline{L})}{\bar{y}}\right) + \frac{\left(\frac{A(\underline{L})}{\bar{y}}\right)^{m_+}\frac{A(\overline{L})}{A(\underline{L})}(1-\alpha)}{m_+\alpha},$$

where the equality is derived by eliminating Y. Since

$$Z \mapsto \frac{A(\overline{L})}{A(\underline{L})}\left(\frac{A(\underline{L})}{Z}\right)^{m_+}(m_+-1) + 1 - m_+\left(\frac{A(\underline{L})}{Z}\right)^{m_+-1} =: G(Z)$$

is strictly decreasing in $Z \in [A(\overline{L}), A(\underline{L}))$ with $\lim_{Z \uparrow A(\underline{L})} G(Z) < 0$, we get $\frac{\partial c_*^{\text{pre}}(x)}{\partial \underline{L}} > 0$. Together with (9.14), we conclude $\frac{\partial \pi_*^{\text{pre}}(x)}{\partial \underline{L}} < 0$.

If $0 < \bar{y} < A(\overline{L})$, on the other hand, then differentiating (6.8) with respect to $\underline{L}$ implies

$$0 = \left(\frac{A(\underline{L})}{\bar{y}}\right)^{m_+-2}\left(Y - \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L})\right)\frac{\partial}{\partial \underline{L}}\left(\frac{A(\underline{L})}{\bar{y}}\right) + \frac{1-\alpha}{\alpha(m_+-1)}\left(\left(\frac{A(\underline{L})}{\bar{y}}\right)^{m_+-1} - \left(\frac{A(\underline{L})}{A(\overline{L})}\right)^{m_+-1}\right).$$

Due to $m_+ > 1$, $A(\overline{L}) < A(\underline{L})$, and $Y < \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L})$, the above equality implies $\frac{\partial c_*^{\text{pre}}(x)}{\partial \underline{L}} > 0$. Together with (9.14), we conclude $\frac{\partial \pi_*^{\text{pre}}(x)}{\partial \underline{L}} < 0$. □

Remark 9.2 (*Economic Interpretations*)**.** We provide economic interpretations of Proposition 9.2 as follows:

The timing of retirement is influenced by the interplay between leisure and income. When income or leisure before retirement increases, the agent finds working more attractive and tends to delay retirement, opting to enjoy the benefits of continued income and utility from leisure while still employed. Conversely, if leisure after retirement becomes more desirable, the agent has a stronger incentive to retire earlier to maximize time spent in this enhanced state of leisure. This balance between pre- and post-retirement utilities drives the optimal retirement decision.

Before retirement, the agent's optimal consumption and portfolio choices are independent of leisure and income levels. This independence indicates that these financial decisions are shaped more by preferences for risk and consumption timing than by external factors like the agent's income or leisure availability. The agent's focus is on achieving an optimal financial trajectory, which remains unaffected by variations in income or leisure (see Fig. 5).

Income and leisure exert nuanced effects on pre-retirement behavior. An increase in income leads to higher consumption, as the agent can afford to allocate more resources to present utility. Greater leisure before retirement motivates the agent to reduce stock investments, prioritizing time spent on leisure activities, which indirectly increases consumption. However, when leisure after

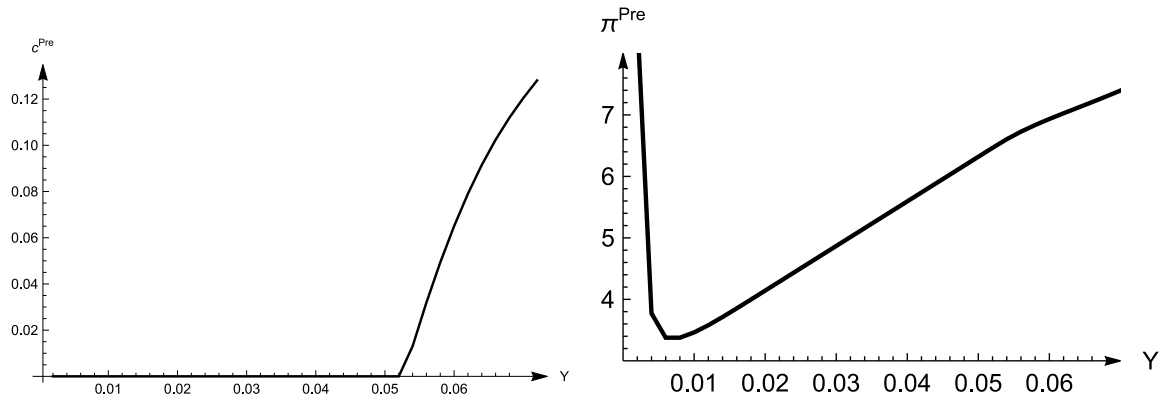


Fig. 3. The two graphs describe optimal consumption and portfolio of pre-retirement as a function of Y. The parameters are $\beta = r = 0.0371$, $\mu = 0.1123$, $\sigma = 0.1954$, $\gamma = 2$, $\underline{L} = 0.05$, $\overline{L} = 0.1$, $\alpha = 0.4$, $x = 1$.

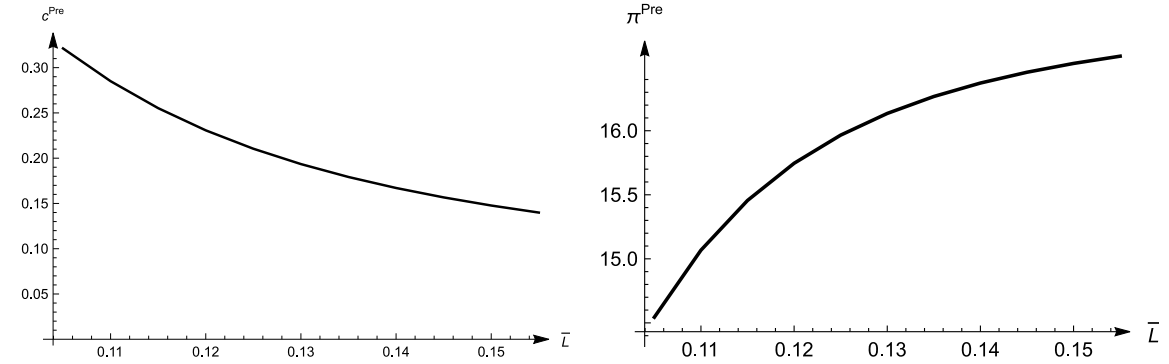


Fig. 4. The two graphs describe optimal consumption and portfolio of pre-retirement as a function of $\overline{L}$. The parameters are $\beta = r = 0.0371$, $\mu = 0.1123$, $\sigma = 0.1954$, $\gamma = 2$, $\underline{L} = 0.05$, $Y = 0.05$, $\alpha = 0.4$, $x = 5$.

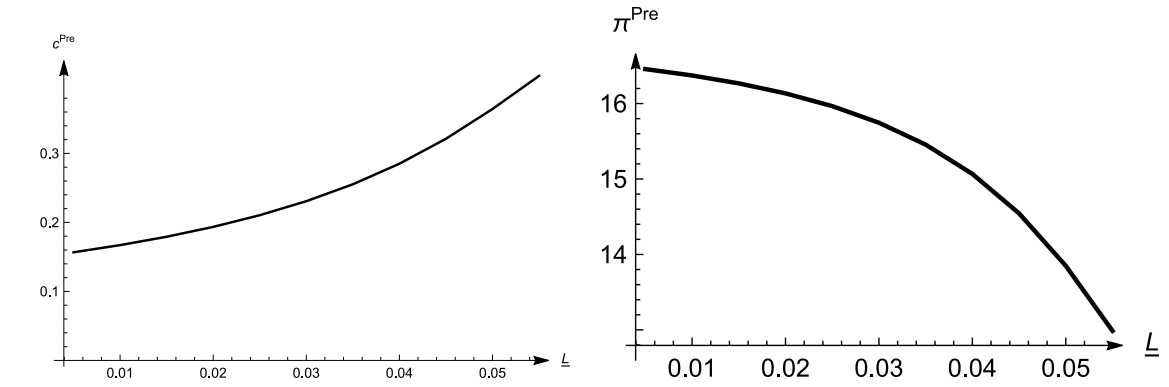


Fig. 5. The two graphs describe optimal consumption and portfolio of pre-retirement as a function of $\underline{L}$. The parameters are $\beta = r = 0.0371$, $\mu = 0.1123$, $\sigma = 0.1954$, $\gamma = 2$, $\overline{L} = 0.1$, $Y = 0.05$, $\alpha = 0.4$, $x = 5$.

retirement becomes more significant, the agent adjusts by investing more in stocks to secure higher returns for the future, which leads to a reduction in current consumption. These adjustments reflect the agents strategy to balance present and future utility optimally (see Fig. 4).

Remark 9.3. Eq. (9.13) shows that

$$\frac{\partial}{\partial Y}\left(I_{\text{pre}}(x)\tilde{V}''_{\text{pre}}(I_{\text{pre}}(x))\right) = \frac{1-m_-}{\gamma r}\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)^{-1}\left(\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right)^{1-m_+} - 1\right)\frac{\partial}{\partial Y}\left(\frac{A(\underline{L})}{I_{\text{pre}}(x)}\right) + \frac{1-m_-}{r},$$

while $\frac{\partial c_*^{\text{pre}}(x)}{\partial Y} > 0$ holds for $I_{\text{pre}}(x) < A(\underline{L})$, the term $\frac{1-m_-}{r}$ can potentially disrupt monotonicity. This nuanced behavior is illustrated in Fig. 3.

10. Conclusions

This paper investigates an optimal control problem involving investment, consumption, and retirement decisions under exponential (CARA-type) utility. Our main contribution lies in analyzing the equivalent condition for no retirement and examining how income Y and the two leisure levels, $\underline{L}$ and $\overline{L}$, influence optimal portfolio, consumption, and retirement decisions. Specifically, the condition for no retirement is given by $Y \geq \frac{1-\alpha}{\alpha}(\overline{L} - \underline{L})$, which defines the threshold income level needed to avoid retirement. The study reveals that the optimal portfolio and consumption decisions before retirement are independent of both income and the two leisure levels. In contrast, post-retirement decisions for the optimal portfolio and consumption demonstrate opposing monotonic relationships with leisure. Specifically, for a given x, $c_*^{\text{pre}}(x)$ is non-decreasing in $\underline{L}$ and non-increasing in $\overline{L}$, while $\pi_*^{\text{pre}}(x)$ shows the opposite trend, being non-increasing in $\underline{L}$ and non-decreasing in $\overline{L}$. These results underscore the distinct effects that leisure before and after retirement have on the agent's financial strategies. While $c_*^{\text{pre}}(x)$ is shown to increase with income Y, the monotonic behavior of $\pi_*^{\text{pre}}(x)$ with respect to Y remains indeterminate. This ambiguity highlights the nuanced interplay between income and investment decisions, reflecting the complex nature of optimal control in this setting.

While our model assumes a constant labor income stream, real-world income is often subject to various risks, such as automation, climate change, and economic fluctuations. Future research could extend our framework by incorporating stochastic labor income to analyze its impact on optimal retirement and investment decisions. In particular, adapting the dual/martingale approach to accommodate income volatility would provide a richer framework for understanding how individuals adjust their financial and retirement strategies under uncertainty. Moreover, linking this analysis to broader studies on optimal consumption and portfolio choices under income uncertainty would enhance our understanding of retirement decision-making in more realistic economic settings.

These findings provide valuable insights into how individuals can balance their financial and leisure preferences across different life stages while offering a foundation for future studies that integrate income uncertainty into retirement planning models.

Acknowledgments

We appreciate the helpful comments and valuable suggestions from two anonymous referees, which have been essential in improving our paper.

Appendix. Subsidiary lemmas with proofs

Lemma A.1.
(i) $\tilde{V}_{\text{post}} < 0$, $\tilde{V}'_{\text{post}} < 0$, $\tilde{V}''_{\text{post}} > 0$.
(ii) $\tilde{V}_{\text{post}} \in C^3((0,\infty))$.
(iii) $\sup_{y>0}(|\tilde{V}_{\text{post}}(y)| + |y\tilde{V}'_{\text{post}}(y)| + |y^2\tilde{V}''_{\text{post}}(y)|) < \infty$.

Proof.
(i) Let us consider the function $h(y) := \frac{m_- - 1}{\gamma r(m_+ - m_-)}\left(\frac{A(\overline{L})}{y}\right)^{1-m_+} + \frac{1}{\gamma r}$, which is strictly decreasing in y, with $h(0) = \frac{1}{\gamma r} > 0$ and $h(A(\overline{L})) = \frac{m_+ - 1}{\gamma r(m_+ - m_-)} > 0$. From this, we deduce that $\tilde{V}''_{\text{post}}(y) > 0$ for all $y > 0$. Combining this with the fact that $\tilde{V}'_{\text{post}}(A(\overline{L})) < 0$ and using (4.10), we conclude $\tilde{V}'_{\text{post}}(y) < 0$ for all $y > 0$. Furthermore, since $\lim_{y\downarrow 0}\tilde{V}_{\text{post}}(y) = 0$ and $\tilde{V}'_{\text{post}}(y) < 0$ for all $y > 0$, it follows that $\tilde{V}_{\text{post}}(y) < 0$ for all $y > 0$.
(ii) It is straightforward to verify that $\tilde{V}_{\text{post}} \in C^2((0,\infty))$. Due to (4.8), $\tilde{V}_{\text{post}} \in C^2((0,\infty))$, and $\overline{U} \in C^1((0,\infty))$, we conclude that $\tilde{V}_{\text{post}} \in C^3((0,\infty))$.
(iii) The finiteness of the supremum is a direct result of the boundedness of the terms. □

Lemma A.2. *For the optimal total wealth* $X_t^{\text{post}} = -\tilde{V}'_{\text{post}}(y_t)$, *we have* $\lim_{t\to\infty}\mathbb{E}^x\left[H_t X_t^{\text{post}}\right] = 0$ *and* $\lim_{t\to\infty}\mathbb{E}^x[e^{-\beta t}\tilde{V}_{\text{post}}(y_t)] = 0$.

Proof. Let $y = (-\tilde{V}'_{\text{post}})^{-1}(x)$, where $(\cdot)^{-1}$ denotes the inverse function, which exists and is well-defined by Lemma A.1. Using this, we have

$$\mathbb{E}^x\left[H_t X_t^{\text{post}}\right] = \mathbb{E}^x\left[-H_t\tilde{V}'_{\text{post}}(y_t)\right] \leq e^{-\beta t}y^{-1}\sup_{z>0}|z\tilde{V}'_{\text{post}}(z)| \xrightarrow{t\to\infty} 0,$$

$$\mathbb{E}^x[e^{-\beta t}\tilde{V}_{\text{post}}(y_t)] \le e^{-\beta t}\sup_{z>0}|\tilde{V}_{\text{post}}(z)| \xrightarrow{t\to\infty} 0.$$

Both limits rely on the finiteness of $|\tilde{V}_{\text{post}}(z)|$, $|z\tilde{V}'_{\text{post}}(z)|$, and properties established in Lemma A.1. □

Lemma A.3. *Let $\tau = \inf\{s > 0 : y_s \le \overline{y}\}$ be a maximizer for (5.6). Then*
(i) $\tilde{V}_{\text{pre}}(y) \le \frac{Y}{r}y$, $\tilde{V}'_{\text{pre}}(y) \le \frac{Y}{r}$, $\tilde{V}''_{\text{pre}}(y) \ge 0$ for all $y > 0$. Equalities become strict if $y > \overline{y}$.
(ii) $\tilde{V}_{\text{pre}} \in C((0,\infty)) \cap C^2((0,\infty)\setminus\{\overline{y}\})$.
(iii) $\sup_{y>0}\left(|\tilde{V}_{\text{pre}}(y)| + |y\tilde{V}'_{\text{pre}}(y)| + |y^2\tilde{V}''_{\text{pre}}(y)|\right) < \infty$.

Proof.
(i) Let us define $\tilde{v}_{\text{pre}}(y) := \tilde{V}_{\text{pre}}(y) - \frac{Y}{r}y$. Then $\tilde{v}_{\text{pre}}$ satisfies

$$\begin{cases} \frac{\theta^2 y^2}{2}\cdot\tilde{v}''_{\text{pre}}(y) + (\beta - r)y\tilde{v}'_{\text{pre}}(y) - \beta\tilde{v}_{\text{pre}}(y) + \underline{U}(y) = 0, & \text{for } y > \overline{y}, \\ \tilde{v}_{\text{pre}}(y) = 0, & \text{for } 0 < y \le \overline{y}. \end{cases} \tag{A.1}$$

Thus we have

$$\begin{aligned} \tilde{V}_{\text{pre}}(y) &= \frac{Y}{r}y + \begin{cases} C_1 y^{m_+} + C_2 y^{m_-} + \left(\frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} - \frac{\ln(A(\underline{L}))+1}{\gamma r}\right)y + \frac{y\ln(y)}{\gamma r}, & \text{if } \overline{y} < y \le A(\underline{L}), \\ C_3 y^{m_-} - \frac{A(\underline{L})}{\gamma\beta}, & \text{if } y > A(\underline{L}), \end{cases} \\ \tilde{V}'_{\text{pre}}(y) &= \frac{Y}{r} + \begin{cases} C_1 m_+ y^{m_+-1} + C_2 m_- y^{m_--1} + \frac{\beta - r + \frac{\theta^2}{2}}{\gamma r^2} + \frac{\ln\left(\frac{y}{A(\underline{L})}\right)}{\gamma r}, & \text{if } \overline{y} < y \le A(\underline{L}), \\ C_3 m_- y^{m_--1}, & \text{if } y > A(\underline{L}), \end{cases} \\ y\tilde{V}''_{\text{pre}}(y) &= \begin{cases} C_1 m_+(m_+ - 1)y^{m_+-1} + C_2 m_-(m_- - 1)y^{m_--1} + \frac{1}{\gamma r}, & \text{if } \overline{y} < y \le A(\underline{L}), \\ C_3 m_-(m_- - 1)y^{m_--1}, & \text{if } y > A(\underline{L}). \end{cases} \end{aligned} \tag{A.2}$$

The coefficients C_1, C_2, and C_3 are determined by ensuring $\tilde{V}_{\text{pre}} \in C^1((\overline{y},\infty))$ and $\tilde{V}_{\text{pre}}$ is continuous at $y = \overline{y}$:

$$\begin{aligned} C_1 &= \frac{m_- - 1}{\gamma r m_+(m_+ - 1)(m_+ - m_-)A(\underline{L})^{m_+-1}} < 0, \\ C_2 &= \frac{r\ln\left(\frac{A(\underline{L})}{\overline{y}}\right) - \left(\beta - 2r + \frac{\theta^2}{2}\right)}{\gamma r^2\overline{y}^{m_--1}} - C_1\overline{y}^{m_+-m_-} > 0, \\ C_3 &= C_2 + \frac{m_+ - 1}{\gamma r m_-(m_- - 1)(m_+ - m_-)A(\underline{L})^{m_--1}} > \frac{m_+ - 1}{\gamma r m_-(m_- - 1)(m_+ - m_-)A(\underline{L})^{m_--1}} > 0. \end{aligned} \tag{A.3}$$

Here, the proof of all signs in (A.3) is complete if $C_2 > 0$ is shown, which can be easily derived from $\tilde{V}_{\text{Mer}}(y) < \tilde{V}_{\text{pre}}(y)$ for $y > \overline{y}$ by considering the stochastic representation (refer to Section 7 for the form of $\tilde{V}_{\text{Mer}}$). The function $k(y) := C_1 m_+(m_+ - 1)y^{m_+-1} + C_2 m_-(m_- - 1)y^{m_--1} + \frac{1}{\gamma r}$ satisfies $k' < 0$ with $k(A(\underline{L})) = \frac{m_+ - 1}{\gamma r(m_+ - m_-)} + C_2 m_-(m_- - 1)A(\underline{L})^{m_--1} > 0$ due to $C_2 > 0$. This result and $C_3 > 0$ imply $\tilde{V}''_{\text{pre}}(y) > 0$ for $y > \overline{y}$. Also, combining the convexity with $C_3 > 0$ and $\tilde{v}'_{\text{pre}}(A(\underline{L})) = \tilde{V}'_{\text{pre}}(A(\underline{L})) - \frac{Y}{r} = C_3 m_- A(\underline{L})^{m_--1} < 0$ gives $\tilde{v}'_{\text{pre}}(y) = \tilde{V}'_{\text{pre}}(y) - \frac{Y}{r} < 0$ for $y > \overline{y}$. Similarly, combining this result with $\tilde{v}_{\text{pre}}(\overline{y}) = \tilde{V}_{\text{pre}}(\overline{y}) - \frac{Y}{r}\overline{y} = 0$ gives $\tilde{v}_{\text{pre}}(y) = \tilde{V}_{\text{pre}}(y) - \frac{Y}{r}y < 0$ for $y > \overline{y}$, that is, $\tilde{V}_{\text{pre}}(y) < \frac{Y}{r}y$ for $y > \overline{y}$.
(ii) From (A.1), the fact that $\tilde{V}_{\text{pre}} \in C^1((\overline{y},\infty))$, and $\underline{U} \in C^1((\overline{y},\infty))$, we conclude that $\tilde{V}_{\text{pre}} \in C^2((\overline{y},\infty))$. $\tilde{V}_{\text{pre}} \in C((0,\infty))$ is already shown by determining C_1, C_2, C_3.
(iii) The finiteness of the supremum is a direct result of the boundedness of the terms. □

Data availability

No data was used for the research described in the article.

References

[1] R.C. Merton, Lifetime portfolio selection under uncertainty: The continuous-time case, Rev. Econ. Stat. 51 (3) (1969) 247–257, http://dx.doi.org/10.2307/1926560.
[2] R.C. Merton, Optimum consumption and portfolio rules in a continuous-time model, J. Econom. Theory 3 (4) (1971) 373–413, http://dx.doi.org/10.1016/0022-0531(71)90038-X.
[3] K.J. Choi, G. Shim, Disutility, optimal retirement, and portfolio selection, Math. Finance 16 (2) (2006) 443–467, http://dx.doi.org/10.1111/j.1467-9965.2006.00278.x.
[4] E. Farhi, S. Panageas, Saving and investing for early retirement: A theoretical analysis, J. Financ. Econ. 83 (1) (2007) 87–121, http://dx.doi.org/10.1016/j.jfineco.2005.10.004.

[5] K.J. Choi, G. Shim, Y.H. Shin, Optimal portfolio, consumption-leisure and retirement choice problem with CES utility, Math. Finance 18 (3) (2008) 445–472, http://dx.doi.org/10.1111/j.1467-9965.2008.00341.x.
[6] P.H. Dybvig, H. Liu, Lifetime consumption and investment: Retirement and constrained borrowing, J. Econom. Theory 145 (3) (2010) 885–907, http://dx.doi.org/10.1016/j.jet.2009.08.003.
[7] I. Karatzas, J.P. Lehoczky, S.P. Sethi, S.E. Shreve, Explicit solution of a general consumption/investment problem, Math. Oper. Res. 11 (2) (1986) 261–294, http://dx.doi.org/10.1287/moor.11.2.261.
[8] I. Karatzas, J.P. Lehoczky, S.E. Shreve, Optimal portfolio and consumption decisions for a "small investor" on a finite horizon, SIAM J. Control Optim. 25 (6) (1987) 1557–1586, http://dx.doi.org/10.1137/0325086.
[9] J.C. Cox, C.F. Huang, Optimal consumption and portfolio policies when asset prices follow a diffusion process, J. Econom. Theory 49 (1) (1989) 33–83, http://dx.doi.org/10.1016/0022-0531(89)90067-7.
[10] I. Karatzas, H. Wang, Utility maximization with discretionary stopping, SIAM J. Control Optim. 39 (1) (2000) 306–329, http://dx.doi.org/10.1137/S0363012998346323.
[11] D. Duffie, W. Fleming, H.M. Soner, T. Zariphopoulou, Hedging in incomplete markets with HARA utility, J. Econom. Dynam. Control 21 (4–5) (1997) 753–782, http://dx.doi.org/10.1016/S0165-1889(97)00002-X.
[12] J.F. Cocco, F.J. Gomes, P.J. Maenhout, Consumption and portfolio choice over the life cycle, Rev. Financ. Stud. 18 (2) (2005) 491–533, http://dx.doi.org/10.1093/rfs/hhi017.
[13] A. Bensoussan, B.-G. Jang, S. Park, Unemployment risks and optimal retirement in an incomplete market, Oper. Res. 64 (4) (2016) 1015–1032, http://dx.doi.org/10.1287/opre.2016.1503.
[14] S. Park, Income disaster model with optimal consumption, Econom. Theory (2025) http://dx.doi.org/10.1007/s00199-024-01629-x, in press.
[15] E. Barucci, D. Marazzina, Optimal investment, stochastic labor income and retirement, Appl. Math. Comput. 218 (9) (2012) 5588–5604, http://dx.doi.org/10.1016/j.amc.2011.11.052.
[16] K. Park, H.Y. Wong, Robust retirement with return ambiguity: optimal G-stopping time in dual space, SIAM J. Control Optim. 61 (3) (2023) 1009–1037, http://dx.doi.org/10.1137/21M1459150.
[17] G. Ding, D. Marazzina, Bankruptcy and retirement: a comparison in an optimal stopping times ordered framework, Comput. Appl. Math. 43 (2024) 53, http://dx.doi.org/10.1007/s40314-023-02566-6.
[18] M. Jeanblanc, P. Lakner, A. Kadam, Optimal bankruptcy time and consumption/investment policies on an infinite horizon with a continuous debt repayment until bankruptcy, Math. Oper. Res. 29 (3) (2004) 649–671, http://dx.doi.org/10.1287/moor.1030.0084.
[19] H.K. Koo, K.H. Roh, Y.H. Shin, Optimal consumption/investment and retirement with necessities and luxuries, Math. Methods Oper. Res. 94 (2) (2021) 281–317, http://dx.doi.org/10.1007/s00186-021-00758-6.
[20] B. Oksendal, Stochastic Differential Equations: An Introduction with Applications, fifth ed., Springer, Berlin, 2003.

Finance Research Letters 89 (2026) 109176

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Household consumption, financial constraints, and size structure of rent☆

Hyun-Tak Lee [a], Bong-Gyu Jang [b,*], Kum-Hwan Roh [c], Yong Hyun Shin [d]

[a] College of Business, Gachon University, Seongnam-si 13120, Republic of Korea
[b] Department of Industrial and Management Engineering, POSTECH, Pohang, Gyeongbuk 37673, Republic of Korea
[c] Department of Mathematics, Hannam University, Ojeong-dong Daedeok-gu, Daejeon 34430, Republic of Korea
[d] Department of Mathematics, Sookmyung Women's University, Seoul 04310, Republic of Korea

ARTICLE INFO

Keywords:
Household finance
Consumption-housing choice
House rent
Financial heterogeneity
Precautionary savings

ABSTRACT

This paper examines how financial constraints shape household consumption and housing choices in rental markets with a non-linear price structure. We show that increases in the elasticity magnitude generate asymmetric behavioral responses. Heightened background risk uniformly lowers consumption across households. Housing responses diverge: financially unconstrained households expand to capture stronger economies of scale (price signal), whereas constrained households downsize to avoid amplified penalties for suboptimal choices (structural signal). This mechanism identifies the market's price structure as a driver of welfare inequality. Policies that reduce the elasticity magnitude and thereby linearize rents raise aggregate welfare and compress disparities by easing structural penalties that bind constrained households.

1. Introduction

How do households with differing financial constraints adjust their consumption and housing choices in rental markets? The literature offers two parallel but disconnected explanations. On one hand, non-linear rent structures create economies of scale (Leung et al., 2022), thereby providing a clear *price incentive* for households to upsize. On the other hand, a separate stream of research shows that financial constraints act as a *structural penalty* (Campbell and Cocco, 2007; Landvoigt et al., 2015). The tension between this incentive and penalty creates varied behavioral responses that representative-agent models do not capture.

To connect these two ideas, we develop an economic model where the rent–size elasticity sends a dual signal. This mechanism is distinct from household hedging against random rent volatility (Sinai and Souleles, 2005), as the risk in our model comes from the deterministic price structure itself. A larger elasticity magnitude increases background risk and creates a greater penalty for suboptimal choices. This strengthens precautionary savings and lowers consumption for all households. Housing choices, however, adjust asymmetrically. For financially unconstrained households, the elasticity acts as a price signal that encourages them to consume more housing. For financially constrained households, it becomes a structural signal of higher risk that prompts them to downsize.

Our work offers two main contributions. First, departing from representative-agent models, we show that financial heterogeneity – driven by households' varying exposure to housing market conditions – is essential for disentangling the dual role of rent–size

☆ We would like to thank Hyeng Keun Koo, Byung Hwa Lim, Seyoung Park, and all participants at Sookmyung-TMU Mathematical Finance Workshop with Young Researchers. All errors and omissions are our own. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea, Republic of Korea (RS-2023-NR077018).

* Corresponding author.
E-mail addresses: hyuntaklee@gachon.ac.kr (H.-T. Lee), bonggyujang@postech.ac.kr (B.-G. Jang), khroh@hnu.kr (K.-H. Roh), yhshin@sookmyung.ac.kr (Y.H. Shin).

https://doi.org/10.1016/j.frl.2025.109176
Received 11 September 2025; Received in revised form 23 November 2025; Accepted 1 December 2025
Available online 12 December 2025
1544-6123/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

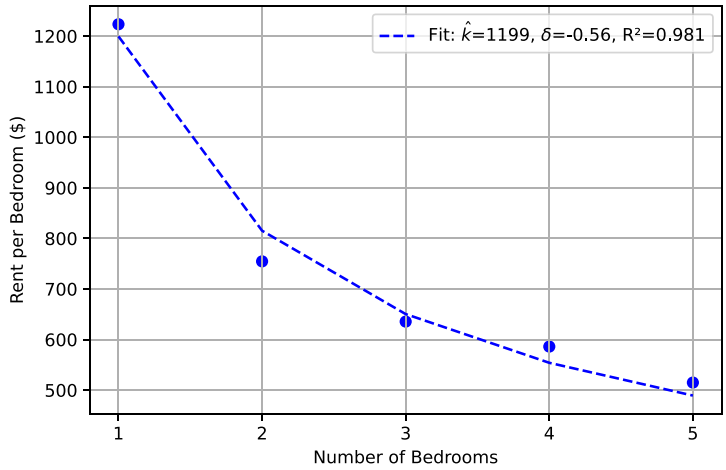


Fig. 1. Estimated size structure of average rent per bedroom in the United States.

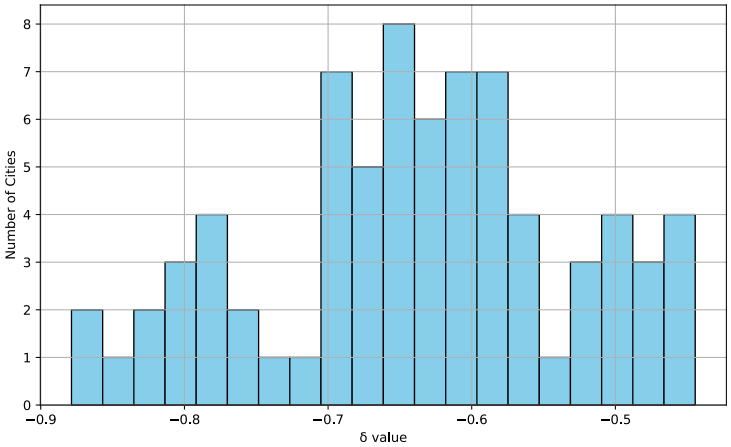


Fig. 2. Distribution of the estimated curvature parameter δ across U.S. cities.

elasticity. This elasticity acts as both a price incentive and a structural penalty, explaining how a uniform consumption decline can coexist with divergent housing adjustments. Such patterns cannot arise in representative-agent frameworks that mute constraint-driven heterogeneity. Second, our analysis provides clear policy guidance. We find that flattening the rent–size structure raises aggregate consumption and reduces welfare inequality. This policy works by easing the structural penalties on financially vulnerable households.

2. Motivation

The U.S. rental market exhibits a clear non-linear price structure. Data from Rentable (March 29, 2025) show that as dwelling size increases, the average rent per bedroom declines at a diminishing rate (Fig. 1). Larger units are cheaper on a per-unit basis — a pattern of economies of scale central to our model.

We fit this per-unit rent curve to a power function,

$$g(h) = \hat{k} \cdot h^{\delta},$$

where h is the number of bedrooms (a proxy for housing size) and the parameter δ captures the curvature of this relationship. The estimation yields $\hat{k} = 1199$ and $\delta = -0.56$ with high explanatory power ($R^2 = 0.981$). The strong fit and tractability of the power-law form make it suitable for our dynamic programming model.

Our model integrates this non-linear size structure of rent into a household's dynamic choice problem. This is critical because the non-homothetic nature of housing demand means such non-linearities disproportionately affect households at different income levels (Albouy et al., 2016), motivating our departure from a representative-agent framework.

The distribution of the estimated δ across U.S. cities (Fig. 2) confirms that this non-linear pattern is a general structural feature at the city level, not just a national aggregate. This evidence highlights the importance of incorporating a non-linear size structure of rent into models of household rental decisions.

While our model is calibrated with U.S. data, such a non-linear rent structure has been observed in other housing markets. For example, a study of Vienna's rental market documents non-linear effects of apartment size on rents (Brunauer et al., 2010). Similarly, research in Hong Kong's subdivided-unit market finds size-sensitive demand and higher rent burdens for smaller units (Leung et al., 2022). These examples suggest that our model's core assumption of rental economies of scale has broad relevance.

3. The model

3.1. House prices and size structure of rent

We consider a financial market with a single risk-free asset to focus on the rental market of our interest. The house price index, w_t, follows a geometric Brownian motion:

$$dw_t = \mu w_t dt + \sigma w_t dB_t, \quad w_0 = w > 0, \tag{1}$$

where μ is the average rate of price growth, σ is its volatility, and B_t is a standard Brownian motion. The tenant is assumed to be a price-taker in partial equilibrium, so choices do not affect the house price index.

Our central specification is the rental price ϵ_t, a function of the house price index w_t and housing size $h_t > 0$:

$$\epsilon(h_t, w_t) := h_t \cdot g(h_t) \cdot w_t,$$

where we assume the *size structure* function $g(h) := k \cdot h^{\delta}$.[1] The parameter k is the rent level for a unit-sized house. The key parameter δ is the elasticity of the marginal rent with respect to housing size, which captures the market's price structure.[2]

The sign of δ determines the nature of this structure. We focus on $\delta \leq 0$, consistent with the economies of scale in Fig. 2, where the marginal cost of housing decreases with size. The magnitude $|\delta|$ governs the intensity of this non-linearity. We assume $-1 < \delta \leq 0$ to ensure a well-posed optimization problem.

3.2. Agent's problem

A forward-looking tenant maximizes expected lifetime utility from non-housing consumption ($c_t > 0$) and housing services ($h_t > 0$) over an infinite horizon.[3] The tenant has Cobb–Douglas preferences:

$$V(x, w) = \max_{(c,h)} \mathbb{E}\left[\int_0^{\infty} e^{-\beta t} \frac{\left(c_t^{\alpha} h_t^{1-\alpha}\right)^{1-\gamma^*}}{1-\gamma^*} dt\right], \tag{2}$$

where β is the subjective discount rate, α ($0 < \alpha < 1$) is the preference weight on consumption, and $\gamma^* > 1$ is the coefficient of relative risk aversion.

Financial wealth X_t evolves according to the dynamic budget constraint:

$$dX_t = (rX_t - c_t - \epsilon(h_t, w_t) + I)dt, \tag{3}$$

where r is the risk-free rate and I ($I \geq 0$) is constant labor income. We assume the initial financial wealth $X_0 = x > -I/r$ to ensure positive total wealth defined as financial wealth plus human capital.[4]

4. Theoretical analysis

The problem yields closed-form solutions for consumption and housing choices:

$$\begin{cases} c^* = & K\left(x + \frac{I}{r}\right), \\ h^* = & \left(\frac{K}{k \cdot M}\right)^{\frac{1}{1+\delta}} \left(\frac{x + \frac{I}{r}}{w}\right)^{\frac{1}{1+\delta}}, \end{cases}$$

where K is the *marginal propensity to consume* (MPC) out of total wealth, and M is a composite parameter. Crucially, M is linear in the elasticity of marginal rent δ. Hence the MPC K, and therefore optimal consumption c^*, depend on δ.[5]

These equations reveal a structural difference between optimal consumption and housing choices. Optimal consumption c^* is a constant fraction K of total wealth $(x + I/r)$, independent of current housing market conditions w, and depends on δ only through the state-independent MPC K. By contrast, optimal housing h^* depends on the ratio of house prices to total wealth, $w/(x + I/r)$, and depends on δ through three channels: the MPC (K), the composite parameter (M), and the non-linear exponent $1/(1+\delta)$. This structural difference foreshadows asymmetric responses to changes in the elasticity magnitude.

[1] We can employ other functions as the size structure function. In this paper, we take one of the most parsimonious functions.

[2] See Appendix A for details.

[3] We adopt an infinite-horizon dynasty setup. Since a finite-horizon model with an exponential death rate behaves identically to this formulation in both its trends and magnitude (Liu and Loewenstein, 2002; Jang et al., 2013), our findings are directly applicable.

[4] See Appendix B for a set of technical assumptions.

[5] See Appendix B for the full derivation from the HJB equation (Theorem B.1).

Table 1
Model calibration.

Parameter	Value	Source/justification
A. Preference Parameters		
Subjective Discount Rate (β)	0.04	Assumption
Risk Aversion (γ^*)	2.0	Benchmark value
Consumption Weight (α)	0.7	U.S. Consumer Expenditure Survey (BLS)
B. Economic Environment		
Real Risk-Free Rate (r)	0.02	Assumption
House Price Growth (μ)	0.015	S&P Case–Shiller long-term real average
House Price Volatility (σ)	0.08	S&P Case–Shiller long-term volatility
C. Rent Structure		
Base Rental Yield (annualized) (k)	0.04	Assumption
Elasticity of Marginal Rent (δ)	$[-0.9, 0.0]$	Avoid the limit case of $\delta \to -1$.

Note: The base rental yield k is normalized. Its relationship to the estimated $\hat{k}$ from Section 2 is $\hat{k} = k\omega$, where ω is the house price.

5. Economic implications

5.1. Calibration

We conduct a quantitative experiment using a calibration based on standard literature values and stylized facts of the U.S. economy, as summarized in Table 1. Among these, a central parameter is the elasticity of marginal rent, δ. While the model is solved for a range of negative δ values, we report results by the elasticity magnitude, $|\delta|$, to focus on the strength of effects. This isolates the novel mechanisms driven by the rent structure's sensitivity. Another key feature of the model is that optimal choices depend on the ratio of house prices to total wealth (HPTW), which is used as a measure of financial constraint. We normalize total wealth to one, so the house price index ω is directly interpretable as HPTW. We examine results across a range of HPTW values to represent diverse financial situations.

5.2. Main results

Our numerical analysis reveals a striking asymmetry in how consumption and housing choices respond to changes in the elasticity magnitude, $|\delta|$.[6]

Fig. 3 highlights two results. First, the consumption response is uniform. Panel A shows that K decreases monotonically with $|\delta|$ for all households, regardless of financial constraints. Second, the housing response is asymmetric and conditional on constraints. Panel B shows that as $|\delta|$ rises, financially unconstrained (low HPTW) households expand housing, whereas constrained households (high HPTW) downsize. For any fixed market structure (given k and $|\delta|$), more constrained households also choose smaller housing — the standard affordability channel.

This combination of a uniform consumption response and an asymmetric housing response is the central puzzle of the paper.

5.3. Asymmetric housing responses and financial constraint

The puzzle arises from the dual role of the elasticity parameter δ. In a concave market ($\partial^2\epsilon/\partial h^2 < 0$ for $\delta < 0$), the elasticity magnitude $|\delta|$ transmits two distinct signals: a *price signal* and a *structural signal*. A household's financial constraint determines which signal dominates its choice.

First, the price signal is the bulk discount incentive. As $|\delta|$ increases, economies of scale strengthen, encouraging households who can afford it to consume larger housing.

Second, the structural signal is the financial penalty for suboptimal choices. This is best understood through the per-unit rent (PUR), defined as

$$\text{PUR} = \epsilon/h = k\omega h^{\delta}.$$

As $|\delta|$ rises, PUR becomes highly sensitive to deviations from the average housing size ($h = 1$), creating a penalty zone for downsizing ($h < 1$) and a reward zone for upsizing ($h > 1$). Table 2 illustrates this effect. For a 50% downsizing, moving from $\delta = -0.3$ to $\delta = -0.8$ raises PUR by 74% instead of 23%, sharply magnifying the penalty.

This dual-signal framework explains the asymmetric housing response:

[6] Our qualitative findings are robust to alternative parameter specifications. We confirm that the asymmetric housing response holds for different values of house price volatility (σ) and risk aversion (γ^*). See Appendix D for the detailed analysis.

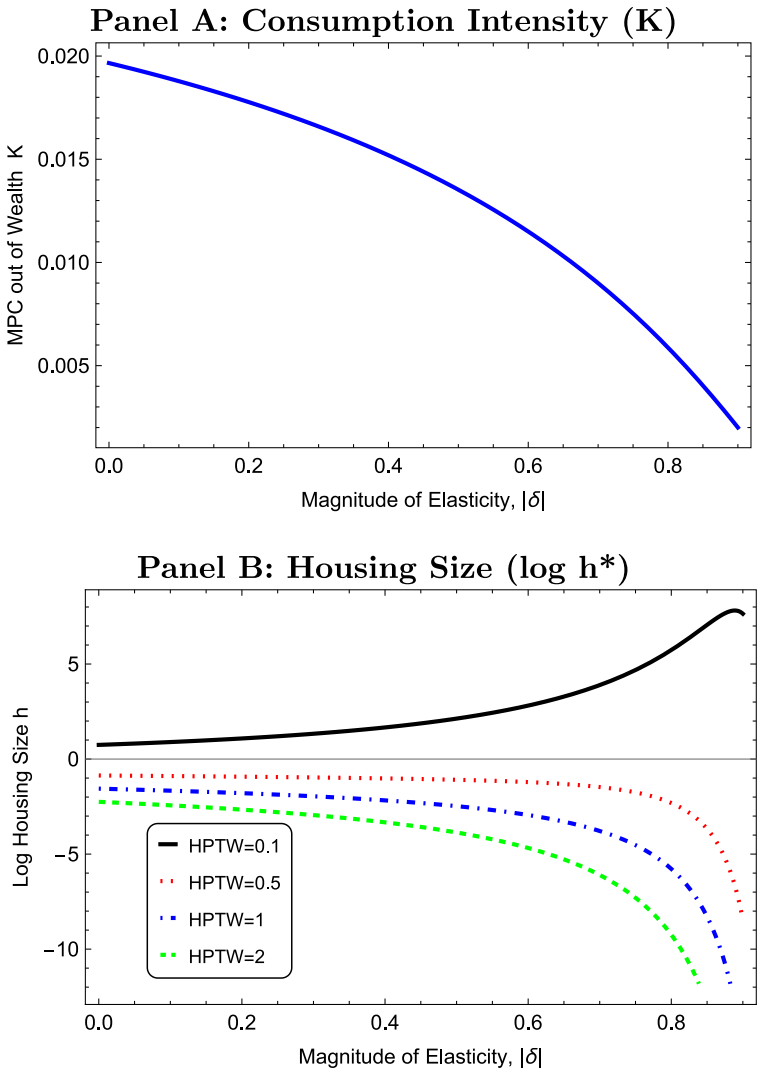


Fig. 3. Asymmetric Impact of $|\delta|$ on Consumption (K) and Housing (h^*)

Table 2
Comparison of pricing rules and budgetary impact.

Variable	Low elasticity ($\delta = -0.3$)	High elasticity ($\delta = -0.8$)
Housing Choice Change	$h = 1.0 \rightarrow 0.5$ (-50%)	$h = 1.0 \rightarrow 0.5$ (−50%)
Per-Unit Rent	Increases by 23% ($= 0.5^{-0.3}$)	Increases by 74% ($= 0.5^{-0.8}$)
Budget Impact	Low Penalty	High Penalty

- Financially unconstrained households (low HPTW) are relatively less concerned with penalties. Their decisions are guided primarily by the price signal, prompting them to expand housing size as $|\delta|$ increases to capture greater bulk discounts in the reward zone.
- Financially constrained households (high HPTW), by contrast, are highly exposed to penalties. For them, the structural signal is decisive. As $|\delta|$ rises, the cost of deviation intensifies, leading them to reduce housing size to avoid the penalty zone.

The same structural signal also explains the uniform consumption response. A larger $|\delta|$ heightens background risk for all households, strengthening the precautionary savings motive and lowering K for everyone. This interpretation is consistent with evidence that reduced ambiguity lowers precautionary savings (Ilut and Schneider, 2022; Enke and Graeber, 2023).

Our analysis identifies a theoretical channel explaining how the rent structure shapes housing choice. We show that a non-linear rent structure alone creates different housing responses based on a household's financial constraints. While we use the HPTW ratio as a proxy for this constraint, our contribution is to isolate this theoretical mechanism, not to make a causal empirical claim. We

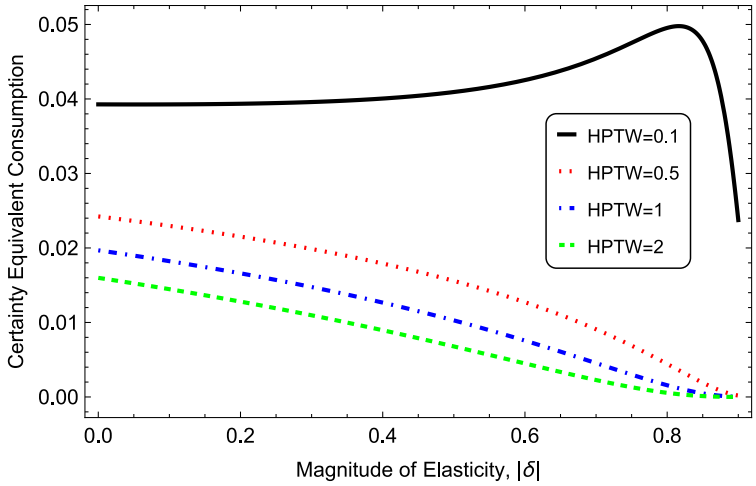


Fig. 4. Certainty equivalent consumption (c_{ec}) as a Function of $|\delta|$ and Financial Constraint (HPTW)

acknowledge that in the data, this link would be confounded by factors like credit access and risk preferences.[7] With this channel now clearly defined, we turn to its welfare consequences.

5.4. Social welfare

Our central finding is that policies making the rental market more linear — i.e., reducing the elasticity magnitude, $|\delta|$ — might reduce welfare inequality by disproportionately benefiting financially constrained households. To see this, we measure welfare using *certainty equivalent consumption* (CEC), which provides a monetary measure of lifetime utility.[8]

Fig. 4 shows that although financially unconstrained households enjoy higher absolute welfare, the welfare gap between unconstrained and constrained households widens as $|\delta|$ increases. The possible reason is that a higher $|\delta|$ corresponds to a more extreme price structure: larger bulk discounts for upsizing and steeper penalties for downsizing.

A welfare decomposition clarifies the divergence:

$$\frac{dV}{d|\delta|} = \frac{\partial V}{\partial c^*}\frac{\partial c^*}{\partial |\delta|} + \frac{\partial V}{\partial h^*}\frac{\partial h^*}{\partial |\delta|}.$$

For constrained households, both the consumption effect ($\partial c^*/\partial|\delta| < 0$) and the housing effect ($\partial h^*/\partial|\delta| < 0$) are negative, yielding lower welfare. For unconstrained households, the housing effect is typically positive ($\partial h^*/\partial|\delta| > 0$). Their welfare generally rises with $|\delta|$ because this positive housing effect, enabled by exploiting economies of scale, dominates the negative consumption effect.

These findings imply clear government policy levers. Interventions that reduce $|\delta|$, such as progressive property taxes, can compress welfare inequality by mitigating structural penalties. However, this approach introduces a critical equity–efficiency trade–off. Smaller rent differences may weaken incentives for large-unit investment and shift demand toward smaller units. When supply is tight, this demand shift can inadvertently push up small-unit rents. Therefore, an effective strategy must pair a flatter rent structure with complementary measures, such as expanding small-unit supply and preserving investment incentives across all unit sizes, to achieve both equity and efficiency.

6. Conclusion

We show that the rental market's price structure elicits asymmetric responses conditional on households' financial constraints. When the magnitude of rent elasticity, $|\delta|$, increases, all households reduce consumption, but their housing choices diverge: the unconstrained upsize while the constrained downsize. This reveals a fundamental trade-off between a price signal (bulk discounts) and a structural signal (financial penalties), with financial constraints determining the dominant force.

The policy implications are significant. Policies that make the rental market more inelastic (i.e., reducing $|\delta|$) not only enhance aggregate welfare but also reduce disparities between constrained and unconstrained households. In doing so, these interventions offer a targeted welfare instrument that moves beyond representative-agent models to account for household heterogeneity.

Our theoretical predictions are consistent with established empirical regularities on rent burden, housing crowding, and housing instability. Empirical work shows that high rent burdens crowd out other essential spending and are associated with smaller, more crowded units and more frequent forced moves (JCHS, 2022; Desmond and Gershenson, 2016). Appendix E documents related patterns in recent U.S. data and links them to our estimated $|\delta|$ and the model's affordability mechanism.

[7] Appendix E provides descriptive evidence from ACS and PSID data that is consistent with this affordability channel and discusses the limits of what can be inferred from the available microdata.

[8] See Appendix C for details.

CRediT authorship contribution statement

Hyun-Tak Lee: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Bong-Gyu Jang:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Kum-Hwan Roh:** Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Yong Hyun Shin:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.frl.2025.109176.

Data availability

Data will be made available on request.

References

Albouy, D., Ehrlich, G., Liu, Y., 2016. Housing demand, cost-of-living inequality, and the affordability crisis. http://dx.doi.org/10.3386/w22816, NBER Working Paper No. 22816.

Brunauer, W.A., Lang, S., Wechselberger, P., Zeugner, S., 2010. Additive hedonic regression models with spatial scaling factors: An application for rents in vienna. J. Real Estate Financ. Econ. 41 (4), 390–411. http://dx.doi.org/10.1007/s11146-009-9177-z.

Campbell, J.Y., Cocco, J.F., 2007. How do house prices affect consumption? Evidence from micro data. J. Monet. Econ. 54 (3), 591–621. http://dx.doi.org/10.1016/j.jmoneco.2005.10.016.

Desmond, M., Gershenson, C., 2016. Housing and employment insecurity among the working. Poor. Soc. Probl. 63 (1), 46–67. http://dx.doi.org/10.1093/socpro/spv025.

Enke, B., Graeber, T., 2023. Cognitive uncertainty. Q. J. Econ. 138 (4), 2021–2067. http://dx.doi.org/10.1093/qje/qjad025.

Ilut, C., Schneider, M., 2022. Modeling uncertainty as ambiguity: A review. http://dx.doi.org/10.3386/w29915, NBER Working Paper No. 29915.

Jang, B.-G., Park, S., Rhee, Y., 2013. Optimal retirement with unemployment risks. J. Bank. Financ. 37 (9), 3585–3604. http://dx.doi.org/10.1016/j.jbankfin.2013.05.017.

Joint Center for Housing Studies of Harvard University, 2022. America's rental housing 2022. In: JCHS Report. jchs.harvard.edu/americas-rental-housing-2022.

Landvoigt, T., Piazzesi, M., Schneider, M., 2015. The housing market(s) of san diego. Am. Econ. Rev. 105 (4), 1371–1417. http://dx.doi.org/10.1257/aer.20111662.

Leung, K.M., Yiu, C.Y., Lai, K., 2022. Responsiveness of sub-divided unit tenants' housing consumption to income: A study of Hong Kong informal housing. Hous. Stud. http://dx.doi.org/10.1080/02673037.2020.1803799.

Liu, H., Loewenstein, M., 2002. Optimal portfolio selection with transaction costs and finite horizons. Rev. Financ. Stud. 15 (3), 805–835. http://dx.doi.org/10.1093/rfs/15.3.805.

Sinai, T., Souleles, N., 2005. Owner-occupied housing as a hedge against rent risk. Q. J. Econ. 120 (2), 763–789. http://dx.doi.org/10.1093/qje/120.2.763.

신진우 교수

논문

- P. T. Ho, J. Shin, Blow-up phenomena for the constant scalar curvature and constant boundary mean curvature equation (after Chen and Wu), NoDEA Nonlinear Differential Equations Appl. 32 (2025), no. 6, Paper No. 130, 17pp.

Nonlinear Differ. Equ. Appl.
(2025) 32:130
© 2025 The Author(s), under exclusive licence
to Springer Nature Switzerland AG
https://doi.org/10.1007/s00030-025-01130-x

Nonlinear Differential Equations
and Applications NoDEA

Blow-up phenomena for the constant scalar curvature and constant boundary mean curvature equation (after Chen and Wu)

Pak Tung Ho and Jinwoo Shin

Abstract. In this paper, the compactness of the solutions to the constant scalar curvature and constant boundary mean curvature equation is considered. Chen and Wu constructed a smooth counterexample showing that the compactness of the set of "lower energy" solutions to the above equation fails when the dimension of the manifold is not less than 62. We prove that a smooth counterexample still exists when the dimension of the manifold is not less than 35.

Mathematics Subject Classification. 53C21, 35J20, 35B33.

Keywords. Yamabe problem, Compactness, Manifolds with boundary.

1. Introduction

Given an n-dimensional closed (i.e. compact without boundary) Riemannian manifold (M, g) with $n \geq 3$, the Yamabe problem is to find a metric $\tilde{g}$ conformal to g such that the scalar curvature $R_{\tilde{g}}$ of $\tilde{g}$ is constant. If we write $\tilde{g} = u^{\frac{4}{n-2}} g$ with $0 < u \in C^\infty(M)$, then $\tilde{g}$ has constant scalar curvature c if and only if

$$\frac{4(n-1)}{n-2} \Delta_g u - R_g u + c u^{\frac{n+2}{n-2}} = 0. \tag{1.1}$$

The Yamabe problem is solved by Aubin [6], Trudinger [34], and Schoen [31]. In other words, there exists at least one solution to (1.1).

Solutions to (1.1) are usually not unique. In [29], Pollack has used gluing techniques to prove the following: given any conformal class with positive Yamabe constant and any positive integer N, there exists a new conformal class which is close to the original one in the C^0-norm and contains at least N metrics of constant scalar curvature.

Published online: 23 September 2025

It is an interesting question whether the set of all solutions to (1.1) is compact in the C^2-topology. A conjecture due to Schoen states that this should be true unless (M, g) is conformally equivalent to the standard sphere (see [30,32,33]). This conjecture has been verified in low dimensions [16,26,27] and in the locally conformally flat case [32,33]. In particular, Khuri, Marques and Schoen [24] proved that the compactness conjecture is true up to dimension 24, assuming that the positive mass theorem holds.

It turns out that the conjecture is false in higher dimension. Brendle [7] constructed Riemannian manifolds (M, g) such that the set of constant scalar curvature metrics in the conformal class of g is non-compact, when the dimension $n \geq 52$. Modifying the arguments in [7], Brendle and Marques [9] were able to construct such Riemannian manifold when the dimension $25 \leq n \leq 52$.

Now suppose that (M, g) is an n-dimensional compact Riemannian manifold with boundary ∂M. The Yamabe problem can also be formulated on Riemannian manifolds with boundary. And there are two types:

(I) Find $\tilde{g}$ conformal to g such that the scalar curvature of $\tilde{g}$ is constant in M and the mean curvature of $\tilde{g}$ is zero on ∂M.

(II) Find $\tilde{g}$ conformal to g such that the scalar curvature of $\tilde{g}$ is zero in M and the mean curvature of $\tilde{g}$ is constant on ∂M.

If we write $\tilde{g} = u^{\frac{4}{n-2}} g$ with $0 < u \in C^\infty(M)$, then

$$\begin{aligned} -\frac{4(n-1)}{n-2}\Delta_g u + R_g u &= R_{\tilde{g}} u^{\frac{n+2}{n-2}} \quad \text{in } M, \\ \frac{2}{n-2}\frac{\partial u}{\partial \nu_g} + H_g u &= H_{\tilde{g}} u^{\frac{n}{n-2}} \quad \text{on } \partial M. \end{aligned} \tag{1.2}$$

Here $\dfrac{\partial}{\partial \nu_g}$ is the outward normal derivative with respect to g, and H_g (and $H_{\tilde{g}}$ respectively) is the mean curvature of g (and $\tilde{g}$ respectively). Therefore, the Yamabe problem with boundary (I) is equivalent to solving (1.2) with $R_{\tilde{g}} \equiv c$ and $H_{\tilde{g}} \equiv 0$, i.e.

$$\begin{aligned} -\frac{4(n-1)}{n-2}\Delta_g u + R_g u &= c u^{\frac{n+2}{n-2}} \quad \text{in } M, \\ \frac{2}{n-2}\frac{\partial u}{\partial \nu_g} + H_g u &= 0 \quad \text{on } \partial M, \end{aligned} \tag{1.3}$$

and the Yamabe problem with boundary (II) is equivalent to solving (1.2) with $R_{\tilde{g}} \equiv 0$ and $H_{\tilde{g}} \equiv c$, i.e.

$$\begin{aligned} -\frac{4(n-1)}{n-2}\Delta_g u + R_g u &= 0 \quad \text{in } M, \\ \frac{2}{n-2}\frac{\partial u}{\partial \nu_g} + H_g u &= c u^{\frac{n}{n-2}} \quad \text{on } \partial M. \end{aligned} \tag{1.4}$$

The Yamabe problem with boundary (I) and (II) have been studied intensively. See [3,8,14,17,18,28] and the references therein.

Inspired by the compactness conjecture in the closed case, one can ask if the set of all solutions to (1.3) (and to (1.4) respectively) is compact in C^2-topology. This was studied in [1,3,15,19–21,25]

More generally, one can try to find a metric $\tilde{g}$ conformal to g such that the scalar curvature of $\tilde{g}$ is equal to a in M and the mean curvature of $\tilde{g}$ is equal to b on ∂M, for some constants a and b. This includes the Yamabe problem with boundary (I) and (II) as special cases. In view of (1.2), it is equivalent to solving the following:

$$\begin{aligned} -\Delta_g u + \frac{n-2}{4(n-1)} R_g u &= c_1 u^{\frac{n+2}{n-2}} \quad \text{in } M, \\ \frac{\partial u}{\partial \nu_g} + \frac{n-2}{2} H_g u &= c u^{\frac{n}{n-2}} \quad \text{on } \partial M \end{aligned} \tag{1.5}$$

where $c_1 = \dfrac{n-2}{4(n-1)} a$ and $c = \dfrac{n-2}{2} b$. This problem of finding conformal metric of constant scalar curvature and constant boundary mean curvature has been studied in [10–12,22,23].

Similar to the Yamabe problem on closed manifolds and the Yamabe problem with boundary (I) and (II), it would be interesting to study the compactness and non-compactness of (1.5). Almaraz and Wang [5] obtained a compactness result of (1.5) when the dimension $n = 3$. Chen and Wu [13] were able to construct Riemannian manifolds (M, g) with boundary ∂M such that the set of all solutions to (1.5) is non-compact, when the dimension $n \geq 62$. In order to state their result, we need some notations. We define the functional

$$\begin{aligned} I_{(M,g)}[u] = &\int_M \left(\frac{4(n-1)}{n-2} |\nabla_g u|^2 + R_g u^2 \right) d\mu_g + 2(n-1) \int_{\partial M} H_g u^2 d\sigma_g \\ &- \frac{4(n-1)}{n} c_1 \int_M u_+^{\frac{2n}{n-2}} d\mu_g - 4c \int_{\partial M} u_+^{\frac{2(n-1)}{n-2}} d\sigma_g \end{aligned} \tag{1.6}$$

for any $u \in H^1(M, g)$, where $u_+ = \max\{u, 0\}$. Let W be a single bubble in $(\mathbb{R}^n_+, |dx|^2)$, i.e.

$$W(x) = \left(\frac{n(n-2)}{c_1} \right)^{\frac{n-2}{4}} (1 + |x - T_c \mathbf{e}_n|^2)^{\frac{2-n}{2}} \quad \text{with } T_c = -\frac{c}{n-2} \sqrt{\frac{n(n-2)}{c_1}},$$

and S_c the energy of W, i.e.

$$\begin{aligned} S_c &= I_{(\mathbb{R}^n_+, |dx|^2)}[W] \\ &= \int_{\mathbb{R}^n_+} \frac{4(n-1)}{n-2} |\nabla W|^2 dx - \frac{4(n-1)}{n} c_1 \int_{\mathbb{R}^n_+} W^{\frac{2n}{n-2}} dx - 4c \int_{\partial \mathbb{R}^n_+} W^{\frac{2(n-1)}{n-2}} d\sigma \\ &= \frac{4}{n-2} \int_{\mathbb{R}^n_+} |\nabla W|^2 dx + \frac{4c_1}{n-2} \int_{\mathbb{R}^n_+} W^{\frac{2n}{n-2}} dx > 0. \end{aligned}$$

The following theorem was proved in [13, Theorem 1.1].

Theorem 1.1. *Let $c \in \mathbb{R}_+$ and $n \geq 62$, there exists a metric g on $\mathbb{S}^n_+$ such that the Yamabe constant with boundary $Y(\mathbb{S}^n_+, \partial\mathbb{S}^n_+, g) > 0$, and a sequence of positive smooth functions $\{u_\nu\}_{\nu\in\mathbb{N}}$ with the following properties:*

(i) *g is not locally conformally flat,*
(ii) *$\partial\mathbb{S}^n_+$ is umbilic with respect to g,*
(iii) *u_ν is a positive smooth solution to (1.5) with $c_1 = n(n-2)$ and $M = \mathbb{S}^n_+$,*
(iv) *$I_{(\mathbb{S}^n_+, g)}[u_\nu] < S_c$, and*
(v) *$\sup_{\partial\mathbb{S}^n_+} u_\nu \to \infty$ as $\nu \to \infty$.*

In this paper, by following and modifying the arguments of Chen and Wu in [13], we can lower the dimension further to $n \geq 35$. The following is our main theorem.

Theorem 1.2. *Let $n \geq 35$. There exists a constant $\bar{c} \in \mathbb{R}^+$ depending only on n, such that for all $c \in \mathbb{R}^+$ with $c < \bar{c}$, there exist a metric g on $\mathbb{S}^n_+$ whose Yamabe constant with boundary $Y(\mathbb{S}^n_+, \partial\mathbb{S}^n_+, g) > 0$, and a sequence of positive smooth functions $\{u_\nu\}_{\nu\in\mathbb{N}}$ with properties (i)-(v) in Theorem 1.1.*

We would like to emphasize the main difference between Theorem 1.1 and Theorem 1.2. Chen and Wu were able to prove that Theorem 1.1 holds for all $c \in \mathbb{R}^+$, while we are only able to prove that Theorem 1.2 holds for all $c \in \mathbb{R}^+$ less than a particular value $\bar{c}$.

In Section 2, we will summarize the argument of Chen and Wu used in the proof of Theorem 1.1. In Section 3, we will then point out the exact place we need to modify the argument of Chen and Wu in order to prove Theorem 1.2.

2. Summary of the proof of Theorem 1.1

In this section, we provide a concise exposition of the principal arguments employed by Chen and Wu in the proof of Theorem 1.1. This overview serves to clarify the framework within which the subsequent modifications will be introduced.

From now on, we let $c_1 = n(n-2)$ and $T_c = -c/(n-2)$, where c is given in (1.5). Given a pair $(\xi, \epsilon) \in \mathbb{R}^{n-1} \times (0, \infty)$, we define

$$u_{(\xi,\epsilon)}(x) = \left(\frac{\epsilon}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x' - \xi|^2}\right)^{\frac{n-2}{2}},$$

where $x = (x', x_n) \in \mathbb{R}^n_+$. Define

$$u_{(\xi,\epsilon,i)} = \left(\frac{\epsilon}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x' - \xi|^2}\right)^{\frac{n+2}{2}} \frac{2\epsilon(x_i - \xi_i)}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x' - \xi|^2},$$

$$\hat{u}_{(\xi,\epsilon,i)} = \left(\frac{\epsilon}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x' - \xi|^2}\right)^{\frac{n}{2}} \frac{2\epsilon(x_i - \xi_i)}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x' - \xi|^2}$$

for $1 \le i \le n-1$, and

$$u_{(\xi,\epsilon,n)} = \left(\frac{\epsilon}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x' - \xi|^2}\right)^{\frac{n+2}{2}} \frac{(1+T_c^2)\epsilon^2 - x_n^2 - |x'-\xi|^2}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x'-\xi|^2},$$

$$\hat{u}_{(\xi,\epsilon,n)} = \left(\frac{\epsilon}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x' - \xi|^2}\right)^{\frac{n}{2}} \frac{(1+T_c^2)\epsilon^2 - x_n^2 - |x'-\xi|^2}{\epsilon^2 + (x_n - T_c\epsilon)^2 + |x'-\xi|^2}.$$

Define

$$\mathcal{E} = \left\{ w \in L^{\frac{2n}{n-2}}(\mathbb{R}^n_+) \cap L^{\frac{2(n-1)}{n-2}}(\partial\mathbb{R}^n_+) \cap H^1_{loc}(\mathbb{R}^n_+) : \int_{\mathbb{R}^n_+} |\nabla w|^2 < \infty \right\}$$

and

$$\mathcal{E}_{(\xi,\epsilon)} = \left\{ w \in \mathcal{E} : 2n \int_{\mathbb{R}^n_+} w u_{(\xi,\epsilon,a)} - T_c \int_{\partial\mathbb{R}^n_+} w \hat{u}_{(\xi,\epsilon,a)} = 0 \text{ for all } 1 \le a \le n \right\}.$$

We define a norm on $\mathcal{E}$ by $\|w\|_{\mathcal{E}} = \left(\int_{\mathbb{R}^n_+} |\nabla w|^2\right)^{\frac{1}{2}}$ Clearly, $u_{(\xi,\epsilon)} \in \mathcal{E}_{(\xi,\epsilon)}$.

We introduce a multi-linear form $\overline{W} : \mathbb{R}^{n-1} \times \mathbb{R}^{n-1} \times \mathbb{R}^{n-1} \times \mathbb{R}^{n-1} \to \mathbb{R}$ satisfying the same algebraic properties of the Weyl tensor on $\partial\mathbb{R}^n_+$. Moreover, we assume

$$\sum_{i,j,k,l=1}^{n-1} (\overline{W}_{ikjl} + \overline{W}_{iljk})^2 > 0.$$

If $x = (x', x_n) \in \mathbb{R}^n_+$, we identify x' with $(x', 0) \in \partial\mathbb{R}^n_+$ and define

$$H_{ij}(x) = H_{ij}(x') = \overline{W}_{ikjl} x_k x_l \quad \text{and} \quad H_{na}(x) = 0, \tag{2.1}$$

as well as $\overline{H}_{ab}(x) = f(|x'|^2) H_{ab}(x)$, where $f(s)$ is a polynomial of degree d for $0 \le d < (n-6)/4$ and to be determined.

We then define

$$\begin{aligned}\mathcal{F}(\xi,\epsilon) = \frac{1}{2} &\int_{\mathbb{R}^n_+} \sum_{i,j,l=1}^{n-1} \overline{H}_{il}\overline{H}_{jl} \partial_i u_{(\xi,\epsilon)} \partial_j u_{(\xi,\epsilon)} \\ &- \frac{n-2}{16(n-1)} \int_{\mathbb{R}^n_+} \sum_{i,j,l=1}^{n-1} \sum_{i,j,l=1}^{n-1} (\partial_l \overline{H}_{ij})^2 u^2_{(\xi,\epsilon)} \\ &+ \int_{\mathbb{R}^n_+} \sum_{i,j=1}^{n-1} \overline{H}_{ij} \partial_i \partial_j u_{(\xi,\epsilon)} z_{(\xi,\epsilon)}\end{aligned}$$

for $(\xi,\epsilon) \in \mathbb{R}^n_+ \times (0,\infty)$, where $z_{(\xi,\epsilon)}(x) = \mu^{-1} w_{(\xi,\epsilon)}(x)$ satisfying

$$\begin{aligned}&\int_{\mathbb{R}^n_+} \left(\langle \nabla z_{(\xi,\epsilon)}, \nabla \varphi \rangle - n(n+2) u^{\frac{4}{n-2}}_{(\xi,\epsilon)} z_{(\xi,\epsilon)} \varphi \right) + nT_c \int_{\partial\mathbb{R}^n_+} u^{\frac{2}{n-2}}_{(\xi,\epsilon)} w_{(\xi,\epsilon)} \varphi \\ &= -\int_{\mathbb{R}^n_+} \mu \lambda^{2d} f(\lambda^{-2}|x'|^2) H_{ij}(x) \partial_i \partial_j u_{(\xi,\epsilon)} \varphi\end{aligned}$$

for all test function $\varphi \in \mathcal{E}_{(\xi,\epsilon)}$.

If $f(s) = \sum_{i=0}^{d} a_i s^i$, $a_i \in \mathbb{R}$ for $1 \le i \le d$, then α_q are constants defined as

$$\sum_{q=0}^{2d} \alpha_q s^q := (n+1)f(s)^2 + 4sf(s)f'(s) + 2s^2 f'(s)^2. \tag{2.2}$$

Also, we define

$$c_q = \int_0^\infty (1 + (t - T_c)^2)^{\frac{5+2q-n}{2}} dt, \ q \in \mathbb{N} \text{ and } 0 \le q \le 2d. \tag{2.3}$$

Then we have

$$I(s) = \sum_{q=0}^{2d} \left(c_q \alpha_q s^{q+2} \prod_{j=0}^{q} \frac{n-1+2j}{n-5-2j} \right). \tag{2.4}$$

Let

$$2f(s)f'(s) + sf'(s) := \sum_{q=0}^{2d-1} \beta_q s^q. \tag{2.5}$$

Then

$$J(s) = \sum_{q=0}^{2q-1} \left(c_q \beta_q s^{q+2} \prod_{j=0}^{q} \frac{n+3+2j}{n-5-2j} \right). \tag{2.6}$$

In order to show that $\mathcal{F}(\xi, \epsilon)$ has a strict local minimum at $(0,1)$, Chen and Wu showed that it suffices to show that $I(1) > 0$, $I'(1) = 0$, $I''(1) < 0$ and $J(1) < 0$. By choosing $d = 1$, i.e. f is a polynomial of degree 1, Chen and Wu was able to show this when $n \ge 62$.

Proposition 2.1. (Proposition 5.9 in [13]) *Let $n \ge 62$. There exists a polynomial $f(s) = -s + a_0$ such that $I(1) > 0$, $I'(1) = 0$, $I''(1) < 0$ and $J(1) < 0$. This implies that $\mathcal{F}(\xi, \epsilon)$ has a strict local minimum at the point $(0,1)$.*

Using Proposition 2.1, Chen and Wu proved the following:

Proposition 2.2. (Proposition 6.1 in [13]) *For $T_c \in \mathbb{R}_-$ and $n \ge n_0$, let $g = \exp(h)$ be a smooth Riemannian metric in $\overline{\mathbb{R}^n_+}$, where h is a symmetric trace-free two tensor in $\overline{\mathbb{R}^n_+}$ satisfying*

$$\begin{cases} h_{ij}(x) = \mu\lambda^2 f(\lambda^{-2}|x'|^2) H_{ij}(x), & \text{in } B^+_\rho; \\ h_{ab}(x) = 0, & \text{in } \overline{\mathbb{R}^n_+} \setminus B^+_\rho; \\ h_{na}(x) = 0, & \text{in } \overline{\mathbb{R}^n_+}, \end{cases}$$

where $0 < \mu \le 1$, $0 < \lambda \le \rho \le 1$, $1 \le i \le n-1$, $1 \le a, b \le n$ and H_{ab} is defined as in (2.1). Assume that $|h(x)| + |\partial h(x)| + |\partial^2 h(x)| \le \alpha$ for all $x \in \overline{\mathbb{R}^n_+}$. If α and $\mu^{-2}\lambda^{n-10}\rho^{2-n}$ are sufficiently small, then there exists a positive smooth solution of

$$\begin{cases} \Delta_g v - \dfrac{n-1}{4(n-2)} R_g v + n(n-2) v^{\frac{n+2}{n-2}} = 0, & \text{in } \mathbb{R}^n_+; \\ \dfrac{\partial v}{\partial x_n} = (n-2) T_c v^{\frac{n}{n-2}}, & \text{on } \partial\mathbb{R}^n_+. \end{cases} \tag{2.7}$$

Moreover, there exists $C = C(n, T_c) > 0$ such that

$$\sup_{B_\lambda^+(0)} v \geq C\lambda^{\frac{2-n}{2}}$$

and

$$\begin{aligned} &2(n-2)\int_{\mathbb{R}^n_+} v^{\frac{2n}{n-2}} - \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} v^{\frac{2(n-1)}{n-2}} \\ &< 2(n-2)\int_{\mathbb{R}^n_+} u_{(0,1)}^{\frac{2n}{n-2}} - \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} u_{(0,1)}^{\frac{2(n-1)}{n-2}}. \end{aligned}$$

Sketch of the Proof. For the sake of completeness, we provide the proof here. However, this is only a sketch, and for the full proof, the reader is referred to [13].

Step 1. If h satisfies the conditions of Proposition 2.2 and $g = \exp(h)$, then we can establish the following: Given any pair $(\xi, \epsilon) \in \mathbb{R}^{n-1} \times (0, \infty)$, there exists a unique function $v_{(\xi,\epsilon)} \in \mathcal{E}$ such that $v_{(\xi,\epsilon)} - u_{(\xi,\epsilon)} \in \mathcal{E}_{(\xi,\epsilon)}$, and

$$\begin{aligned} &\int_{\mathbb{R}^n_+} \left(\langle \nabla v_{(\xi,\epsilon)}, \nabla\varphi\rangle_g, c_n R_g v_{(\xi,\epsilon)}\varphi - n(n-2)|v_{(\xi,\epsilon)}|^{\frac{4}{n-2}} v_{(\xi,\epsilon)}\varphi\right) \\ &\quad + \int_{\partial\mathbb{R}^n_+} \left(d_n h_g v_{(\xi,\epsilon)} + (n-2)T_c |v_{(\xi,\epsilon)}^{\frac{2}{n-2}} v_{(\xi,\epsilon)}\right)\varphi = 0 \end{aligned}$$

for all $\varphi \in \mathcal{E}_{(\xi,\epsilon)}$. Moreover, there exists a positive constant C, depending only on T_c and n such that

$$\begin{aligned} &\|v_{(\xi,\epsilon)} - u_{(\xi,\epsilon)}\|_{\mathcal{E}} \\ &\quad \leq C\|\Delta_g u_{(\xi,\epsilon)} - c_n R_g u_{(\xi,\epsilon)} + n(n-2)u_{(\xi,\epsilon)}^{\frac{n+2}{n-2}}\|_{L^{\frac{2n}{n+2}}(\mathbb{R}^n_+)} \\ &\qquad + C\|h_g u_{(\xi,\epsilon)}\|_{L^{\frac{2(n-1)}{n}}(\partial\mathbb{R}^n_+)}. \end{aligned} \tag{2.8}$$

Step 2. Given a pair $(\xi, \epsilon) \in \mathbb{R}^{n-1} \times (0, \infty)$, we define the following energy functional

$$\begin{aligned} &\mathcal{F}_g(\xi, \epsilon) \\ &:= \int_{\mathbb{R}^n_+} \left(|\nabla v_{(\xi,\epsilon)}|_g^2 + c_n R_g v_{(\xi,\epsilon)}^2 - (n-2)^2 |v_{(\xi,\epsilon)}^{\frac{2n}{n-2}}\right) + d_n \int_{\partial\mathbb{R}^n_+} h_g v_{(\xi,\epsilon)}^2 \\ &\quad + \frac{(n-2)^2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} |v_{(\xi,\epsilon)}|^{\frac{2(n-1)}{n-2}} - 2(n-2)\int_{\mathbb{R}^n_+} u_{(\xi,\epsilon)}^{\frac{2n}{n-2}} + \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} u_{(\xi,\epsilon)}^{\frac{2(n-1)}{n-2}}. \end{aligned}$$

Then the functional $\mathcal{F}_g$ is continuously differentiable. Moreover, if (ξ, ϵ) is a critical point of $\mathcal{F}_g$, then the function $v_{(\xi,\epsilon)}$ is a positive smooth solution of

$$\begin{cases} -\Delta_g v_{(\xi,\epsilon)} + c_n R_g v_{(\xi,\epsilon)} = n(n-2)v_{(\xi,\epsilon)}^{\frac{n+2}{n-2}}, \text{ in } \mathbb{R}^n_+, \\ \frac{\partial v_{(\xi,\epsilon)}}{\partial x_n} - d_n h_g v_{(\xi,\epsilon)} = (n-2)T_c v_{(\xi,\epsilon)}^{\frac{n}{n-2}}, \quad \text{on } \partial\mathbb{R}^n_+. \end{cases}$$

Step 3. Applying Step 1 to each pair $(\xi,\epsilon)\in\mathbb{R}^{n-1}\times(0,\infty)$, we choose $v_{(\xi,\epsilon)}$ to be the unique element of $\mathcal{E}$ such that $v_{(\xi,\epsilon)}-u_{(\xi,\epsilon)}\in\mathcal{E}_{(\xi,\epsilon)}$ and

$$\int_{\mathbb{R}^n_+}\Big(\langle\nabla v_{(\xi,\epsilon)},\nabla\varphi\rangle_g+c_nR_gv_{(\xi,\epsilon)}\varphi-n(n-2)|v_{(\xi,\epsilon)}|^{\frac{4}{n-2}}v_{(\xi,\epsilon)}\varphi\Big)$$
$$+(n-2)T_c\int_{\partial\mathbb{R}^n_+}|v_{(\xi,\epsilon)}|^{\frac{2}{n-2}}v_{(\xi,\epsilon)}\varphi=0$$

for all $\varphi\in\mathcal{E}_{(\xi,\epsilon)}$. Let $\Omega=\big\{(\xi,\epsilon)\in\mathbb{R}^{n-1}\times(0,\infty):|\xi|<1,\frac{1}{2}<\epsilon<1\big\}$. We define the function $w_{(\xi,\epsilon)}$ as the unique element of $\mathcal{E}_{(\xi,\epsilon)}$ satisfying

$$\int_{\mathbb{R}^n_+}\Big(\langle\nabla w_{(\xi,\epsilon)},\nabla\varphi\rangle-n(n+2)u_{(\xi,\epsilon)}^{\frac{4}{n-2}}w_{(\xi,\epsilon)}\varphi\Big)+nT_c\int_{\partial\mathbb{R}^n_+}u_{(\xi,\epsilon)}^{\frac{2}{n-2}}w_{(\xi,\epsilon)}\varphi$$
$$=-\int_{\mathbb{R}^n_+}\mu\lambda^{2d}f(\lambda^{-2}|x'|^2H_{ij}(x)\partial_i\partial_ju_{(\xi,\epsilon)}\varphi$$

for all $\varphi\in\mathcal{E}_{(\xi,\epsilon)}$. Then we can establish the following: For any $(\xi,\epsilon)\in\lambda\Omega$, there holds

$$\Big|\mathcal{F}_g(\xi,\epsilon)-\frac{1}{2}\int_{B^+_\rho(0)}\sum_{a,b,c=1}^n h_{ac}h_{bc}\partial_au_{(\xi,\epsilon)}\partial_bu_{(\xi,\epsilon)}+\frac{c_n}{4}\int_{B^+_\rho(0)}\sum_{a,b,c=1}^n(\partial_ch_{ab})^2u^2_{(\xi,\epsilon)}$$
$$-\mu\lambda^{2d}\int_{\mathbb{R}^n_+}f(\lambda^{-2}|x'|^2)w_{(\xi,\epsilon)}\sum_{a,b=1}^n H_{ab}(x)\partial_a\partial_bu_{(\xi,\epsilon)}\Big|$$
$$\le C\mu^{\frac{2(n-1)}{n-2}}\lambda^{\frac{(4d+4)(n-1)}{n-2}}+C\mu\lambda^{2d+2+\frac{n-2}{2}}\rho^{\frac{2-n}{2}}+C\lambda^{n-2}\rho^{2-n}.$$

Step 4. It follows from Proposition 2.1 that $(0,1)$ is a strict local minimum point of $\mathcal{F}(\xi,\epsilon)$. Hence, we can find an open set $\Omega'\subset\Omega$ such that $(0,1)\in\Omega'$ and $\mathcal{F}(0,1)<\inf_{(\xi,\epsilon)\in\partial\Omega'}\mathcal{F}(\xi,\epsilon)<0$. By Step 3 with $d=1$, we have

$$\big|\mathcal{F}_g(\lambda\xi,\lambda\epsilon)-\lambda^8\mu^2\mathcal{F}(\xi,\epsilon)\big|$$
$$\le C\mu^{\frac{2(n-1)}{n-2}}\lambda^{\frac{8(n-1)}{n-2}}+C\mu\lambda^{\frac{n+6}{2}}\rho^{\frac{2-n}{2}}+C\lambda^{n-2}\rho^{2-n}$$

for all $(\xi,\epsilon)\in\Omega$, equivalently,

$$\big|\lambda^{-8}\mu^{-2}\mathcal{F}_g(\lambda\xi,\lambda\epsilon)-\mathcal{F}(\xi,\epsilon)\big|$$
$$\le C\mu^{\frac{2}{n-2}}\lambda^{\frac{8}{n-2}}+C\mu^{-1}\lambda^{\frac{n-10}{2}}\rho^{\frac{2-n}{2}}+C\mu^{-2}\lambda^{n-10}\rho^{2-n}$$

for all $(\xi,\epsilon)\in\Omega$. If $\mu^{-2}\lambda^{n-10}\rho^{2-n}$ is sufficiently small, then we have

$$\mathcal{F}_g(0,\lambda)<\int_{(\xi,\epsilon)\in\partial\Omega'}\mathcal{F}_g(\lambda\xi,\lambda\epsilon)<0.$$

Consequently, there exists $(\bar\xi,\bar\epsilon)\in\Omega'$ such that

$$\mathcal{F}_g(\lambda\bar\xi,\lambda\bar\epsilon)=\inf_{(\xi,\epsilon)\in\Omega'}\mathcal{F}_g(\lambda\xi,\lambda\epsilon)<0.$$

It follows from Step 2 that that the function $v=v_{(\lambda\bar\xi,\lambda\bar\epsilon)}$ obtained in Step 1 is a positive smooth solution to (2.7). By definition of $\mathcal{F}_g$ we

have

$$2(n-2)\int_{\mathbb{R}^n_+} v^{\frac{2n}{n-2}} - \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} v^{\frac{2(n-1)}{n-2}} = \mathcal{F}_g(\lambda\bar{\xi},\lambda\bar{\epsilon}) + 2(n-2)\int_{\mathbb{R}^n_+} u_{(\lambda\bar{\xi},\lambda\bar{\epsilon})}^{\frac{2n}{n-2}} - \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} u_{(\lambda\bar{\xi},\lambda\bar{\epsilon})}^{\frac{2(n-1)}{n-2}},$$

whence

$$2(n-2)\int_{\mathbb{R}^n_+} v^{\frac{2n}{n-2}} - \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} v^{\frac{2(n-1)}{n-2}} < 2(n-2)\int_{\mathbb{R}^n_+} u_{(0,1)}^{\frac{2n}{n-2}} - \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} u_{(0,1)}^{\frac{2(n-1)}{n-2}}.$$

Using (2.8), we estimate

$$\|v - u_{(\lambda\bar{\xi},\lambda\bar{\epsilon})}\|_{L^{\frac{2n}{n-2}}(B^+_\lambda(0))} \le \|v - u_{(\lambda\bar{\xi},\lambda\bar{\epsilon})}\|_{L^{\frac{2n}{n-2}}(\partial\mathbb{R}^n_+)} \le C\alpha.$$

Then

$$|B^+_\lambda(0)|^{\frac{n-2}{2n}} \sup_{B^+_\lambda(0)} v \ge \|v\|_{L^{\frac{2n}{n-2}}(B^+_\lambda(0))} \ge -C\alpha + \|u_{(\lambda\bar{\xi},\lambda\bar{\epsilon})}\|_{L^{\frac{2(n-1)}{n-2}}(B^+_\lambda(0))}.$$

Hence, if α is sufficiently small, then we obtain

$$\sup_{B^+_\lambda(0)} v \ge C\lambda^{\frac{2-n}{2}}.$$

This completes the proof.

□

We remark that in the original statement of Proposition 6.1 in [13], it is assumed that $f(s) = -s + a_0$ and $n_0 = 62$. However, one can see from the proof of Proposition 6.1 in [13] that Proposition 2.2 is still true.

With Proposition 2.2, one can follow the proof of Theorem 6.2 in [13] to prove the following:

Theorem 2.3. (Theorem 6.2 in [13]) *For $T_c \in \mathbb{R}_-$ and $n \ge n_0$, there exists a smooth Riemannian metric g in $\overline{\mathbb{R}^n_+}$ with the following properties:*
(i) $g_{ab}(x) = \delta_{ab}$ for $x \in \overline{\mathbb{R}^n_+} \setminus B^+_{1/2}(0)$,
(ii) g is not locally conformally flat,
(iii) $\partial\mathbb{R}^n_+$ is totally geodesic with respect to g,
(iv) there exists a sequence of positive smooth functions $\{v_\nu\}_{\nu\in\mathbb{N}}$ satisfying

$$\begin{cases} \Delta_g v_\nu - \dfrac{n-1}{4(n-2)}R_g v_\nu + n(n-2)v_\nu^{\frac{n+2}{n-2}} = 0, \text{ in } \mathbb{R}^n_+; \\ \dfrac{\partial v_\nu}{\partial x_n} = (n-2)T_c v_\nu^{\frac{n}{n-2}}, \qquad \text{on } \partial\mathbb{R}^n_+ \end{cases}$$

for all ν. Moreover, there hold

$$2(n-2)\int_{\mathbb{R}^n_+} v_\nu^{\frac{2n}{n-2}} - \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} v_\nu^{\frac{2(n-1)}{n-2}} < 2(n-2)\int_{\mathbb{R}^n_+} u_{(0,1)}^{\frac{2n}{n-2}} - \frac{n-2}{n-1}T_c\int_{\partial\mathbb{R}^n_+} u_{(0,1)}^{\frac{2(n-1)}{n-2}}.$$

for all ν, i.e. $I_{(\mathbb{R}^n_+,|dx|^2)}[v_\nu] < S_c$, and $\sup_{B_1(0)^+} v_\nu \to \infty$ as $\nu \to \infty$.

Now Theorem 1.1 is a direct consequence of Theorem 2.3 by taking $n_0 = 62$.

3. Proof of Theorem 1.2

In order to lower the dimension further, we are going to choose $f(s)$ to be a polynomial of degree 6. In addition, we will establish the corresponding version of Proposition 2.1. More precisely, we prove in this section the following:

Proposition 3.1. *Let $n \geq 35$. There exists a polynomial $f(s) = \sum_{i=0}^{6} a_i s^i$ such that $I(1) > 0$, $I'(1) = 0$, $I''(1) < 0$ and $J(1) < 0$, provided that $T_c = -\frac{c}{n-2}$ is sufficiently close to 0. This implies that $\mathcal{F}(\xi, \epsilon)$ has a strict local minimum at the point $(0, 1)$.*

With Proposition 3.1, Theorem 1.2 now follows from Proposition 2.2 and Theorem 2.3 with $n_0 = 35$. Thus it boils down to proving Proposition 3.1.

Now we choose $d = 6$ such that $f(s) = a_0 + a_1 s + a_2 s^2 + a_3 s^3 + a_4 s^4 + a_5 s^5 + a_6 s^6$, where a_1, a_3, a_5 are negative constants and a_2, a_4, a_6 are positive constants to be chosen later. It follows from (2.2) that

$$\begin{aligned}
\alpha_0 &= (n+1)a_0^2,\\
\alpha_1 &= 2(n+3)a_0a_1,\\
\alpha_2 &= (n+7)a_1^2 + 2(n+5)a_0a_2,\\
\alpha_3 &= 2(n+11)a_1a_2 + 2(n+7)a_0a_3,\\
\alpha_4 &= (n+17)a_2^2 + 2(n+15)a_1a_3 + 2(n+9)a_0a_4,\\
\alpha_5 &= 2(n+23)a_2a_3 + 2(n+19)a_1a_4 + 2(n+11)a_0a_5,\\
\alpha_6 &= (n+31)a_3^2 + 2(n+29)a_2a_4 + 2(n+23)a_1a_5 + 2(n+13)a_0a_6, \qquad (3.1)\\
\alpha_7 &= 2(n+39)a_3a_4 + 2(n+35)a_2a_5 + 2(n+27)a_1a_6,\\
\alpha_8 &= (n+49)a_4^2 + 2(n+47)a_3a_5 + 2(n+41)a_2a_6,\\
\alpha_9 &= 2(n+59)a_4a_5 + 2(n+55)a_3a_6,\\
\alpha_{10} &= (n+71)a_5^2 + 2(n+69)a_4a_6,\\
\alpha_{11} &= 2(n+83)a_5a_6,\\
\alpha_{12} &= (n+97)a_6^2.
\end{aligned}$$

Differentiating (2.4) with respect to s yields

$$I'(s) = \sum_{q=0}^{12} \left(c_q \alpha_q (q+2) s^{q+1} \prod_{j=0}^{q} \frac{n-1+2j}{n-5-2j} \right). \tag{3.2}$$

Combining this with (3.1), one can see that $I'(1)$ can be considered as a quadratic polynomial in a_0. More precisely, $p_n(a_0) := I'(1) = A_n a_0^2 + B_n a_0 + C_n$ where

$$\begin{aligned}
A_n &= \frac{2(n-1)(n+1)}{n-5} c_0, \\
B_n &= 2(n-1) \sum_{l=1}^{6} \left[(l+2) a_l c_l \prod_{j=0}^{l} \frac{n+1+2j}{n-5-2j} \right], \\
C_n &= 4(n+7) a_1^2 c_2 \prod_{0}^{2} \frac{n-1+2j}{n-5-2j} + 10(n+11) a_1 a_2 c_3 \prod_{j=0}^{3} \frac{n-1+2j}{n-5-2j} \\
&\quad + 6 \left[(n+17) a_2^2 + 2(n+15) a_1 a_3 \right] c_4 \prod_{j=0}^{4} \frac{n-1+2j}{n-5-2j} \\
&\quad + 14 \left[(n+23) a_2 a_3 + (n+19) a_1 a_4 \right] c_5 \prod_{j=0}^{5} \frac{n-1+2j}{n-5-2j} \\
&\quad + 8 \left[(n+31) a_3^2 + 2(n+29) a_2 a_4 + 2(n+23) a_1 a_5 \right] c_6 \prod_{j=0}^{6} \frac{n-1+2j}{n-5-2j} \\
&\quad + 18 \left[(n+39) a_3 a_4 + (n+35) a_2 a_5 + (n+27) a_1 a_6 \right] c_7 \prod_{j=0}^{7} \frac{n-1+2j}{n-5-2j} \\
&\quad + 10 \left[(n+49) a_4^2 + 2(n+47) a_3 a_5 + 2(n+41) a_2 a_6 \right] c_8 \prod_{j=0}^{8} \frac{n-1+2j}{n-5-2j} \\
&\quad + 22 \left[(n+59) a_4 a_5 + (n+55) a_3 a_6 \right] c_9 \prod_{j=0}^{9} \frac{n-1+2j}{n-5-2j} \\
&\quad + 12 \left[(n+71) a_5^2 + 2(n+69) a_4 a_6 \right] c_{10} \prod_{j=0}^{10} \frac{n-1+2j}{n-5-2j} \\
&\quad + 26(n+83) a_5 a_6 c_{11} \prod_{j=0}^{11} \frac{n-1+2j}{n-5-2j} \\
&\quad + 14(n+97) a_6^2 c_{12} \prod_{j=0}^{12} \frac{n-1+2j}{n-5-2j}.
\end{aligned} \tag{3.3}$$

Note that it follows from (2.3) that c_q is a continuous function of T_c, i.e. $c_q = c_q(T_c)$. Hence, $A_n = A_n(T_c)$, $B_n = B_n(T_c)$, and $C_n = C_n(T_c)$ can also

be regarded as continuous functions of T_c. Let $discrim(p_n)(T_c) := B_n(T_c)^2 - 4A_n(T_c)C_n(T_c)$, which is the discriminant of $I'(1)$, the quadratic polynomial in a_0. We will show that $discrim(p_n)(0) > 0$. By continuity, this implies that $discrim(p_n)(T_c) > 0$ when $T_c = -\dfrac{c}{n-2}$ is sufficiently close to 0. When $T_c = 0$, the explicit value of $c_q(0)$ can be computed by using (2.3):

$$c_q(0) = \frac{\sqrt{\pi}\Gamma(\frac{n-6-2q}{2})}{2\Gamma(\frac{n-5-2q}{2})}, \tag{3.4}$$

where $\Gamma(s)$ denotes the gamma function.

Now we choose $a_1 = -10$, $a_2 = 10^{-4}$, $a_3 = -10^{-3}$, $a_4 = 1.84 \times 10^{-1}$, $a_5 = -2.65 \times 10^{-2}$, and $a_6 = 7.37 \times 10^{-4}$. By (3.3) and (3.4), a direct computation with the help of `Mathematica` shows that $discrim(p_n)(0) > 0$ for $35 \le n \le 62$ (c.f. [35]). This tells us that there exists $a_0 \in \mathbb{R}$ such that $I'(1) = 0$, when $T_c = -\dfrac{c}{n-2}$ is sufficiently close to 0.

Differentiating (3.2) with respect to s yields

$$I''(s) = \sum_{q=0}^{12} \left(c_q \alpha_q (q+2)(q+1) s^q \prod_{j=0}^{q} \frac{n-1+2j}{n-5-2j} \right).$$

Combining this with (3.1), we have the following:

$$\begin{aligned}
I''(1) = {} & \frac{2(n-1)(n+1)}{n-5} a_0^2 c_0 + 2(n-1)a_0 \\
& \sum_{l=1}^{6} \left[(l+1)(l+2) a_l c_l \prod_{j=0}^{l} \frac{n+1+2j}{n-5-2j} \right] \\
& + 12(n+7) a_1^2 c_2 \prod_{j=0}^{2} \frac{n-1+2j}{n-5-2j} + 40(n+11) a_1 a_2 c_3 \prod_{j=0}^{3} \frac{n-1+2j}{n-5-2j} \\
& + 30 \left[(n+17) a_2^2 + 2(n+15) a_1 a_3 \right] c_4 \prod_{j=0}^{4} \frac{n-1+2j}{n-5-2j} \\
& + 84 \left[(n+23) a_2 a_3 + (n+19) a_1 a_4 \right] c_5 \prod_{j=0}^{5} \frac{n-1+2j}{n-5-2j} \\
& + 56 \left[(n+31) a_3^2 + 2(n+29) a_2 a_4 + 2(n+23) a_1 a_5 \right] c_6 \prod_{j=0}^{6} \frac{n-1+2j}{n-5-2j} \\
& + 144 \left[(n+39) a_3 a_4 + (n+35) a_2 a_5 + (n+27) a_1 a_6 \right] c_7 \prod_{j=0}^{7} \frac{n-1+2j}{n-5-2j} \\
& + 90 \left[(n+49) a_4^2 + 2(n+47) a_3 a_5 + 2(n+41) a_2 a_6 \right] c_8 \prod_{j=0}^{8} \frac{n-1+2j}{n-5-2j}
\end{aligned}$$

$$+ 220\left[(n+59)a_4a_5 + (n+55)a_3a_6\right] c_9 \prod_{j=0}^{9} \frac{n-1+2j}{n-5-2j}$$

$$+ 132\left[(n+71)a_5^2 + 2(n+69)a_4a_6\right] c_{10} \prod_{j=0}^{10} \frac{n-1+2j}{n-5-2j}$$

$$+ 312(n+83)a_5a_6c_{11} \prod_{j=0}^{11} \frac{n-1+2j}{n-5-2j}$$

$$+ 182(n+97)a_6^2c_{12} \prod_{j=0}^{12} \frac{n-1+2j}{n-5-2j}.$$

Note that $I''(1) = I''(1)(T_c)$ can also be regarded as a continuous function of T_c. We will show that $I''(1)(0) < 0$, which implies that $I''(1)(T_c) < 0$ when $T_c = -\dfrac{c}{n-2}$ is close to 0. Let $a_0 = a_0(n)$ be the largest root of the quadratic polynomial p_n when $T_c = 0$, i.e.

$$a_0(n) = \frac{-B_n(0) + \sqrt{B_n(0)^2 - 4A_n(0)C_n(0)}}{2A_n(0)}.$$

We remark that, unlike other a_i's that have a fixed value, a_0 depends on n. For example, we have

$$a_0(35) = \frac{388694358052 + \sqrt{31382012570285343694}}{21443906250}$$

when $n = 35$, while

$$a_0(62) = \frac{46081869321940760 + \sqrt{76861908977503143130771763741035}}{2750533632000000}$$

when $n = 62$. Then direct computations with the help of `Mathematica` show that $I''(1)(0) < 0$ for $35 \le n \le 62$ (c.f. [35]), which by continuity implies that $I''(1)(T_c) < 0$ for $35 \le n \le 62$ when T_c is sufficiently close to 0. Likewise, with the help of `Mathematica`, we can show that $I(1) > 0$ when $T_c = 0$ (c.f. [35]), which implies that $I(1) > 0$ when T_c is sufficiently close to 0.

It remains to show that $J(1) < 0$ when T_c is sufficiently close to 0. It follows from (2.5) that

$$\begin{aligned}
\beta_0 &= 2a_0a_1,\\
\beta_1 &= a_1 + 2a_1^2 + 4a_0a_2,\\
\beta_2 &= 2a_2 + 6a_1a_2 + 6a_0a_3,\\
\beta_3 &= 4a_2^2 + 3a_3 + 8a_1a_3 + 8a_0a_4,\\
\beta_4 &= 10a_2a_3 + 4a_4 + 10a_1a_4 + 10a_0a_5,\\
\beta_5 &= 6a_3^2 + 12a_2a_4 + 5a_5 + 12a_1a_5 + 12a_0a_6,\\
\beta_6 &= 14a_3a_4 + 14a_2a_5 + 6a_6 + 14a_1a_6,\\
\beta_7 &= 8a_4^2 + 16a_3a_5 + 16a_2a_6,\\
\beta_8 &= 18a_4a_5 + 18a_3a_6,
\end{aligned}$$

$$\begin{aligned}\beta_9 &= 10a_5^2 + 20a_4a_6,\\ \beta_{10} &= 20a_5a_6,\\ \beta_{11} &= 12a_6^2.\end{aligned}$$

Combining this with (2.6), we have

$$\begin{aligned}J(1) = {} & 2a_1a_0c_0\prod_{j=0}^{0}\frac{n+3+2j}{n-5-2j} + (a_1+2a_1^2+4a_2a_0)c_1\prod_{j=0}^{1}\frac{n+3+2j}{n-5-2j}\\ & + (2a_2+6a_1a_2+6a_3a_0)c_2\prod_{j=0}^{2}\frac{n+3+2j}{n-5-2j}\\ & + (4a_2^2+3a_3+8a_1a_3+8a_4a_0)c_3\prod_{j=0}^{3}\frac{n+3+2j}{n-5-2j}\\ & + (10a_2a_3+4a_4+10a_1a_4+10a_5a_0)c_4\prod_{j=0}^{4}\frac{n+3+2j}{n-5-2j}\\ & + (6a_3^2+12a_2a_4+5a_5+12a_1a_5+12a_6a_0)c_5\prod_{j=0}^{5}\frac{n+3+2j}{n-5-2j}\\ & + (14a_3a_4+14a_2a_5+6a_6+14a_1a_0)c_6\prod_{j=0}^{6}\frac{n+3+2j}{n-5-2j}\\ & + (8a_4^2+16a_3a_5+16a_2a_6)c_7\prod_{j=0}^{7}\frac{n+3+2j}{n-5-2j}\\ & + (18a_4a_5+18a_3a_6)c_8\prod_{j=0}^{8}\frac{n+3+2j}{n-5-2j}\\ & + (10a_5^2+20a_4a_6)c_9\prod_{j=0}^{9}\frac{n+3+2j}{n-5-2j}\\ & + 22a_5a_6c_{10}\prod_{j=0}^{10}\frac{n+3+2j}{n-5-2j} + 12a_6^2c_{11}\prod_{j=0}^{11}\frac{n+3+2j}{n-5-2j}.\end{aligned}$$

Viewing $J(1)$ as a continuous function of T_c, with our choice of a_i, we can show that $J(1)(0) < 0$ for $35 \le n \le 62$ with the help of `Mathematica` (c.f. [35]). This implies that, by continuity, $J(1) < 0$ when T_c is sufficiently close to 0. This finishes the proof of Proposition 3.1.

Acknowledgements

The first author was supported by the National Science and Technology Council (NSTC), Taiwan, with grant Number: 112-2115-M-032-006-MY2.

Author contributions The authors have contributed equally to the entirety of this manuscript

Data Availability Statement No datasets were generated or analysed during the current study.

Declarations
Competing interests The authors declare no competing interests.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

References

[1] Almaraz, S.: A compactness theorem for scalar-flat metrics on manifolds with boundary. Calc. Var. Partial. Differ. Equ. **41**(3–4), 341–386 (2011)

[2] Almaraz, S.: Blow-up phenomena for scalar-flat metrics on manifolds with boundary. J. Differ. Equ. **251**(7), 1813–1840 (2011)

[3] Almaraz, S.: An existence theorem of conformal scalar-flat metrics on manifolds with boundary. Pacific J. Math. **248**(1), 1–22 (2010)

[4] Almaraz, S., de Queiroz, O., Wang, S.: A compactness theorem for scalar-flat metrics on 3-manifolds with boundary. J. Funct. Anal. **277**(7), 2092–2116 (2019)

[5] Almaraz, S., Wang, S.: A compactness theorem for conformal metrics with constant scalar curvature and constant boundary mean curvature in dimension three. (2023), preprint. https://arxiv.org/abs/2306.07088

[6] Aubin, T.: Équations différentielles non linéaires et problème de yamabe concernant la courbure scalaire. J. Math. Pures Appl.(9) **55**(3), 269–296 (1976)

[7] Brendle, S.: Blow-up phenomena for the yamabe equation. J. Amer. Math. Soc. **21**(4), 951–979 (2008)

[8] Brendle, S., Chen, S.S.: An existence theorem for the yamabe problem on manifolds with boundary. J. Eur. Math. Soc. (JEMS) **16**(5), 991–1016 (2014)

[9] Brendle, S., Marques, F.C.: Blow-up phenomena for the yamabe equation. II. J. Differential Geom. **81**(2), 225–250 (2009)

[10] Chen, X., Ho, P.T., Sun, L.: Prescribed scalar curvature plus mean curvature flows in compact manifolds with boundary of negative conformal invariant. Ann. Global Anal. Geom. **53**(1), 121–150 (2018)

[11] Chen, X., Ruan, Y., Sun, L.: The han-li conjecture in constant scalar curvature and constant boundary mean curvature problem on compact manifolds. Adv. Math. **358**, 106854 (2019)

[12] Chen, X., Sun, L.: Existence of conformal metrics with constant scalar curvature and constant boundary mean curvature on compact manifolds. Commun. Contemp. Math. **21**(3), 1850021 (2019)

[13] Chen, X., Wu, N.: Blow-up phenomena for the constant scalar curvature and constant boundary mean curvature equation. J. Differential Equations **269**(11), 9432–9470 (2020)

[14] Cherrier, P.: Problèmes de Neumann non linéaires sur les variétés riemanniennes. J. Funct. Anal. **57**(2), 154–206 (1984)

[15] Disconzi, M., Khuri, M.A.: Compactness and non-compactness for the Yamabe problem on manifolds with boundary. J. Reine Angew. Math. **724**, 145–201 (2017)

[16] Druet, O.: Compactness for yamabe metrics in low dimensions. Int. Math. Res. Not. **23**, 1143–1191 (2004)

[17] Escobar, J.F.: Conformal deformation of a Riemannian metric to a scalar flat metric with constant mean curvature on the boundary. Ann. of Math. (2) **136**(1), 1–50 (1992)

[18] Escobar, J.F.: The yamabe problem on manifolds with boundary. J. Differential Geom. **35**(1), 21–84 (1992)

[19] Felli, V., Ould Ahmedou, M.: Compactness results in conformal deformations of Riemannian metrics on manifolds with boundaries. Math. Z. **244**(1), 175–210 (2003)

[20] Felli, V., Ould Ahmedou, M.: A geometric equation with critical nonlinearity on the boundary. Pacific J. Math. **218**(1), 75–99 (2005)

[21] Ghimenti, M., Micheletti, A.M.: Compactness for conformal scalar-flat metrics on umbilic boundary manifolds. Nonlinear Anal. **200**, 111992 (2020)

[22] Han, Z.C., Li, Y.Y.: The existence of conformal metrics with constant scalar curvature and constant boundary mean curvature. Comm. Anal. Geom. **8**(4), 809–869 (2000)

[23] Han, Z.C., Li, Y.Y.: The yamabe problem on manifolds with boundary: existence and compactness results. Duke Math. J. **99**(3), 489–542 (1999)

[24] Khuri, M.A., Marques, F.C., Schoen, R.: A compactness theorem for the yamabe problem. J. Differential Geom. **81**(1), 143–196 (2009)

[25] Kim, S., Musso, M., Wei, J.C.: Compactness of scalar-flat conformal metrics on low-dimensional manifolds with constant mean curvature on boundary. Ann. Inst. H. Poincaré C Anal. Non Linéaire **38**(6), 1763–1793 (2021)

[26] Li, Y.Y., Zhu, M.: Yamabe type equations on three-dimensional Riemannian manifolds. Commun. Contemp. Math. **1**(1), 1–50 (1999)

[27] Marques, F.C.: A priori estimates for the yamabe problem in the non-locally conformally flat case. J. Differential Geom. **71**(2), 315–346 (2005)

[28] Mayer, M., Ndiaye, C.B.: Barycenter technique and the riemann mapping problem of cherrier-escobar. J. Differential Geom. **107**(3), 519–560 (2017)

[29] Pollack, D.: Nonuniqueness and high energy solutions for a conformally invariant scalar equation. Comm. Anal. Geom. **1**(3–4), 347–414 (1993)

[30] Schoen, R.: A report on some recent progress on nonlinear problems in geometry. Surveys in differential geometry (Cambridge, MA,: 201–241, p. 1991. Lehigh Univ, Bethlehem, PA (1990)

[31] Schoen, R.: Conformal deformation of a riemannian metric to constant scalar curvature. J. Differential Geom. **20**(2), 479–495 (1984)

[32] Schoen, R.: On the number of constant scalar curvature metrics in a conformal class. Differential geometry, 311–320, Pitman Monogr. Surveys Pure Appl. Math., **52**, Longman Sci. Tech., Harlow, (1991)

[33] Schoen, R.: Variational theory for the total scalar curvature functional for Riemannian metrics and related topics. Topics in calculus of variations (Montecatini Terme, 1987), 120–154, Lecture Notes in Math., **1365**, Springer, Berlin, (1989)

[34] Trudinger, N.S.: Remarks concerning the conformal deformation of riemannian structures on compact manifolds. Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3) **22**, 265–274 (1968)

[35] The Mathematica script for this paper: https://sites.google.com/view/paktungho/code

Pak Tung Ho
Department of Mathematics
Tamkang University
Tamsui, New Taipei City 251301
Taiwan
e-mail: paktungho@yahoo.com.hk

Jinwoo Shin
Department of Mathematics & Research Institute of Natural Science
Sookmyung Women's University
Cheongpa-ro 47-gil 100
Youngsan-gu, Seoul 04310
Korea
e-mail: shinjin@sookmyung.ac.kr

Received: 7 May 2025.
Revised: 9 July 2025.
Accepted: 28 August 2025.

최진원 교수

논문

- Jinwon Choi, Young-Hoon Kiem and Donggun Lee (2025). Characteristic polynomial of $\overline{M_{0,n}}$ and log-concavity. Proceedings of the London Mathematical Society 131 (2025)

Check for updates

Received: 2 May 2025 | Revised: 12 October 2025 | Accepted: 16 October 2025

DOI: 10.1112/plms.70096

RESEARCH ARTICLE

Proceedings of the London Mathematical Society

Characteristic polynomial of $\overline{\mathcal{M}}_{0,n}$ and log-concavity

Jinwon Choi[1] | **Young-Hoon Kiem**[2] | **Donggun Lee**[3]

[1]Department of Mathematics and Research Institute of Natural Sciences, Sookmyung Women's University, Seoul, South Korea

[2]School of Mathematics, Korea Institute for Advanced Study, 85 Hoegiro, Dongdaemun-gu, Seoul, South Korea

[3]Center for Complex Geometry, Institute for Basic Science (IBS), 55 Expo-ro, Yuseong-gu, Daejeon, South Korea

Correspondence
Donggun Lee, Center for Complex Geometry, Institute for Basic Science (IBS), 55 Expo-ro, Yuseong-gu, Daejeon 34126, Korea.
Email: dglee@ibs.re.kr

Funding information
NRF, Grant/Award Numbers: 2018R1C1B6005600, 2021R1A2C1093787; Institute for Basic Science, Grant/Award Number: IBS-R032-D1

Abstract

Motivated by Stanley's generalization of the chromatic polynomial of a graph to the chromatic symmetric function, we introduce the characteristic polynomial of a representation of the symmetric group, or more generally, of a symmetric function. When the representation arises from geometry, the coefficients of its characteristic polynomial tend to form a log-concave sequence. To illustrate, we investigate explicit examples, including the n-fold products of the projective spaces, the GIT moduli spaces of points on $\mathbb{P}^1$ and Hessenberg varieties. Our main focus lies on the cohomology of the moduli space of pointed rational curves, for which we prove asymptotic formulas of its characteristic polynomial and establish asymptotic log-concavity.

MSC 2020
14H10, 14C15, 20C30, 05E05, 14N10

Contents

1. INTRODUCTION 2
2. TWO-DIMENSIONAL LOG-CONCAVITY 7
3. CHARACTERISTIC POLYNOMIALS OF SYMMETRIC FUNCTIONS AND REPRESENTATIONS 9
4. CHARACTERISTIC POLYNOMIALS OF VARIETIES AND LOG-CONCAVITY 18
5. CHARACTERISTIC POLYNOMIAL OF $\overline{\mathcal{M}}_{0,n}$ 27
6. ASYMPTOTIC BEHAVIOR AND LOG-CONCAVITY 38

© 2025 The Author(s). The publishing rights in this article are licensed to the London Mathematical Society under an exclusive licence.

7. PROOFS OF ASYMPTOTIC FORMULAS 42
ACKNOWLEDGMENTS 52
REFERENCES 52

1 | INTRODUCTION

In [26], Stanley introduced the chromatic symmetric function of a graph, as a symmetric function generalization of the chromatic polynomial of a graph. In this paper, by taking the inverse approach, we associate a polynomial, called the *characteristic polynomial*, to each symmetric function. Although the polynomials we consider are not the chromatic polynomials of graphs in general, those obtained from geometric contexts appear to share one of the most interesting features of chromatic polynomials, namely the *log-concavity* (cf. §2).

1.1 | Chromatic polynomial and characteristic polynomial

The *chromatic polynomial* of a graph Γ is a polynomial $\mathscr{X}_\Gamma(m) \in \mathbb{Q}[m]$ such that for each $m_0 \in \mathbb{N} = \mathbb{Z}_{\geqslant 1}$, the value $\mathscr{X}_\Gamma(m_0)$ equals the number of proper colorings of the vertices of Γ using m_0 colors. Here, a coloring is called *proper* if any two adjacent vertices (i.e., vertices connected by an edge) take different colors. In this paper, we regard $\mathscr{X}_\Gamma(m)$ as a polynomial in the variable m, and write $\mathscr{X}_\Gamma(m_0)$ for its evaluation at a positive integer m_0.

The *chromatic symmetric function* csf_Γ of Γ is defined by

$$\mathsf{csf}_\Gamma = \sum_{\substack{\kappa: V(\Gamma)\to\mathbb{Z}_{\geqslant 1},\\ \text{proper}}} \prod_{v\in V(\Gamma)} x_{\kappa(v)},$$

where $\{x_i\}_{i\in\mathbb{Z}_{\geqslant 1}}$ are formal variables and $V(\Gamma)$ is the set of vertices of Γ. This is obviously a symmetric function in x_i and the substitution

$$x_1 = x_2 = \cdots = x_{m_0} = 1 \quad \text{and} \quad x_{m_0+1} = x_{m_0+2} = \cdots = 0 \tag{1.1}$$

to csf_Γ gives the integer $\mathscr{X}_\Gamma(m_0)$. In this sense, csf_Γ can be considered as a generalization of $\mathscr{X}_\Gamma$. However, the log-concavity proved in [17] appears only after the substitution (1.1), and hence, it makes sense to extend the chromatic polynomials to more general contexts when one is interested in the log-concavity.

In this paper, we first observe that the process of producing a polynomial out of a symmetric function by the substitution (1.1) works not only for chromatic symmetric functions of graphs but also for all symmetric functions. Indeed, the substitution (1.1) produces a map $\mathscr{S}_F : \mathbb{N} \to \mathbb{Q}$ for any symmetric function F in the variables x_i with rational coefficients. It is straightforward to check that $\mathscr{S}_F$ is a polynomial function whose associated polynomial in $\mathbb{Q}[m]$ is called the *characteristic polynomial* $\mathscr{S}_F(m)$ of F (Definition 3.1).

It turns out that this definition is quite natural. To explain this, we first write it in a different but equivalent form. Let Λ be the power series ring of the symmetric functions in x_i over $\mathbb{Q}$, that is,

$$\Lambda = \varprojlim \mathbb{Q}[\![x_1, \dots, x_n]\!]^{\mathbb{S}_n},$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and let $\Lambda_n \subset \Lambda$ be the subspace of elements of degree n. Let $\mathsf{p}_n = \sum_{i\geqslant 1} x_i^n$ be the nth power sum symmetric function for $n \geqslant 1$ and let $\mathsf{p}_0 = 1$. It is well known that $\Lambda \cong \mathbb{Q}[\![\mathsf{p}_1, \mathsf{p}_2, \ldots]\!]$. The map associating the characteristic polynomial to each symmetric function

$$\widehat{\mathcal{S}} : \Lambda \longrightarrow \mathbb{Q}[m][\![q]\!], \quad (F \in \Lambda_n) \mapsto \mathcal{S}_F(m)q^n$$

is precisely the $\mathbb{Q}$-algebra homomorphism which sends p_n to mq^n for every $n \geqslant 1$, where q is a formal variable.

A remarkable feature of this map is that it behaves nicely with *plethysm* $\circ$ on Λ (see §3.1). In fact, it converts the plethysm $F \circ G$ of symmetric functions $F \in \Lambda_n$ and $G \in \Lambda_r$ to the composition $(\mathcal{S}_F \circ \mathcal{S}_G)(m) = \mathcal{S}_F(\mathcal{S}_G(m))$ (Proposition 3.2). Conversely, this feature uniquely determines $\widehat{\mathcal{S}}$ as a $\mathbb{Q}$-algebra homomorphism $\Lambda \to \mathbb{Q}[m][\![q]\!]$ Theorem 3.3.

We can also define the characteristic polynomial of a symmetric group representation. Let $\mathbb{S}_n$ denote the symmetric group of n letters and R_n be the $\mathbb{Q}$-vector space spanned by irreducible $\mathbb{S}_n$-representations. Let

$$\mathrm{ch} = \mathrm{ch}_{\mathbb{S}_n} : R_n \xrightarrow{\cong} \Lambda_n$$

be the Frobenius characteristic map (cf. (3.16)). The *characteristic polynomial* of an $\mathbb{S}_n$-module V is defined by $\mathcal{S}_V(m) = \mathcal{S}_{\mathrm{ch}_{\mathbb{S}_n}(V)}(m)$.

1.2 | Characteristic polynomials of varieties

In this paper, we work out the characteristic polynomials of examples from geometry, especially the cohomology of varieties equipped with an action of $\mathbb{S}_n$. These include the n-fold products of the projective spaces (cf. §4.1) and the GIT moduli space of n points on $\mathbb{P}^1$ for odd n (cf. §4.2). In both cases, the $\mathbb{S}_n$-action that permutes the n ordered points induces graded $\mathbb{S}_n$-representations on the cohomology. We compute the characteristic polynomials of these representations and establish suitable log-concavity.

Another remarkable example is a regular semisimple *Hessenberg variety* X_h introduced in [12]. It is a subvariety in the flag variety of $\mathbb{C}^n$, defined by a function $h : \{1, \ldots, n\} \to \{1, \ldots, n\}$ satisfying $h(i) \geqslant i$ and $h(i) \geqslant h(i-1)$ for all $i > 1$. There is a natural action of $\mathbb{S}_n$ on its cohomology $H(X_h)$ by [30] whose characteristic polynomial satisfies

$$\omega \mathrm{ch}_{\mathbb{S}_n} H(X_h) = \mathsf{csf}_{\Gamma_h} \tag{1.2}$$

where ω is the involution on Λ that sends p_n to $(-1)^n \mathsf{p}_n$ and Γ_h is the incomparability graph associated to h (cf. Theorem 4.13). While this equality can be written in a graded version, we stick to this simpler form here. In particular, applying $\mathcal{S}$ to (1.2), we get

$$(-1)^n \mathcal{S}_{H(X_h)}(-m) = \mathcal{X}_{\Gamma_h}(m). \tag{1.3}$$

By [17], the characteristic polynomial $\mathcal{S}_{H(X_h)}(m)$ of the Hessenberg variety is log-concave. See §4.3 for further details.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1.3 | Characteristic polynomial of $\overline{\mathcal{M}}_{0,n}$

The main focus of this paper lies on the moduli space $\overline{\mathcal{M}}_{0,n}$ of stable rational curves with n marked points. See [10, 11] for basic facts about this moduli space and its cohomology. This is a case where the characteristic polynomial is not the chromatic polynomial of any graph. The $\mathbb{S}_n$-action on $\overline{\mathcal{M}}_{0,n}$ permuting the n marked points gives rise to a graded $\mathbb{S}_n$-representation on its cohomology $H^*(\overline{\mathcal{M}}_{0,n})$. Consequently, we obtain a bivariate polynomial

$$\mathscr{P}_n(m,t) = \sum_{k=0}^{n-3} \mathscr{P}_{n,k}(m)t^k \in \mathbb{Q}[m,t], \quad \text{where } \mathscr{P}_{n,k}(m) = \mathcal{S}_{\mathrm{ch}_{\mathbb{S}_n} H^{2k}(\overline{\mathcal{M}}_{0,n})}(m)$$

for each $k \geqslant 0$. We write the generating series of $\mathscr{P}_n$ as

$$\mathscr{P} = 1 + mq + \frac{m(m+1)}{2}q^2 + \sum_{n\geqslant 3} \mathscr{P}_n(m,t)q^n \in \mathbb{Q}[m,t]\llbracket q\rrbracket$$

for a formal parameter q recording n.

Recursive formulas for computing $\mathscr{P}_n$ can be deduced from [10, 11], where we developed recursive formulas for $\mathrm{ch}_{\mathbb{S}_n} H^*(\overline{\mathcal{M}}_{0,n})$ that use only arithmetic operations and plethysm. See [11, Theorem 1.3] or §5.1 and 5.2.

To present explicit formulas, let $\mathscr{Q}_{n,k}$ denote the characteristic polynomial

$$\mathscr{Q}_{n,k}(m) = \mathcal{S}_{\mathrm{ch}_{\mathbb{S}_n} H^{2k}(\overline{\mathcal{M}}_{0,n+1})}(m)$$

of the $\mathbb{S}_n$-representation on $H^{2k}(\overline{\mathcal{M}}_{0,n+1})$ induced by the action that permutes the first n markings while keeping the last fixed. The advantage of $\mathscr{Q}_{n,k}$ is that there is a rather straightforward combinatorial formula

$$\mathrm{ch}_{\mathbb{S}_n} H^{2k}(\overline{\mathcal{M}}_{0,n+1}) = \sum_{T\in\mathscr{T}_{n,k}} \mathrm{ch}_{\mathbb{S}_n}(U_T)$$

for the Frobenius characteristic in terms of *weighted rooted trees* T and the associated permutation representations U_T by [10, Proposition 5.12]. In particular, we have

$$\mathscr{Q}_{n,k} = \sum_{T\in\mathscr{T}_{n,k}} \mathcal{S}_{\mathrm{ch}_{\mathbb{S}_n}(U_T)}. \tag{1.4}$$

For precise definitions of T, $\mathscr{T}_{n,k}$, and U_T, we refer the reader to §5.

We define the generating series of $\mathscr{Q}_{n,k}$ by

$$\mathscr{Q} = 1 + mq + \sum_{n\geqslant 2} \mathscr{Q}_n(m,t)q^n \in \mathbb{Q}[m,t]\llbracket q\rrbracket, \quad \text{where } \mathscr{Q}_n(m,t) = \sum_{k=0}^{n-2} \mathscr{Q}_{n,k}(m)t^k.$$

Moreover, we consider a partial sum

$$\mathscr{Q}^+_{n,k} = \sum_{T\in\mathscr{T}^+_{n,k}} \mathcal{S}_{\mathrm{ch}_{\mathbb{S}_n}(U_T)}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of (1.4) where $\mathcal{T}^+_{n,k} \subset \mathcal{T}_{n,k}$ consists of weighted rooted trees having positive weights at the root vertex. The generating series for $\mathcal{Q}^+_{n,k}$ is defined as

$$\mathcal{Q}^+ = mq + \sum_{n\geqslant 2} \mathcal{Q}^+_n q^n \in \mathbb{Q}[m,t][\![q]\!], \quad \text{where } \mathcal{Q}^+_n = \sum_{k=0}^{n-2} \mathcal{Q}^+_{n,k} t^k.$$

Then, $\mathcal{P}$ can be computed from $\mathcal{Q}$, $\mathcal{Q}$ can be computed from $\mathcal{Q}^+$ and $\mathcal{Q}^+$ can be computed by its own recursion as follows (cf. §5).

Theorem 1.1 (Theorem 5.11). *$\mathcal{P}$ is determined by the following.*

(1) *$\mathcal{P}$ and $\mathcal{Q}$ are related by the formula*

$$(1+t)\mathcal{P} = (1+t+mqt)\mathcal{Q} - \frac{t}{2}(\mathcal{Q}^2 - \mathcal{Q}^{[2]}),$$

where $\mathcal{Q}^{[2]}(m,q,t) = \mathcal{Q}(m,q^2,t^2)$.

(2) *$\mathcal{Q}$ is the plethystic exponential of $\mathcal{Q}^+$: $\mathcal{Q} = \mathrm{Exp}(\mathcal{Q}^+)$.*

(3) *$\mathcal{Q}^+$ satisfies the following two equivalent formulas.*

$$\text{(Recursive)} \quad \mathcal{Q}^+ = mq + \sum_{r\geqslant 3} \left(\sum_{i=1}^{r-2} t^i\right)(\mathsf{h}_r \circ \mathcal{Q}^+),$$

$$\text{(Exponential)} \quad \mathrm{Exp}(t\mathcal{Q}^+) = t^2\,\mathrm{Exp}(\mathcal{Q}^+) + (1-t)(1+t+mqt).$$

An equivalent polynomial version of this theorem is the following.

Corollary 1.2 (Corollary 5.12). *For $n \geqslant 3$, $\mathcal{P}_n$ and $\mathcal{Q}_n$ are related by*

$$(1+t)\mathcal{P}_n = \mathcal{Q}_n - \frac{1}{2}t\left(\sum_{h=2}^{n-2} \mathcal{Q}_h \mathcal{Q}_{n-h} - \mathcal{Q}^{[2]}_{\frac{n}{2}}\right),$$

where $\mathcal{Q}_{\frac{n}{2}} = 0$ for n odd, and $\mathcal{Q}^{[2]}_{\frac{n}{2}}(m,t) = \mathcal{Q}_{\frac{n}{2}}(m,t^2)$.

Set $\mathcal{Q}^+_1 = m$. For $n \geqslant 2$, $\mathcal{Q}^+_n$ and $\mathcal{Q}_n$ satisfy

$$\mathcal{Q}^+_n = \sum_{\lambda\vdash n} \left(\sum_{i=1}^{\ell(\lambda)-2} t^i\right) \prod_{j=1}^{m} \left(\mathsf{h}_{r_j} \circ \mathcal{Q}^+_{n_j}\right) \quad \text{and} \quad \mathcal{Q}_n = \sum_{\lambda\vdash n} \prod_{j=1}^{m} \left(\mathsf{h}_{r_j} \circ \mathcal{Q}^+_{n_j}\right),$$

where n_j denote the parts of λ with multiplicities r_j so that $\lambda = (n_1^{r_1}, \dots, n_s^{r_s})$ with $n_1 > \cdots > n_s > 0$ and $\ell(\lambda) = \sum_{j=1}^{s} r_j$.

By using the above formulas, we implemented a computer program to explicitly compute the characteristic polynomials, and carried out computations for $n \leqslant 72$. The computational results allow us to investigate the log-concavity properties of the characteristic polynomials. Since $\mathcal{P}_n$ and $\mathcal{Q}_n$ are bivariate polynomials, various types of log-concavity can be considered (cf. §2), and our computations indicate that they exhibit several such properties. Based on this numerical

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

evidence, we propose a conjecture concerning the log-concavity of the characteristic polynomials (Conjecture 5.14).

1.4 | Asymptotic formulas and log-concavity

We study asymptotic formulas of $\mathscr{P}_{n,k}(m)$, as n grows, in two directions:

(1) the values $\mathscr{P}_{n,k}(m_0)$ at a given $m_0 \geqslant 1$ for given $k \geqslant 0$;
(2) the coefficients of m^{n-j} in $\mathscr{P}_{n,k}$, denoted by $\mathscr{P}_{n,k}^{n-j}$, for given $j, k \geqslant 0$.

These formulas lead us to corresponding asymptotic log-concavity properties that are far reaching generalizations of the corresponding results in [3, 11].

For the asymptotics, we use the constants

$$c_k = \frac{(k+1)^{k-1}}{k!} \quad \text{and} \quad d_k = \frac{(k+1)^{k-2}}{k!}.$$

Theorem 1.3 (Theorem 6.1). *Fix $m_0 \geqslant 1$ and $k \geqslant 0$. Then as n grows,*

$$\mathscr{Q}_{n,k}(m_0) = \frac{c_k}{((k+1)m_0-1)!} n^{(k+1)m_0-1} + o(n^{(k+1)m_0-1}) \text{ and}$$
$$\mathscr{P}_{n,k}(m_0) = \frac{d_k}{((k+1)m_0-1)!} n^{(k+1)m_0-1} + o(n^{(k+1)m_0-1}).$$

From this, the following asymptotic log-concavity follows.

Corollary 1.4 (Corollary 6.9). *Fix $m_0, k \geqslant 1$. Then,*

$$\mathscr{P}_{n,k}(m_0)^2 \geqslant \mathscr{P}_{n,k-1}(m_0)\mathscr{P}_{n,k+1}(m_0) \text{ and}$$
$$\mathscr{Q}_{n,k}(m_0)^2 \geqslant \mathscr{Q}_{n,k-1}(m_0)\mathscr{Q}_{n,k+1}(m_0)$$

for sufficiently large n. Moreover, when $m_0 > 1$, the inequalities are strict.

Theorem 1.3 and Corollary 1.4 generalize the asymptotic formulas and the asymptotic log-concavity for the Betti numbers of $\overline{\mathcal{M}}_{0,n}/\mathbb{S}_n$ in [11, §6] that correspond to the case of $m_0 = 1$ (cf. Remark 6.2).

Next, we investigate the asymptotic behaviors of the coefficients in $\mathscr{P}_{n,k}$ and $\mathscr{Q}_{n,k}$. Recall that the *signless Stirling number of the first kind* $c(n, j)$ is the number of $\sigma \in \mathbb{S}_n$ with j disjoint cycles.

Write $\mathscr{P}_{n,k}(m) = \sum_j \mathscr{P}_{n,k}^j m^j$ and $\mathscr{Q}_{n,k}(m) = \sum_j \mathscr{Q}_{n,k}^j m^j$.

Theorem 1.5 (Theorem 6.6). *Let $j, k \in \mathbb{Z}_{\geqslant 0}$. Then, as n grows,*

$$\mathscr{Q}_{n,k}^{n-j} = c_k \cdot \frac{(k+1)^{n-j} c(n, n-j)}{n!} + o\left(\frac{(k+1)^n n^{2j}}{n!}\right) \text{ and}$$
$$\mathscr{P}_{n,k}^{n-j} = d_k \cdot \frac{(k+1)^{n-j} c(n, n-j)}{n!} + o\left(\frac{(k+1)^n n^{2j}}{n!}\right).$$

Based on this, we obtain the following two asymptotic log-concavity results about the coefficients when we vary the powers of t and m, respectively. In fact, in the first case, they are asymptotically *ultra* log-concave.

Corollary 1.6 (Corollary 6.10). *Fix $j \geqslant 0$ and $k \geqslant 1$. Then,*

$$\left(\frac{\mathscr{P}_{n,k}^{n-j}}{\binom{n-3}{k}}\right)^2 \geqslant \left(\frac{\mathscr{P}_{n,k-1}^{n-j}}{\binom{n-3}{k-1}}\right)\left(\frac{\mathscr{P}_{n,k+1}^{n-j}}{\binom{n-3}{k+1}}\right) \text{ and}$$

$$\left(\frac{\mathscr{Q}_{n,k}^{n-j}}{\binom{n-2}{k}}\right)^2 \geqslant \left(\frac{\mathscr{Q}_{n,k-1}^{n-j}}{\binom{n-2}{k-1}}\right)\left(\frac{\mathscr{Q}_{n,k+1}^{n-j}}{\binom{n-2}{k+1}}\right)$$

for any sufficiently large n.

Corollary 1.7 (Corollary 6.11). *Fix $j \geqslant 1$ and $k \geqslant 0$. Then,*

$$\left(\mathscr{P}_{n,k}^{n-j}\right)^2 \geqslant \left(\mathscr{P}_{n,k}^{n-j-1}\right) \cdot \left(\mathscr{P}_{n,k}^{n-j+1}\right) \text{ and}$$

$$\left(\mathscr{Q}_{n,k}^{n-j}\right)^2 \geqslant \left(\mathscr{Q}_{n,k}^{n-j-1}\right) \cdot \left(\mathscr{Q}_{n,k}^{n-j+1}\right)$$

for any sufficiently large n.

Theorem 1.5 and Corollary 1.6 generalize the results by Aluffi–Chen–Marcolli in [3] on the Betti numbers of $\overline{\mathcal{M}}_{0,n+1}$, which correspond to the case where $j = 0$ (Remark 6.7). It remains open whether the ultra log-concavity for $j = 0$ in Corollary 1.6 holds for all n, whereas explicit formulas for the Betti numbers have recently been obtained in [4, 13].

All cohomology groups in this paper are singular cohomology with rational coefficients and all $\mathbb{S}_n$-representations are defined over $\mathbb{Q}$.

2 | TWO-DIMENSIONAL LOG-CONCAVITY

A sequence $\{a_k\}$ of nonnegative real numbers is called *log-concave* if

$$a_k^2 \geqslant a_{k-1}a_{k+1} \quad \text{for all } k. \tag{2.1}$$

The log-concavity extends to arbitrary real sequences by taking the sequence of absolute values $|a_k|$. A nonnegative finite sequence $\{a_k\}_{k=0}^n$ is *ultra log-concave* if the normalized sequence

$$\left\{\frac{a_k}{\binom{n}{k}}\right\}_{0\leqslant k\leqslant n}$$

is log-concave. In what follows, any property defined in terms of log-concavity is similarly extended to define the corresponding notion of ultra log-concavity.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

A polynomial $f(m) = \sum_{k=0}^{n} a_k m^k \in \mathbb{R}_{\geqslant 0}[m]$ with nonnegative coefficients is said to be *log-concave* if the sequence $\{a_k\}$ is log-concave. We say $f(m)$ has *no internal zeros* if there do not exist integers $i < j < k$ such that $a_i \neq 0$, $a_j = 0$ and $a_k \neq 0$.

In this section, we generalize the notion of log-concavity to a two-dimensional array $\{A_k^j\}_{0 \leqslant j,k \leqslant N}$ of nonnegative real numbers or to a bivariate polynomial

$$A(m,t) = \sum_{k \geqslant 0} \sum_{j \geqslant 0} A_k^j m^j t^k \ \in \mathbb{R}_{\geqslant 0}[m,t]. \tag{2.2}$$

Let $A_k(m) = \sum_{j \geqslant 0} A_k^j m^j$ and $A^j(t) = \sum_{k \geqslant 0} A_k^j t^k$ so that we have

$$A(m,t) = \sum_k A_k(m) t^k = \sum_j A^j(t) m^j.$$

From the matrix $\{A_k^j\}$, there are many ways to extract a sequence of numbers and define log-concavity. In this paper, we will consider the following natural choices.

Definition 2.1. Let $A(m,t) = \sum_{j,k \geqslant 0} A_k^j m^j t^k \ \in \mathbb{R}_{\geqslant 0}[m,t]$ and $t_0, m_0 \in \mathbb{R}$.

(1) $A(m,t)$ is *length log-concave in degree* k if $A_k(m) = \sum_{j \geqslant 0} A_k^j m^j \in \mathbb{R}_{\geqslant 0}[m]$ is log-concave.
(2) $A(m,t)$ is *length log-concave at* $t = t_0$ if $A(m,t_0) \in \mathbb{R}_{\geqslant 0}[m]$ is log-concave.
(3) $A(m,t)$ is *degree log-concave in length* j if $A^j(t) = \sum_{k \geqslant 0} A_k^j t^k \in \mathbb{R}_{\geqslant 0}[t]$ is log-concave.
(4) $A(m,t)$ is *degree log-concave at* $m = m_0$ if $A(m_0,t) \in \mathbb{R}_{\geqslant 0}[t]$ is log-concave.

This definition extends to arbitrary bivariate real polynomials $A(m,t) \in \mathbb{R}[m,t]$ by taking the absolute values $|A_k^j|$ of the coefficients.

We call t the *degree* variable and m the *length* variable since we will see below that if $A(m,t)$ comes from the cohomology $H^*(X)$ of a space X acted on by the symmetric group $\mathbb{S}_n$, the exponent of the variable m corresponds to the length of a partition and the exponent of the variable t keeps track of the cohomological degree.

Remark 2.2. In §3.4, we will associate a polynomial $\mathcal{S}_V(m) \in \mathbb{Q}[m]$ of degree n to each finite-dimensional $\mathbb{S}_n$-representation $V \neq 0$ by applying the principal specialization map in [29, §7.8] to the Frobenius characteristic of V. By Lemma 3.10, the coefficient of m^n in $\mathcal{S}_V(m)$ is $\frac{1}{n!}\dim V$ and the dimension of the invariant part equals

$$\dim V^{\mathbb{S}_n} = \mathcal{S}_V(1).$$

When the $\mathbb{S}_n$-representation $V = \bigoplus_{k \geqslant 0} V_k$ is graded, we associate the polynomial

$$\mathcal{S}_V(m,t) = \sum_k \mathcal{S}_{V_k}(m) t^k \in \mathbb{Q}[m,t].$$

In this case, the degree log-concavity in length n is equivalent to the log-concavity of the sequence $\{\dim V_k\}_k$ while the degree log-concavity at $m = 1$ is equivalent to the log-concavity of the sequence $\{\dim V_k^{\mathbb{S}_n}\}_k$ of the dimensions of the invariant parts.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In particular, when

$$V = \bigoplus_{k\geqslant 0} H^{2k}(\overline{\mathcal{M}}_{0,n})$$

is the cohomology of the moduli space $\overline{\mathcal{M}}_{0,n}$ of n-pointed stable curves of genus 0, we note that the degree log-concavity in length n is the log-concavity of the Betti numbers proved asymptotically by Aluffi–Chen–Marcolli [3], while the degree log-concavity at $m = 1$ is the log-concavity of the invariant part proved asymptotically in [11]. Below we will find more log-concavity results about the graded $\mathbb{S}_n$-module $H^*(\overline{\mathcal{M}}_{0,n})$ in terms of the log-concavities in Definition 2.1.

Remark 2.3. In §3.2, we will associate a polynomial $\mathcal{S}_F(m) \in \mathbb{Q}[m]$ of degree at most n to each symmetric function $F \in \Lambda_n$ of degree n by the principal specialization map in [29, §7.8]. When $F = \sum_{k\geqslant 0} F_k t^k \in \Lambda_n[t]$ with $F_k \in \Lambda_n$, we associate $\mathcal{S}_F(m,t) = \sum_{k\geqslant 0} \mathcal{S}_{F_k}(m)t^k \in \mathbb{Q}[m,t]$.

Given a graph Γ, we have the chromatic quasi-symmetric function $\mathrm{cqsf}_\Gamma(t)$ defined in [23]. When Γ is the graph associated to a Hessenberg function in §4.3, $F = \mathrm{cqsf}_\Gamma(t) \in \Lambda_n[t]$ and $\mathcal{S}_F(m,1) = \mathcal{S}_{\mathrm{csf}_\Gamma}(m)$ is the usual chromatic polynomial $\mathcal{X}_\Gamma(m)$ of Γ. Therefore, the length log-concavity at $t = 1$ of $\mathcal{S}_F(m,t)$ is the log-concavity of the chromatic polynomial of the graph Γ that is a special case of the celebrated theorem of June Huh in [17].

We end this section with the following well-known results that will be used below.

Lemma 2.4 [25, Proposition 2]. *If $f, g \in \mathbb{R}_{\geqslant 0}[m]$ are log-concave with no internal zeros, then so is $fg \in \mathbb{R}_{\geqslant 0}[m]$.*

Lemma 2.5 [25, Theorem 2]. *If $f \in \mathbb{R}_{\geqslant 0}[m]$ has only real roots, then f is log-concave.*

3 | CHARACTERISTIC POLYNOMIALS OF SYMMETRIC FUNCTIONS AND REPRESENTATIONS

In this section, we investigate useful properties of the principal specialization map or the *Stanley map* in [29, §7.8] that associates a polynomial $\mathcal{S}_F(m) \in \mathbb{Q}[m]$ to each symmetric function $F \in \Lambda_n$ of degree n. A remarkable feature of the Stanley map

$$\mathcal{S} : \Lambda_n \longrightarrow \mathbb{Q}[m], \quad F \mapsto \mathcal{S}_F(m)$$

is that it sends the plethysm operation on Λ_n to the composition on $\mathbb{Q}[m]$ (Proposition 3.2). In fact, this property uniquely determines the Stanley map (Theorem 3.3). Since the plethysm operation usually causes computational difficulties, the Stanley map levitates computational burdens while keeping core information, just like the characteristic polynomial of a matrix. In this vein, we call $\mathcal{S}_F(m) \in \mathbb{Q}[m]$ the *characteristic polynomial* of F (Definition 3.1). When the symmetric function is the chromatic symmetric function of a graph defined by Stanley in [26], its characteristic polynomial is the usual chromatic polynomial of Birkhoff's, which is log-concave by [17].

By applying the Stanley map to the Frobenius characteristic $\mathrm{ch}(V) \in \Lambda_n$ of an $\mathbb{S}_n$-module V, we obtain a polynomial $\mathcal{S}_V(m) \in \mathbb{Q}[m]$ which we call the *characteristic polynomial* of V. When

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$V = \bigoplus_{k\geqslant 0} V_k$ is graded, its characteristic polynomial

$$\mathcal{S}_V(m,t) = \sum_k \mathcal{S}_{V_k}(m)t^k \in \mathbb{Q}[m,t]$$

allows us to consider the two-dimensional log-concavities in Definition 2.1. In the subsequent sections, we will find interesting examples where the log-concavities hold.

3.1 | Symmetric functions and plethysm

In this subsection, we set up our notation on symmetric functions and recall basic facts about plethysm. For further details and undefined terms, we refer the reader to [21].

Let

$$\Lambda = \varprojlim \mathbb{Q}[\![x_1, \dots, x_n]\!]^{\mathbb{S}_n}$$

be the power series ring of symmetric functions in variables $x_1, x_2, \dots$, with *rational* coefficients. Let $\Lambda_n \subset \Lambda$ denote the subspace of elements of degree n, so that $\Lambda = \prod_{n\geqslant 0} \Lambda_n$. For an integer $n \geqslant 1$, let

$$\mathsf{p}_n = \sum_{i\geqslant 1} x_i^n, \quad \mathsf{h}_n = \sum_{1\leqslant i_1\leqslant\cdots\leqslant i_n} x_{i_1}\cdots x_{i_n} \quad \text{and} \quad \mathsf{e}_n = \sum_{1\leqslant i_1<\cdots<i_n} x_{i_1}\cdots x_{i_n}$$

be the nth power sum symmetric function, homogeneous symmetric function and elementary symmetric function, respectively. For convenience, we set $\mathsf{p}_0 = \mathsf{h}_0 = \mathsf{e}_0 = 1$. For a partition $\lambda = (\lambda_1, \dots, \lambda_\ell) \vdash n$, let

$$\mathsf{p}_\lambda = \mathsf{p}_{\lambda_1}\cdots\mathsf{p}_{\lambda_\ell}, \quad \mathsf{h}_\lambda = \mathsf{h}_{\lambda_1}\cdots\mathsf{h}_{\lambda_\ell} \quad \text{and} \quad \mathsf{e}_\lambda = \mathsf{e}_{\lambda_1}\cdots\mathsf{e}_{\lambda_\ell}.$$

Then, each of $\{\mathsf{p}_\lambda\}_{\lambda\vdash n}$, $\{\mathsf{h}_\lambda\}_{\lambda\vdash n}$ and $\{\mathsf{e}_\lambda\}_{\lambda\vdash n}$ is a basis for Λ_n.

For a partition λ, define

$$z_\lambda = \prod_{i\geqslant 1} i^{m_i} m_i!,$$

where m_i is the number of parts of λ equal to i. We let $z_0 = 1$. Define a symmetric bilinear form $\langle -,-\rangle$ on Λ by

$$\langle \mathsf{p}_\lambda, \mathsf{p}_\mu\rangle = z_\lambda \delta_{\lambda\mu} \tag{3.1}$$

so that $\{\mathsf{p}_\lambda\}$ is an orthogonal basis.

There is a ring homomorphism $\omega : \Lambda \to \Lambda$ defined by

$$\omega(\mathsf{p}_n) = (-1)^{n-1}\mathsf{p}_n. \tag{3.2}$$

Clearly ω is an isometric involution and satisfies $\omega(\mathsf{e}_\lambda) = \mathsf{h}_\lambda$ for any $\lambda \vdash n$.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

On Λ, there is an associative binary operation $\circ$, called the *plethysm*, which is uniquely determined by the following properties: for $F, G, H \in \Lambda$,

(1) $F \circ \mathsf{p}_n = \mathsf{p}_n \circ F = F(x_1^n, x_2^n, \ldots)$ for $n \geqslant 1$;
(2) $(F+G) \circ H = F \circ H + G \circ H$;
(3) $(FG) \circ H = (F \circ H)(G \circ H)$.

Note that p_1 is the two-sided identity with respect to $\circ$ by (1). This naturally extends to such an operation $\circ$ on $\Lambda[\![t]\!]$ satisfying (2) and (3) with Λ replaced by $\Lambda[\![t]\!]$ as well as

(1′) $F \circ \mathsf{p}_n = \mathsf{p}_n \circ F = F(t^n, x_1^n, x_2^n, \ldots)$ for $n \geqslant 1$, and $t \circ F = t$.

For $F \in \Lambda$ or $\Lambda[\![t]\!]$ with no constant term, the *plethystic exponential* of F is defined to be

$$\mathrm{Exp}(F) = \sum_{r \geqslant 0} \mathsf{h}_r \circ F = \exp \sum_{r \geqslant 1} \frac{1}{r} \mathsf{p}_r \circ F, \tag{3.3}$$

where the second equality holds since $\sum_{r\geqslant 0} \mathsf{h}_r = \exp \sum_{r \geqslant 1} \frac{1}{r} \mathsf{p}_r$.

3.2 | Characteristic polynomial of a symmetric function

In this subsection, we introduce the Stanley map and the characteristic polynomials of symmetric functions. We further characterize the Stanley map as the only algebra homomorphism that converts plethysm to composition.

Definition 3.1. Let $F \in \Lambda_n$ be a symmetric function of degree n such that $F = \sum_{\lambda \vdash n} c_\lambda \mathsf{p}_\lambda$. The *characteristic polynomial* of F is defined to be

$$\mathcal{S}_F(m) = \sum_{\lambda \vdash n} c_\lambda m^{\ell(\lambda)} \quad \in \mathbb{Q}[m], \tag{3.4}$$

where $\ell(\lambda)$ denotes the length of a partition λ. For $F = \sum_k F_k t^k \in \Lambda_n[t]$ with $F_k \in \Lambda_n$ and a formal variable t, we define

$$\mathcal{S}_F(m,t) = \sum_k \mathcal{S}_{F_k}(m) t^k \quad \in \mathbb{Q}[m,t].$$

Note that the degree of $\mathcal{S}_F$ is at most n. It is easy to see that for any positive integer m_0, $\mathcal{S}_F(m_0)$ is obtained from F by the substitution (1.1). By (3.1), (3.4) can be also written as

$$\mathcal{S}_F(m) = \sum_{\lambda \vdash n} z_\lambda^{-1} \langle F, \mathsf{p}_\lambda \rangle m^{\ell(\lambda)}. \tag{3.5}$$

By Definition 3.1, we have the map

$$\mathcal{S} : \Lambda_n \longrightarrow \mathbb{Q}[m], \quad F \mapsto \mathcal{S}_F(m) \tag{3.6}$$

and its extension

$$\widehat{\mathcal{S}} : \Lambda \longrightarrow \mathbb{Q}[m][\![q]\!], \quad (F \in \Lambda_n) \mapsto \mathcal{S}_F(m) q^n, \tag{3.7}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where q is a formal variable recording the degree n of F. Note that (3.7) is a $\mathbb{Q}$-algebra homomorphism, called the *principal specialization* in [29, §7.8], or the *Stanley map*.

A remarkable feature of (3.6) and (3.7) is that they convert plethysm to composition.

Proposition 3.2. *For any $F \in \Lambda_n$ and $G \in \Lambda_{n'}$ with $n, n' \geqslant 0$,*

$$\mathcal{S}_{F\circ G}(m) = (\mathcal{S}_F \circ \mathcal{S}_G)(m) = \mathcal{S}_F(\mathcal{S}_G(m)). \tag{3.8}$$

Proof. By definition, $\mathcal{S}_{\mathsf{p}_r\circ G} = \mathcal{S}_G$ for any $r \geqslant 1$ and $G \in \Lambda_{n'}$. Moreover, if $F = \sum_{\lambda\vdash n} a_\lambda \mathsf{p}_\lambda$, then $\mathcal{S}_F(m) = \sum_{\lambda\vdash n} a_\lambda m^{\ell(\lambda)}$. Consequently,

$$\mathcal{S}_{F\circ G}(m) = \sum_{\lambda\vdash n} a_\lambda \prod_{i=1}^{\ell(\lambda)} \mathcal{S}_{\mathsf{p}_{\lambda_i}\circ G}(m) = \sum_{\lambda\vdash n} a_\lambda (\mathcal{S}_G(m))^{\ell(\lambda)} = \mathcal{S}_F(\mathcal{S}_G(m))$$

for any $G \in \Lambda_{n'}$, where $\lambda = (\lambda_1, \dots, \lambda_{\ell(\lambda)})$. □

If we extend the composition $(f\circ g)(m) = f(g(m))$ to an operation on $\mathbb{Q}[m][\![q]\!]$ by

$$\left(\sum_i f_i(m)q^i\right)\circ\left(\sum_j g_j(m)q^j\right) = \sum_{i,j} f_i(g_j(m))q^{ij}. \tag{3.9}$$

Proposition 3.2 is equivalent to

$$\widehat{\mathcal{S}}_{F\circ G} = \widehat{\mathcal{S}}_F \circ \widehat{\mathcal{S}}_G \in \mathbb{Q}[m][\![q]\!] \quad \text{for } F, G \in \Lambda. \tag{3.10}$$

Conversely, (3.8) or (3.10) determines $\widehat{\mathcal{S}}$ and thus $\mathcal{S}$.

Theorem 3.3. *The map $\widehat{\mathcal{S}} : \Lambda \to \mathbb{Q}[m][\![q]\!]$ is the unique $\mathbb{Q}$-algebra homomorphism that sends Λ_n to $\mathbb{Q}[m]q^n$ for every n and satisfies (3.10).*

Proof. Let $\widehat{\mathcal{S}}'$ be another such homomorphism and let $\mathcal{S}' : \Lambda_n \to \mathbb{Q}[m]$ be defined by $\widehat{\mathcal{S}}'_F = \mathcal{S}'_F(m)q^n$ for $F \in \Lambda_n$. Since $\mathsf{p}_n \circ a = a$ and $\mathcal{S}'_a = a$ for any $n \geqslant 1$ and $a \in \Lambda_0 = \mathbb{Q}$,

$$\widehat{\mathcal{S}}'_{\mathsf{p}_n}(a) = \widehat{\mathcal{S}}'_{\mathsf{p}_n}(\widehat{\mathcal{S}}'_a(m)) = \widehat{\mathcal{S}}'_{\mathsf{p}_n\circ a}(m) = \mathcal{S}'_a(m)q^n = aq^n.$$

Therefore, $\mathcal{S}'_{\mathsf{p}_n}(m) = m$ and $\widehat{\mathcal{S}}'_{\mathsf{p}_n} = mq^n$ for $n \geqslant 1$. The uniqueness follows from this because $\widehat{\mathcal{S}}'$ is a $\mathbb{Q}$-algebra homomorphism. □

All these extend to a graded version with a new variable t as

$$\widehat{\mathcal{S}} : \Lambda[\![t]\!] \longrightarrow \mathbb{Q}[m][\![q,t]\!], \quad \mathsf{p}_n t^k \mapsto mq^n t^k. \tag{3.11}$$

By abuse of notation, we denote this map also by $\widehat{\mathcal{S}}$. If we extend the composition to an operation $\mathbb{Q}[m][\![q]\!] \times \mathbb{Q}[m][\![q,t]\!] \to \mathbb{Q}[m][\![q,t]\!]$ by

$$\left(\sum_i f_i(m)q^i\right)\circ\left(\sum_{j,k} g_{jk}(m)q^j t^k\right) = \sum_{i,j,k} f_i(g_{jk}(m))q^{ij}t^{ik}. \tag{3.12}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Proposition 3.2 upgrades to

$$\widehat{\mathcal{S}}_{F\circ G} = \widehat{\mathcal{S}}_F \circ \widehat{\mathcal{S}}_G \in \mathbb{Q}[m][\![q,t]\!] \quad \text{for } F \in \Lambda,\ G \in \Lambda[\![t]\!]. \tag{3.13}$$

Theorem 3.4. *The map (3.11) is the unique $\mathbb{Q}$-algebra homomorphism that sends $\Lambda_n[\![t]\!]$ to $\mathbb{Q}[m][\![t]\!]q^n$ for every n and satisfied (3.13).*

Proof. The proof is basically the same as that for Theorem 3.3. □

Finally, we observe that $\mathcal{S}$ satisfies a duality with respect to the involution (3.2) on Λ (cf. [28, Exercise 24]).

Lemma 3.5 (ω-duality). *For $F \in \Lambda_n$, $\mathcal{S}_{\omega F}(m) = (-1)^n \mathcal{S}_F(-m)$.*

Proof. This follows immediately from $\omega \mathsf{p}_\lambda = (-1)^{n-\ell(\lambda)}\mathsf{p}_\lambda$ for $\lambda \vdash n$. □

3.3 | Plethysm on polynomials

For convenience, we introduce the plethysm operation on polynomials as follows.

Definition 3.6. For any $F \in \Lambda_n$, we define the plethysm

$$F\circ(-) : \mathbb{Q}[m][\![q,t]\!] \longrightarrow \mathbb{Q}[m][\![q,t]\!], \quad F\circ\big(f(m)q^i t^j\big) = \mathcal{S}_F(f(m))q^{ni}t^{nj}.$$

For $F \in \Lambda$, we extend it linearly.

It is straightforward to check that $F\circ(-)$ is uniquely characterized by the following properties: for $f \in \mathbb{Q}[m][\![q,t]\!]$ and $F, G \in \Lambda$,

(1) $\mathsf{p}_n \circ f = f^{[n]}$ for $n \geqslant 1$;
(2) $(F+G)\circ f = F\circ f + G\circ f$,
(3) $(FG)\circ f = (F\circ f)(G\circ f)$,

where $f^{[n]}(m,q,t) = f(m,q^n,t^n)$. Note that $\mathsf{p}_n\circ(-)$ leaves m unchanged, while q and t are raised to the nth power. If $F = \sum_{\lambda\vdash n} c_\lambda \mathsf{p}_\lambda$, then

$$F\circ f = \sum_{\lambda\vdash n} c_\lambda f^{[\lambda]},$$

where we write $f^{[\lambda]} := f^{[\lambda_1]}\cdots f^{[\lambda_\ell]}$ for $\lambda = (\lambda_1,\dots,\lambda_\ell)$.

For $f \in \mathbb{Q}[m][\![q,t]\!]$ with no constant term, as in (3.3), we define the *plethystic exponential* of f to be

$$\mathrm{Exp}(f) = \sum_{r\geqslant 0} \mathsf{h}_r \circ f = \exp\sum_{r\geqslant 1}\frac{1}{r}\mathsf{p}_r\circ f = \exp\sum_{r\geqslant 1}\frac{1}{r}f^{[r]}. \tag{3.14}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Example 3.7 (Monomial). For integers a, j, n, k with $j, n, k \geqslant 0$, we have

$$\begin{aligned}\mathrm{Exp}(am^j q^n t^k) &= \exp \sum_{r\geqslant 1} \frac{1}{r} am^j q^{rn} t^{rk} \\ &= \exp(-am^j \log(1-q^n t^k)) = (1-q^n t^k)^{-am^j},\end{aligned}$$

provided $(n,k) \neq (0,0)$.

Remark 3.8 (Alternative description of $\mathrm{Exp}(f)$). Recall that

$$\mathsf{h}_n = \sum_{\lambda \vdash n} z_\lambda^{-1} \mathsf{p}_\lambda \tag{3.15}$$

for any $n \geqslant 1$. Then, $\mathrm{Exp}(f)$ can also be written as

$$\mathrm{Exp}(f) = \sum_{r\geqslant 0} \mathsf{h}_r \circ f = \sum_{\lambda \in \mathrm{Par}} z_\lambda^{-1} f^{[\lambda]},$$

where Par denotes the set of partitions of nonnegative integers and $f^{[0]} = 1$.

3.4 | Characteristic polynomial of an $\mathbb{S}_n$-representation

In this subsection, we define the characteristic polynomial of an $\mathbb{S}_n$-module V by applying the Stanley map $\mathcal{S}$ to the Frobenius characteristic of V.

Let $R_n = \mathrm{Rep}(\mathbb{S}_n)_{\mathbb{Q}}$ be the $\mathbb{Q}$-vector space spanned by irreducible $\mathbb{S}_n$-representations and let $R_0 = \mathbb{Q}$. Then, $R = \prod_{n\geqslant 0} R_n$ is the ring of virtual representations of the symmetric groups whose ring structure is given by

$$V.W = \mathrm{Ind}_{\mathbb{S}_n \times \mathbb{S}_m}^{\mathbb{S}_{n+m}} V \otimes W$$

for an $\mathbb{S}_n$-module V and an $\mathbb{S}_m$-module W. Then, it is well known that the Frobnenius characteristic map

$$\mathrm{ch} : R \xrightarrow{\cong} \Lambda, \quad \mathrm{ch}(V) = \mathrm{ch}_{\mathbb{S}_n}(V) = \frac{1}{n!} \sum_{\sigma \in \mathbb{S}_n} \mathrm{Tr}_V(\sigma) \mathsf{p}_\sigma \quad \text{for } V \in R_n \tag{3.16}$$

is a ring isomorphism where $\mathsf{p}_\sigma := \mathsf{p}_{\lambda(\sigma)}$ and $\lambda(\sigma) \vdash n$ is the cycle partition of σ. This sends the Specht module S^λ to the Schur function s_λ.

For $\sigma \in \mathbb{S}_n$, let $\ell(\sigma)$ denote the length of the partition $\lambda(\sigma)$.

Definition 3.9. The *characteristic polynomial* of an $\mathbb{S}_n$-representation V is

$$\mathcal{S}_V(m) = \mathcal{S}_{\mathrm{ch}(V)}(m) = \frac{1}{n!} \sum_{\sigma \in \mathbb{S}_n} \mathrm{Tr}_V(\sigma) m^{\ell(\sigma)}.$$

When $V = \bigoplus_{k\in\mathbb{Z}} V_k$ is a graded $\mathbb{S}_n$-module, we let $\mathcal{S}_V(m,t) = \sum_k \mathcal{S}_{V_k}(m) t^k$.

The coefficients of $\mathcal{S}_V$ keep essential information about V.

Lemma 3.10. *Let $V \in R_n$ be an $\mathbb{S}_n$-representation.*

(1) *The coefficient of m^n in $\mathcal{S}_V(m)$ is $\frac{1}{n!}\dim V$.*
(2) *The value $\mathcal{S}_V(1)$ of $\mathcal{S}_V$ at $m = 1$ is the dimension of the invariant subspace $V^{\mathbb{S}_n}$.*
(3) *$(-1)^n \cdot \mathcal{S}_V(-1)$ is the multiplicity of the sign representation* sgn *in the decomposition of V into irreducible representations.*

Proof.

(1) is a direct consequence of the facts that $\ell(\sigma) = n$ if and only if σ is the identity e and $\mathrm{Tr}_V(e) = \dim V$.
(2) follows from (3.5) and the identity $\mathrm{ch}_{\mathbb{S}_n}(\mathbb{1}) = \mathsf{h}_n = \sum_{\lambda\vdash n} z_\lambda^{-1}\mathsf{p}_\lambda$ where $\mathbb{1}$ denotes the trivial representation. Indeed,

$$\mathcal{S}_V(1) = \sum_{\lambda\vdash n} z_\lambda^{-1}\langle \mathrm{ch}_{\mathbb{S}_n}(V), \mathsf{p}_\lambda\rangle = \langle \mathrm{ch}_{\mathbb{S}_n}(V), \mathrm{ch}_{\mathbb{S}_n}(\mathbb{1})\rangle = \langle V, \mathbb{1}\rangle = \dim V^{\mathbb{S}_n}.$$

(3) follows from (2) and Lemma 3.5, since $\omega\mathrm{ch}_{\mathbb{S}_n}(V) = \mathrm{ch}_{\mathbb{S}_n}(V \otimes \mathrm{sgn})$. □

3.5 | First examples

In this subsection, we work out examples of the characteristic polynomials of $\mathbb{S}_n$-modules.

Example 3.11 [21, Example I.2.1]. When $V = \mathbb{1}$ is the trivial representation, $\mathrm{ch}_{\mathbb{S}_n}(\mathbb{1}) = \mathsf{h}_n$. By (1.1), we have

$$\mathcal{S}_{\mathsf{h}_n}(m) = \left(\!\!\binom{m}{n}\!\!\right) := \binom{n+m-1}{n} = \frac{m(m+1)\cdots(m+n-1)}{n!}$$

since $\mathsf{h}_n = \sum_{1\leqslant i_1\leqslant\cdots\leqslant i_n} x_{i_1}\cdots x_{i_n}$. Similarly, when $V = \mathrm{sgn}$ is the sign representation, $\mathrm{ch}_{\mathbb{S}_n}(\mathrm{sgn}) = \mathsf{e}_n$ and

$$\mathcal{S}_{\mathsf{e}_n} = \binom{m}{n} = \frac{m(m-1)\cdots(m-n+1)}{n!}$$

since $\mathsf{e}_n = \sum_{1\leqslant i_1<\cdots<i_n} x_{i_1}\cdots x_{i_n}$.

Example 3.12 (Permutation modules). The *permutation $\mathbb{S}_n$-module* associated to a partition $\lambda = (\lambda_1, \ldots, \lambda_\ell)$ of n is the induced representation

$$M^\lambda := \mathrm{Ind}_{\mathbb{S}_\lambda}^{\mathbb{S}_n}\mathbb{1}$$

of the trivial representation of the Young subgroup $\mathbb{S}_\lambda = \mathbb{S}_{\lambda_1}\times\cdots\times\mathbb{S}_{\lambda_\ell}$. Its dimension is $n!/\lambda!$ where $\lambda! = \lambda_1!\cdots\lambda_\ell!$ and we have

$$\mathrm{ch}_{\mathbb{S}_n}(M^\lambda) = \mathsf{h}_\lambda = \mathsf{h}_{\lambda_1}\cdots\mathsf{h}_{\lambda_\ell}.$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hence, by multiplicativity,

$$\mathcal{S}_{M^\lambda}(m) = \prod_{i=1}^{\ell} \left(\!\!\binom{m}{\lambda_i}\!\!\right) = \frac{1}{\lambda!} m^{\lambda_1^t} (m+1)^{\lambda_2^t} \cdots (m+\lambda_1-1)^{\lambda_{\lambda_1}^t}, \tag{3.17}$$

where λ^t denotes the dual partition of λ with parts $\lambda_1^t \geqslant \cdots \geqslant \lambda_{\lambda_1}^t > 0$. Note that $\lambda_1^t = \ell(\lambda)$ is the length of λ and (3.17) is divisible by $m^{\ell(\lambda)}$.

From this, we deduce the following.

Lemma 3.13. *Let V be an $\mathbb{S}_n$-representation with $V = \sum_{\lambda \vdash n} b_\lambda M^\lambda$. Then,*

(1) *$b_{(n)}/n$ is equal to the coefficient of m in $\mathcal{S}_V(m)$ and*
(2) *$b_{(1^n)}$ is equal to $(-1)^n \cdot \mathcal{S}_V(-1) = \dim (V \otimes \mathrm{sgn})^{\mathbb{S}_n}$.*

Proof. By (3.17), we find that, for $\lambda = (\lambda_1, \dots, \lambda_\ell)$,

(1′) $\mathcal{S}_{M^\lambda}(m)$ is divisible by m^ℓ and the coefficient of m^ℓ is $1/(\lambda_1 \cdots \lambda_n)$;
(2′) $\mathcal{S}_{M^\lambda}(-1) \neq 0$ if and only if $\lambda = (1^n)$ and $\mathcal{S}_{M^{(1^n)}}(-1) = (-1)^n$.

The statements (1) and (2) follow immediately from (1′) and (2′). □

Example 3.14 (Permutation representations). A *permutation representation* of $\mathbb{S}_n$ is a representation of $\mathbb{S}_n$ that admits an $\mathbb{S}_n$-invariant basis. It is called *transitive* if an (or equivalently any) $\mathbb{S}_n$-invariant basis consists of a single orbit. Every permutation representation of $\mathbb{S}_n$ decomposes as a direct sum of transitive ones and every transitive permutation representation is the induced representation

$$U_H := \mathrm{Ind}_H^{\mathbb{S}_n} \mathbb{1} \tag{3.18}$$

of the trivial representation of some subgroup H of $\mathbb{S}_n$. Moreover, any $\mathbb{S}_n$-invariant basis of U_H is isomorphic to $\mathbb{S}_n/H$ as an $\mathbb{S}_n$-set. In particular,

$$\begin{aligned}
\mathrm{ch}_{\mathbb{S}_n}(U_H) &= \frac{1}{n!} \sum_{\sigma \in \mathbb{S}_n} \left|(\mathbb{S}_n/H)^\sigma\right| \cdot \mathsf{p}_\sigma \\
&= \frac{1}{n!} \sum_{\sigma \in \mathbb{S}_n} \left|\{\tau H \in \mathbb{S}_n/H : \sigma \in \tau H \tau^{-1}\}\right| \cdot \mathsf{p}_\sigma \\
&= \frac{1}{n!} \sum_{\tau H \in \mathbb{S}_n/H} \sum_{\sigma \in \tau H \tau^{-1}} \mathsf{p}_\sigma \\
&= \frac{1}{n!} \sum_{\tau H \in \mathbb{S}_n/H} \sum_{\sigma' \in H} \mathsf{p}_{\sigma'} = \frac{1}{|H|} \sum_{\sigma' \in H} \mathsf{p}_{\sigma'},
\end{aligned}$$

where the first equality holds since the σ acts on U_H as a permutation on the basis $\mathbb{S}_n/H$ and the second equality holds since $\sigma\tau H = \tau H$ if and only if $\sigma \in \tau H \tau^{-1}$. The fourth equality follows from setting $\sigma = \tau\sigma'\tau^{-1}$, since $\mathsf{p}_\sigma = \mathsf{p}_{\sigma'}$, while the third and the last equalities are immediate.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Consequently, the characteristic polynomial of U_H is

$$\mathcal{S}_{U_H}(m) = \frac{1}{|H|}\sum_{\sigma\in H} m^{\ell(\sigma)}. \tag{3.19}$$

In other words, the coefficient of m^ℓ in $\mathcal{S}_{U_H}(m)$ is the number of permutations contained in H that have precisely ℓ orbits in $[n] := \{1,\dots,n\}$, divided by $|H|$.

We note that for a positive integer m_0, the value $\mathcal{S}_{U_H}(m_0)$ is equal to the number of ways to color the elements in $[n]$ using m_0 colors, where two colorings are considered equivalent if they belong to the same H-orbit under the natural H-action on $[n]$ (see [29, Corollary 7.24.6]).

Remark 3.15. For a permutation representation V of $\mathbb{S}_n$, $\mathcal{S}_V$ always has nonnegative coefficients. On the other hand, it may have internal zeros in its coefficients. For example, for $H = \langle(1,2,3)\rangle \subset \mathbb{S}_3$, $\mathcal{S}_{U_H} = \frac{1}{3}(m^3+m)$.

Example 3.16 (Specht module). Let $\lambda = (\lambda_1,\dots,\lambda_\ell)$ be a partition of n with $\lambda_1 \geqslant \cdots \geqslant \lambda_\ell > 0$. Then, for $m \geqslant \ell$,

$$\mathcal{S}_{\mathsf{s}_\lambda}(m) = \prod_{1\leqslant i<j\leqslant m}\left(1+\frac{\lambda_i-\lambda_j}{j-i}\right).$$

In particular, $\mathcal{S}_{\mathsf{s}_\lambda}(n)$ is equal to the dimension of the irreducible GL_n-representation V_λ with the highest weight λ, by the Weyl dimension formula. See [29, (7.106) or (A.2.4)].

Example 3.17. Suppose that an $\mathbb{S}_n$-representation V is spanned by M^λ with $\ell(\lambda)\leqslant 2$. In this case, the representation V is completely determined by its characteristic polynomial $\mathcal{S}_V(m)$. To see this, write

$$V = \sum_{0\leqslant j\leqslant \frac{n}{2}} c_j M^{(n-j,j)}, \quad c_j \in \mathbb{Z}$$

so that

$$\mathcal{S}_V(m) = \sum_{1\leqslant j\leqslant \frac{n}{2}} c_j \mathcal{S}_{\mathsf{h}_{(n-j,j)}}(m), \quad \text{where } \mathcal{S}_{\mathsf{h}_{(n-j,j)}} = \left(\!\!\binom{m}{n-j}\!\!\right)\left(\!\!\binom{m}{j}\!\!\right).$$

Letting $r = \lfloor \frac{n}{2} \rfloor$, we have

$$B_{ij} := \mathcal{S}_{\mathsf{h}_{(n-j,j)}}(i-n) = \begin{cases} 0 & \text{if } i>j \\ (-1)^n\binom{n-i}{n-j}\binom{n-i}{j} & \text{if } i\leqslant j \end{cases}$$

for $0\leqslant i\leqslant r$, since $\left(\!\!\binom{i-n}{n-j}\!\!\right) = (-1)^{n-j}\binom{n-i}{n-j}$ and $\left(\!\!\binom{i-n}{j}\!\!\right) = (-1)^j\binom{n-i}{j}$. The upper triangular matrix $B = (B_{ij})_{0\leqslant i,j\leqslant r}$ has nonzero determinant $(-1)^{n(r+1)}\prod_{i=0}^r\binom{n-i}{i}\neq 0$. Since $\mathcal{S}_V(i-n) = \sum_{j=0}^r B_{ij}c_j$ for every $0\leqslant i\leqslant r$, c_j are determined by the matrix multiplication

$$(c_0,\dots,c_r)^t = B^{-1}\cdot(\mathcal{S}_V(-n),\dots,\mathcal{S}_V(r-n))^t.$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4 | CHARACTERISTIC POLYNOMIALS OF VARIETIES AND LOG-CONCAVITY

In this section, we present examples derived from geometry. Let Z be a smooth projective pure-dimensional variety with no odd degree cohomology, that is, $H^{2k+1}(Z) = 0$ for every k. Suppose that Z is equipped with an $\mathbb{S}_n$-action, or more generally, $H^{2k}(Z)$ is equipped with an $\mathbb{S}_n$-action for every k. We denote the characteristic polynomial of the induced graded $\mathbb{S}_n$-module $H^*(Z)$ by

$$\mathcal{S}_Z(m,t) = \sum_{k\geqslant 0} \mathcal{S}_{\mathrm{ch}_{\mathbb{S}_n} H^{2k}(Z)}(m) t^k.$$

We call this *the characteristic polynomial of* Z.

Note that $\mathcal{S}_Z(m,1)$ is the characteristic polynomial of the ungraded $\mathbb{S}_n$-representation $H(Z) = \bigoplus_k H^{2k}(Z) \in R_n$, while $\mathcal{S}_Z(1,t)$ is the Poincaré polynomial of the invariant part $H^*(Z)^{\mathbb{S}_n}$, which is isomorphic to $H^*(Z/\mathbb{S}_n)$ when there is an $\mathbb{S}_n$-action on Z.

4.1 | *n*-fold products

Let X be a smooth projective variety with no odd cohomology. Consider the $\mathbb{S}_n$-action on the n-fold product $X^n = X \times \cdots \times X$ of X given by permuting the product components. This can be viewed as the space of ordered n points in X acted on by an $\mathbb{S}_n$-action reordering them.

We compute $\mathcal{S}_{X^n}(m,t)$ and explore the four types of log-concavities introduced in Definition 2.1. We begin with an example where $X = \mathbb{P}^1$.

Example 4.1. Let $X = \mathbb{P}^1$. Observe that

$$\sum_{k\geqslant 0} \mathrm{ch}_{\mathbb{S}_n} H^{2k}(X^n) t^k = \mathsf{h}_n \circ (\mathsf{h}_1 + \mathsf{h}_1 t) = \sum_{k=0}^{n} \mathsf{h}_{n-k}\mathsf{h}_k t^k.$$

Applying the Stanley map $\mathcal{S}$, we obtain

$$\mathcal{S}_{(\mathbb{P}^1)^n}(m,t) = \sum_{k=0}^{n} \left(\!\!\binom{m}{n-k}\!\!\right)\left(\!\!\binom{m}{k}\!\!\right) t^k. \tag{4.1}$$

In particular, when evaluated at $t = 1$ or $m = 1$, one gets

$$\begin{aligned} \mathcal{S}_{(\mathbb{P}^1)^n}(m,1) &= \sum_{k=0}^{n} \left(\!\!\binom{m}{n-k}\!\!\right)\left(\!\!\binom{m}{k}\!\!\right) = \left(\!\!\binom{2m}{n}\!\!\right); \\ \mathcal{S}_{(\mathbb{P}^1)^n}(1,t) &= \sum_{k=0}^{n} t^k, \end{aligned} \tag{4.2}$$

respectively. By (4.1) and (4.2), we find that

- $\mathcal{S}_{(\mathbb{P}^1)^n}(m,t)$ is length log-concave in degree k for any $0 \leqslant k \leqslant n$;
- $\mathcal{S}_{(\mathbb{P}^1)^n}(m,t)$ is length log-concave at $t = 1$;

- $\mathcal{S}_{(\mathbb{P}^1)^n}(m,t)$ is degree log-concave at $m=1$,

respectively, where the first and the second statements follow by Lemma 2.4 or 2.5. However, $\mathcal{S}_{(\mathbb{P}^1)^n}(m,t)$ is not degree log-concave in a given length in general. For example, when $n=4$,

$$\mathcal{S}_{(\mathbb{P}^1)^4}(m,t) = \frac{1}{4}(t^4+1)m + \frac{1}{24}(11t^4+8t^3+6t^2+8t+11)m^2 + \frac{1}{4}(1+2t+2t^2+2t^3+t^4)m^3 + \frac{1}{24}(t^4+4t^3+6t^2+4t+1)m^4,$$

which is not degree log-concave in length 2.

On the other hand, interestingly, it seems to hold that

- $\mathcal{S}_{(\mathbb{P}^1)^n}(m,t)$ is degree log-concave in length ℓ whenever the inequalities $\lceil \log_2(n\log_2 n)\rceil \leqslant \ell \leqslant n$ are satisfied.

We have checked this for $n \leqslant 200$ by numerical computation. We also note that the bound $\lceil \log_2(n\log_2 n)\rceil$ is not sharp.

For general X, we can explicitly calculate $\mathcal{S}_{X^n}(m,t) \in \mathbb{Q}[m,t]$, and study the log-concavity. Let $p_X = \sum_{k=0}^{d} b_k t^k$ denote the Poincaré polynomial of X, where $d = \dim X$ and $b_k = \dim H^{2k}(X)$.

Theorem 4.2. *We set X^0 to be a point. Then,*

$$\sum_{n\geqslant 0}\sum_{k\geqslant 0} \mathrm{ch}_{\mathbb{S}_n} H^{2k}(X^n)t^k = \mathrm{Exp}(p_X \mathsf{h}_1). \tag{4.3}$$

Consequently, the generating series of $\mathcal{S}_{X^n}(m,t)$ is

$$\sum_{n\geqslant 0} \mathcal{S}_{X^n}(m,t)q^n = \prod_{i=0}^{d} \left(1-qt^i\right)^{-b_i m}. \tag{4.4}$$

Proof. By the Künneth formula and the property of the plethysm by h_n, we have

$$\sum_{k\geqslant 0} \mathrm{ch}_{\mathbb{S}_n} H^{2k}(X^n)t^k = \mathsf{h}_n \circ \left(\sum_{k\geqslant 0} \mathrm{ch}_{\mathbb{S}_1} H^{2k}(X)t^k\right) = \mathsf{h}_n \circ (p_X \mathsf{h}_1). \tag{4.5}$$

Hence, we obtain (4.3) by the definition of plethystic exponential.

The formula (4.4) follows from (4.3), by applying $\widehat{\mathcal{S}}_{(-)}$ since

$$\widehat{\mathcal{S}}_{\mathrm{Exp}(p_X \mathsf{h}_1)} = \exp \sum_{r\geqslant 1} \frac{1}{r} \mathsf{p}_r \circ \widehat{\mathcal{S}}(p_X \mathsf{h}_1) = \exp m \sum_{r\geqslant 1} \frac{1}{r} \mathsf{p}_r \circ (p_X q) = (\mathrm{Exp}(p_X q))^m,$$

which is equal to $\prod_{i=0}^{d}(1-qt^i)^{-b_i m}$ by Example 3.7. □

Corollary 4.3. *Let X be as above.*

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(1) *As an element of* $\mathbb{Q}[m][t]$, $\mathcal{S}_{X^n}(m,t)$ *is written as*

$$\mathcal{S}_{X^n}(m,t) = \sum_{n_0+\cdots+n_d=n} \left(\!\!\binom{b_0 m}{n_0}\!\!\right) \cdots \left(\!\!\binom{b_d m}{n_d}\!\!\right) t^{1\cdot n_1+\cdots+d\cdot n_d}. \tag{4.6}$$

(2) *When evaluated at* $t = 1$, *we have*

$$\mathcal{S}_{X^n}(m,1) = \left(\!\!\binom{Nm}{n}\!\!\right) = \frac{1}{n!} Nm(Nm+1)\cdots(Nm+n-1), \tag{4.7}$$

where $N = \sum_{k=0}^{d} b_k$.

(3) *When evaluated at* $m = 1$, *we have*

$$\mathcal{S}_{X^n}(1,t) = \sum_{n_0+\cdots+n_d=n} \left(\!\!\binom{b_0}{n_0}\!\!\right) \cdots \left(\!\!\binom{b_d}{n_d}\!\!\right) t^{1\cdot n_1+\cdots+d\cdot n_d}. \tag{4.8}$$

(4) *As an element of* $\mathbb{Q}[t][m]$, $\mathcal{S}_{X^n}(m,t)$ *is written as*

$$\mathcal{S}_{X^n}(m,t) = \sum_{\lambda\vdash n} z_\lambda^{-1} p_X^{[\lambda]} m^{\ell(\lambda)}, \tag{4.9}$$

where $p_X^{[\lambda]} = \prod_{i=1}^{\ell(\lambda)} p_X(t^{\lambda_i})$ *for a partition* $\lambda = (\lambda_1, \ldots, \lambda_{\ell(\lambda)})$ *of* n.

Proof. (1), (2), and (3) are immediate from (4.4) by reading the coefficient of q^n. For (4), observe that from (4.5),

$$\sum_{k\geqslant 0} \mathrm{ch}_{\mathbb{S}_n} H^{2k}(X^n) t^k = \mathsf{h}_n \circ (p_X \mathsf{h}_1) = \sum_{\lambda\vdash n} z_\lambda^{-1} \mathsf{p}_\lambda \circ (p_X \mathsf{h}_1) = \sum_{\lambda\vdash n} z_\lambda^{-1} p_X^{[\lambda]} \mathsf{p}_\lambda.$$

Applying $\mathcal{S}$ to this, (4.9) follows. □

By Corollary 4.3, we obtain the following straightforward log-concavity results.

Corollary 4.4.

(1) $\mathcal{S}_{X^n}(m,t)$ *is length log-concave at* $t = 1$.
(2) $\mathcal{S}_{X^n}(m,t)$ *is degree log-concave in length* n, *provided that* p_X *is log-concave with no internal zeros.*
(3) *The coefficient of* m *in* $\mathcal{S}_{X^n}(m,t)$ *has internal zeros whenever* $n > 1$.

Proof. (1) follows from (4.7) and Lemma 2.5. For (2), by (4.9), the coefficient of m^n in $\mathcal{S}_{X^n}(m,t)$ is $\frac{1}{n!}(p_X(t))^n$. This is log-concave with no internal zeroes when p_X is. For (3), by (4.9), the coefficient of m in $\mathcal{S}_{X^n}$ is $\frac{1}{n} p_X(t^n) = \frac{1}{n}\sum_{i=0}^{d} b_i t^{ni}$. □

Note that there are examples of X with non-log-concave Poincaré polynomial. For example, when $X = \mathrm{Gr}(2,4)$, $p_X(t) = 1 + t + 2t^2 + t^3 + t^4$ is not log-concave.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Example 4.5. When $X = \mathbb{P}^d$, (4.8) reads

$$\mathcal{S}_{X^n}(1,t) = \sum_{n_0+\cdots+n_d=n} t^{1\cdot n_1+\cdots+d\cdot n_d} = \sum_{\substack{\lambda\in\mathrm{Par},\\ \ell(\lambda)\leqslant n,\lambda_1\leqslant d}} t^{|\lambda|} = \begin{bmatrix} n+d \\ n \end{bmatrix}_t, \tag{4.10}$$

where λ runs over all the partitions with length at most n and all parts at most d and for $a \geqslant b \geqslant 0$,

$$\begin{bmatrix} a \\ b \end{bmatrix}_t := \frac{[a]_t!}{[a-b]_t![b]_t!} \in \mathbb{Z}[t]$$

with $[a]_t! = \prod_{i=1}^{a}[i]_t$ and $[a]_t = \sum_{i=0}^{a-1} t^i$ for all $a \geqslant 0$. The second equality in (4.10) holds by identifying the $(d+1)$-tuple $(n_0, \dots, n_d)$ with the partition $\lambda = 1^{n_1}\cdots d^{n_d}$ and the last equality holds by [27, Proposition 1.7.3].

Unlike the case of $X = \mathbb{P}^1$, this is not log-concave in general. For example, when $n = d = 2$, one gets $\mathcal{S}_{(\mathbb{P}^2)^2}(1,t) = \begin{bmatrix} 4 \\ 2 \end{bmatrix}_t = 1 + t + 2t^2 + t^3 + t^4$. In other words, $\mathcal{S}_{X^n}(m,t)$ is not degree log-concave at $m = 1$ in general.

Recall that when $X = \mathbb{P}^1$, we verified in Example 4.1 that $\mathcal{S}_{X^n}(m,t)$ is degree log-concave in length $\lceil \log_2(n\log_2 n)\rceil \leqslant \ell \leqslant n$ for $n \leqslant 200$, while it fails to be degree log-concave in smaller lengths. Our goal is to investigate the degree log-concavity of $\mathcal{S}_{X^n}(m,t)$ in sufficiently large length. Additionally, we aim to examine the length log-concavity of $\mathcal{S}_{X^n}(m,t)$ at a given degree.

Question 4.6. Let $\mathcal{S}^{\ell}_{X^n,k} \in \mathbb{Q}$ denote the coefficient of $m^\ell t^k$ in $\mathcal{S}_{X^n}(m,t)$ for each ℓ, k.

(1) Is it true that

$$\left(\mathcal{S}^{\ell}_{X^n,k}\right)^2 \geqslant \mathcal{S}^{\ell+1}_{X^n,k}\cdot\mathcal{S}^{\ell-1}_{X^n,k} \quad \text{for any } \ell, k?$$

(2) Assume that p_X is log-concave with no internal zeros. Is there a lower bound for ℓ as a function of n such that

$$\left(\mathcal{S}^{\ell}_{X^n,k}\right)^2 \geqslant \mathcal{S}^{\ell}_{X^n,k+1}\cdot\mathcal{S}^{\ell}_{X^n,k-1}$$

holds for all k?

Here, we give a partial asymptotic answer to these questions. Indeed, both inequalities hold when k is sufficiently small and ℓ is sufficiently large with respect to n.

Proposition 4.7. *Let $j, k \geqslant 0$. The coefficient of $m^{n-j}t^k$ in $\mathcal{S}_{X^n}(m,t)$ is*

$$\mathcal{S}^{n-j}_{X^n,k} = \frac{b_0^{n-k-j}b_1^k}{2^j j!k!}\cdot\frac{n^{2j+k}}{n!} + o\left(\frac{n^{2j+k}}{n!}\right).$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In particular, these coefficients satisfy the log-concavities

$$\left(\mathcal{S}^{n-j}_{X^n,k}\right)^2 \geqslant \mathcal{S}^{n-j}_{X^n,k-1}\cdot\mathcal{S}^{n-j}_{X^n,k+1} \quad \textit{and} \quad \left(\mathcal{S}^{n-j}_{X^n,k}\right)^2 \geqslant \mathcal{S}^{n-j-1}_{X^n,k}\cdot\mathcal{S}^{n-j+1}_{X^n,k}$$

for sufficiently large n.

Proof. Recall that $\left(\!\!\binom{m}{n}\!\!\right) = \sum_{j=0}^{n-1}\frac{c(n,n-j)}{n!}m^{n-j}$, where $c(n,j)$ is the signless Stirling number of the first kind. It follows from (4.6) that

$$\mathcal{S}^{n-j}_{X^n,k} = \sum_{\substack{j_0+\cdots+j_d=j,\\ n_0+\cdots+n_d=n,\\ 1\cdot n_1+\cdots+d\cdot n_d=k}} \prod_{i=0}^{d}\frac{c(n_i,n_i-j_i)}{n_i!}\cdot b_i^{n_i-j_i}. \tag{4.11}$$

Since $j_0,\dots,j_d$ and $n_1,\dots,n_d$ are bounded by j and k, respectively, and it is known that $c(n,n-j)=\frac{1}{2^j j!}n^{2j}+o(n^{2j})$, for any given $j\geqslant 0$, one can easily check that the summands in (4.11) are dominated by those with n_0 minimal and j_0 maximal, namely, those with

$$(n_0,\dots,n_d)=(n-k,k,0,\dots,0) \quad \text{and} \quad (j_0,\dots,j_d)=(j,0,\dots,0).$$

Indeed, $\frac{c(n_0,n_0-j_0)}{n_0!}=\frac{1}{2^{j_0}j_0!}\cdot\frac{n_0^{2j_0}}{n_0!}+o(\frac{n_0^{2j_0}}{n_0!})$ is maximal if $2j_0-n_0$ is maximal. In particular, the coefficient (4.11) is equal to

$$\frac{c(n-k,n-k-j)b_0^{n-k-j}}{(n-k)!}\cdot\frac{c(k,k)b_1^k}{k!}=\frac{b_0^{n-k-j}b_1^k}{2^j j!k!}\cdot\frac{(n-k)^{2j}}{(n-k)!}$$

modulo $o(\frac{n^{2j+k}}{n!})$. The first formula in the proposition now follows from $\frac{(n-k)^{2j}}{(n-k)!}=\frac{n^{2j+k}}{n!}+o(\frac{n^{2j+k}}{n!})$. The inequalities are immediate from the asymptotic formula. □

Remark 4.8 (Zeta function). When X is defined over $\mathbb{Z}$, and has cohomology of pure Tate type, so that $|X(\mathbb{F}_q)|=p_X(q)$ for any prime power q (e.g., this holds if X admits a cell decomposition), the generating series (4.4) is equal to the mth power of the Zeta function

$$\zeta_X(t)=\exp\sum_{n\geqslant 1}\frac{1}{n}|X(\mathbb{F}_{q^n})|t^n=\exp\sum_{n\geqslant 1}\frac{1}{n}\mathrm{p}_n\circ(p_X(q)t)=\mathrm{Exp}(p_X(q)t)$$

of X up to interchanging the roles of q and t. Here, q is regarded as a variable when applying the plethysm. Moreover, the equality (4.4) corresponds to the identity appearing in the Weil conjecture

$$\zeta_X(t)=\frac{P_1(t)\cdots P_{2d-1}(t)}{P_0(t)\cdots P_{2d}(t)}$$

with polynomials $P_{2i+1}(t)=1$ and $P_{2i}(t)=(1-q^i t)^{b_i}$ for all i.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.2 | GIT moduli space of n points on $\mathbb{P}^1$

Let $n = 2r + 1$ be odd with $r \geqslant 1$. We let

$$Y_n = (\mathbb{P}^1)^n /\!\!/ \mathrm{SL}_2(\mathbb{C})$$

be the GIT quotient (cf. [22]) of $(\mathbb{P}^1)^n$ by the diagonal $\mathrm{SL}_2(\mathbb{C})$ action with respect to the linearization $\mathcal{O}_{(\mathbb{P}^1)^n}(1, \dots, 1)$. It is a smooth projective variety whose cohomology was computed in [20, Example 5.18].

The natural action of $\mathbb{S}_n$ on $(\mathbb{P}^1)^n$ permuting the factors commutes with that of $\mathrm{SL}_2(\mathbb{C})$, and hence, we have the induced action of $\mathbb{S}_n$ on Y_n. It is straightforward to apply Kirwan's method to compute the graded $\mathbb{S}_n$-module $H^*(Y_n)$ to obtain (cf. [6, Lemma 4.7])

$$\sum_{k=0}^{2r-2} \mathrm{ch}_{\mathbb{S}_n} H^{2k}(Y_n) t^k = \sum_{j=0}^{r-1} t^j (1 + t^2 + \cdots + t^{2(r-j-1)}) \mathsf{h}_{(n-j,j)}. \tag{4.12}$$

Applying the Stanley map $\mathcal{S}$, we obtain the characteristic polynomial $\mathcal{S}_{Y_n}(m, t)$ of Y_n as follows.

Proposition 4.9. *For $n = 2r + 1$,*

$$\mathcal{S}_{Y_n}(m, t) = \sum_{j=0}^{r-1} t^j (1 + t^2 + \cdots + t^{2(r-j-1)}) \left(\!\!\binom{m}{n-j}\!\!\right) \left(\!\!\binom{m}{j}\!\!\right).$$

In particular,

$$\mathcal{S}_{Y_n}(m, 1) = \sum_{j=0}^{r-1} (r - j) \left(\!\!\binom{m}{n-j}\!\!\right) \left(\!\!\binom{m}{j}\!\!\right), \text{ and}$$

$$\mathcal{S}_{Y_n}(1, t) = \sum_{j=0}^{r-1} t^j (1 + t^2 + \cdots + t^{2(r-j-1)}).$$

Proof. This is immediate from (4.12), since $\mathcal{S}_{\mathsf{h}_{(a,b)}} = \left(\!\!\binom{m}{a}\!\!\right) \left(\!\!\binom{m}{b}\!\!\right)$. □

Numerical calculations lead us to the following conjecture.

Conjecture 4.10. *Let $n \geqslant 3$ be an odd integer.*

(1) *$\mathcal{S}_{Y_n}(m, t)$ is length log-concave in degree k for any $0 \leqslant k \leqslant n$.*
(2) *$\mathcal{S}_{Y_n}(m, t)$ is length log-concave at $t = 1$, that is, $\mathcal{S}_{Y_n}(m, 1) \in \mathbb{Q}[m]$ is log-concave.*

We have checked that (1) and (2) hold for $n \leqslant 400$.

Example 4.11. When $n = 5$, the graded $\mathbb{S}_n$-representation on $H^*(Y_5)$ is $\sum_{k=0}^{2} \mathrm{ch}_{\mathbb{S}_5} H^{2k}(Y_5) t^k = \mathsf{h}_5(1 + t^2) + \mathsf{h}_{(4,1)} t$. Hence, by applying $\mathcal{S}$, we obtain

$$\mathcal{S}_{Y_5}(m, t) = \frac{1}{120}(m^5 + 10m^4 + 35m^3 + 50m^2 + 24m)(1 + t^2) + \frac{1}{24}(m^5 + 6m^4 + 11m^3 + 6m^2)t.$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Evaluating it at $t = 1$, we get

$$\mathcal{S}_{Y_5}(m, 1) = \frac{1}{120}(7m^5 + 50m^4 + 125m^3 + 130m^2 + 48).$$

It is easy to see that Conjecture 4.10 holds in this case.

On the other hand, the degree log-concavities fail for Y_n. More precisely, $\mathcal{S}_{Y_n}(m, t)$ is degree log-concave neither at $m = 1$ nor in length 1.

Example 4.12. $\mathcal{S}_{Y_n}(1, t)$ is not log-concave whenever $n \geqslant 7$. Indeed,

$$\mathcal{S}_{Y_7}(1, t) = 1 + t + 2t^2 + t^3 + t^4,$$

$$\mathcal{S}_{Y_9}(1, t) = 1 + t + 2t^2 + 2t^3 + 2t^4 + t^5 + t^6,$$

$$\mathcal{S}_{Y_{11}}(1, t) = 1 + t + 2t^2 + 2t^3 + 3t^4 + 2t^5 + 2t^6 + t^7 + t^8,$$

$$\cdots$$

One can also easily check that the coefficient of m in $\mathcal{S}_{Y_n}(m, t)$ is

$$\frac{1}{n}\sum_{j=0}^{r-1} t^{2j} = \frac{1}{n}(1 + t^2 + \cdots + t^{2r-4} + t^{2r-2}),$$

which is not log-concave due to the internal zeros at odd degrees.

Interestingly, numerical calculation shows that $\mathcal{S}_{Y_n}(m, t)$ is degree log-concave in length ℓ when ℓ is *sufficiently close* to n. For example, $\mathcal{S}_{Y_n}(m, t)$ is degree log-concave in length ℓ for any ℓ in the range $\sqrt{n} + 2 \leqslant \ell \leqslant n$ for $5 \leqslant n \leqslant 400$, although the bound $\sqrt{n} + 2$ is definitely not sharp. Computing a sharp bound and analyzing its asymptotic behavior as n grows would also be of interest.

4.3 | Hessenberg varieties

Regular semisimple Hessenberg varieties are subvarieties in the flag variety, naturally equipped with an $\mathbb{S}_n$-action on its cohomology. For further details on the discussion below, we refer the reader to a survey [2].

Denote the variety of flags in $\mathbb{C}^n$ by $\mathrm{Fl}_n = \{F_\bullet = (F_1 \subset \cdots \subset F_n = \mathbb{C}^n)\}$ where F_i are i-dimensional subspaces of $\mathbb{C}^n$ for each $i \in [n] := \{1, \ldots, n\}$.

A *Hessenberg function* is a nondecreasing function $h : [n] \to [n]$ with $h(i) \geqslant i$ for all i. Choosing a regular semisimple $n \times n$ matrix s (or equivalently, an $n \times n$ matrix with n distinct eigenvalues), one obtains a subvariety in Fl_n, called a *(regular semisimple) Hessenberg variety*, defined as

$$X_h = \{F_\bullet \in \mathrm{Fl}_n : sF_i \subset F_{h(i)} \text{ for all } i\}.$$

14602244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Here, we suppress s in the notation, as its diffeomorphism type (hence its cohomology) does not depend on the choice of s.

These varieties were introduced in [12], where it is also proved that they are smooth of dimension $\sum_{i=1}^{n}(h(i)-i)$ when s is regular semisimple. In this case, X_h admits a natural action of the maximal torus T in $\mathrm{GL}_n(\mathbb{C})$ centralizing s. Since its T-fixed locus is finite (and isomorphic to $\mathbb{S}_n$), X_h admits a cell decomposition by the Białynicki–Birula theorem and $H^{2k+1}(X_h)=0$ for all k. Later in [30], Tymoczko defined an $\mathbb{S}_n$-action on the T-equivariant cohomology $H_T^*(X_h)$ using the GKM theory [14] and $X_h^T \cong \mathbb{S}_n$, and showed that it descends to an $\mathbb{S}_n$-action on $H^*(X_h)$ via the canonical projection $H_T^*(X_h) \to H^*(X_h)$. We note that this $\mathbb{S}_n$-action on $H^*(X_h)$ may not lift to an $\mathbb{S}_n$-action on X_h [8].

We thus obtain

$$\sum_{k\geqslant 0} \mathrm{ch}_{\mathbb{S}_n}(H^{2k}(X_h))t^k \in \Lambda_n[t] \tag{4.13}$$

whose characteristic polynomial is

$$\mathcal{S}_{X_h}(m,t) = \sum_{k\geqslant 0} \mathcal{S}_{H^{2k}(X_h)}(m)t^k \in \mathbb{Q}[m,t]. \tag{4.14}$$

On the other hand, there is a purely combinatorial way to associate a symmetric function to each Hessenberg function h. The chromatic symmetric function introduced by Stanley in [26] associates a symmetric function

$$\mathsf{csf}_\Gamma = \sum_{\kappa:[n]\to\mathbb{Z}_{>0} \text{ proper}} x_{\kappa(1)}\cdots x_{\kappa(n)} \in \Lambda_n \tag{4.15}$$

to each graph Γ with n vertices, where κ runs over all the coloring maps from the vertex set $[n]$ to the set of colors indexed by positive integers that are *proper*, that is, $\kappa(i)\neq\kappa(j)$ whenever $(i,j)\in E(\Gamma)$. By the substitution (1.1), its characteristic polynomial

$$\mathcal{S}_{\mathsf{csf}_\Gamma}(m) = \mathcal{X}_\Gamma(m)$$

is the usual chromatic polynomial.

In [23], Shareshian and Wachs introduced the notion of the chromatic *quasi*symmetric function cqsf_Γ of a graph Γ, which is a polynomial in a variable t whose coefficients are quasi-symmetric functions, that is, functions invariant under order-preserving change of variable indices. The chromatic quasi-symmetric function reduces to the original chromatic symmetric function when evaluated at $t=1$.

For a Hessenberg function $h:[n]\to[n]$, let Γ_h be the *incomparability graph associated to* h, that is, the graph whose vertex set and edge set are

$$V(\Gamma_h)=[n] \quad \text{and} \quad E(\Gamma_h)=\{(i,j)\in[n]^2 : i<h(i)\leqslant j\}$$

respectively. Shareshian and Wachs in [23] showed that the chromatic quasi-symmetric function of the graph Γ_h is, in fact, a graded symmetric function

$$\mathsf{cqsf}_{\Gamma_h}(t)\in\Lambda_n[t] \tag{4.16}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

so that we have the characteristic polynomial

$$\mathcal{S}_{\Gamma_h}(m,t) := \mathcal{S}_{\mathsf{cqsf}_{\Gamma_h}}(m,t) \in \mathbb{Q}[m,t] \tag{4.17}$$

that satisfies

$$\mathcal{S}_{\Gamma_h}(m,1) = \mathcal{S}_{\mathsf{csf}_{\Gamma_h}}(m) = \mathcal{X}_{\Gamma_h}(m). \tag{4.18}$$

A remarkable conjecture by Shareshian–Wachs in [23], proved by Brosnan–Chow and independently by Guay–Paquet tells us that (4.16) is the ω-dual of (4.13) where $\omega : \Lambda \to \Lambda$ is the involution sending e_λ to h_λ.

Theorem 4.13 [7, 15]. *For any Hessenberg function h,*

$$\omega \sum_{k\geqslant 0} \mathrm{ch}_{\mathbb{S}_n}(H^{2k}(X_h))t^k = \mathsf{cqsf}_{\Gamma_h}.$$

See [19] for an elementary proof based on the modular law.

By applying the Stanley map $\mathcal{S}$, we get the following.

Corollary 4.14. *Let*

$$\mathcal{S}^{\omega}_{Z}(m,t) = \mathcal{S}\left(\omega \sum_{k\geqslant 0} \mathrm{ch}_{\mathbb{S}_n}(H^{2k}(Z))t^k\right) = \sum_{k\geqslant 0} \mathcal{S}_{H^{2k}(Z)\otimes \mathsf{sgn}_n} t^k$$

for any connected smooth projective variety Z with an $\mathbb{S}_n$-action on each even degree cohomology group of Z. Then,

$$\mathcal{S}^{\omega}_{X_h}(m,t) = \mathcal{S}_{\Gamma_h}(m,t). \tag{4.19}$$

By (4.18) and (4.19), we have

$$\mathcal{S}^{\omega}_{X_h}(m,1) = \mathcal{X}_{\Gamma_h}(m) = \prod_{i=1}^{n}(m-(h(i)-i)), \tag{4.20}$$

where the second equality is an easy calculation based on the fact that Γ_h is directed. In particular, this is log-concave as it should be by a general result of Huh [17].

By Lemma 3.5, we have

$$\mathcal{S}^{\omega}_{F}(m) = (-1)^n \mathcal{S}_F(-m) \tag{4.21}$$

for any $F \in \Lambda_n$ and the following is immediate.

Proposition 4.15. *The characteristic polynomial of a Hessenberg variety X_h satisfies*

$$\mathcal{S}_{X_h}(m,1) = \prod_{i=1}^{n}(m+h(i)-i) \tag{4.22}$$

and $\mathcal{S}_{X_h}(m,t)$ is length log-concave at $t=1$.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

There are combinatorial formulas for the coefficients of p_λ in $\mathsf{cqsf}_{\Gamma_h}(t)$ in [5, Theorem 4 and Corollary 7]. These immediately induce combinatorial formulas for $\mathcal{S}_{X_h}(m,t)$ by (4.19) and (4.21). We do not know yet how to deduce or disprove the other log-concavities from them, except the following two results. Write $[r]_t := \sum_{i=0}^{r-1} t^i$ for $r \geqslant 0$.

(1) The evaluation of $\mathcal{S}_{X_h}(m,t)$ at $m=1$ is

$$\mathcal{S}_{X_h}(1,t) = \sum_{k\geqslant 0} \dim(H^{2k}(X_h)^{\mathbb{S}_n})t^k = \prod_{i=1}^{n}[h(i)-i+1]_t.$$

In particular, $\mathcal{S}_{X_h}(m,t)$ is degree log-concave at $m=1$.

(2) The coefficient of m in $\mathcal{S}_{X_h}(m,t)$ is

$$\frac{1}{n}[n]_t \prod_{i=1}^{n-1}[h(i)-i]_t.$$

In particular, $\mathcal{S}_{X_h}(m,t)$ is degree log-concave at length 1.

Both formulas follow by combining the results of [1, 23, 24], together with Lemmas 3.10 and 3.13. See also [19, Proposition 5.10].

We end this section with the following natural question that may be regarded as a graded version of Huh's theorem in [17].

Question 4.16. Is $\mathcal{S}_{X_h}(m,t)$ or $\mathcal{S}^{\omega}_{X_h}(m,t) = \mathcal{S}_{\Gamma_h}(m,t) = \mathcal{S}_{\mathsf{cqsf}_{\Gamma_h}}(m,t)$ length log-concave in degree k for any k?

5 | CHARACTERISTIC POLYNOMIAL OF $\overline{\mathcal{M}}_{0,n}$

The moduli space $\overline{\mathcal{M}}_{0,n}$ of n-pointed stable curves of genus 0 is one of the most important varieties in algebraic geometry and has been much studied (cf. [10, §1]). It comes with a natural action of the symmetric group $\mathbb{S}_n$ by permuting the marked points, and thus, its cohomology $H^*(\overline{\mathcal{M}}_{0,n})$ is a graded $\mathbb{S}_n$-module. The rest of this paper is devoted to the study of the characteristic polynomial $\mathcal{S}_{\overline{\mathcal{M}}_{0,n}}$ of $\overline{\mathcal{M}}_{0,n}$ as well as that of $\overline{\mathcal{M}}_{0,n+1}$ on which $\mathbb{S}_n$ is acting by permuting the first n marked points with the last one fixed. Based on our pervious results in [10, 11], we provide recursive formulas for the characteristic polynomials (Theorem 5.11). In the last subsection §5.3, we find an interpretation of the characteristic polynomials of permutation representations that constitute the $\mathbb{S}_n$-representation $H^*(\overline{\mathcal{M}}_{0,n+1})$, in terms of colorings (Corollary 5.21).

5.1 | Recursive formulas for the representations

This section reviews the results in [10, 11], with a focus on formulas. For an overview together with related geometry, we refer the reader to [11, §3].

The moduli space $\overline{\mathcal{M}}_{0,n}$ is a compactification of the variety

$$(\mathbb{P}^1)^n - \Delta/\operatorname{Aut}(\mathbb{P}^1)$$

of n distinct ordered points on $\mathbb{P}^1$ up to the automorphism group action. It certainly comes with the natural action of $\mathbb{S}_n$ permuting the marked points. When studying the cohomology of $\overline{\mathcal{M}}_{0,n}$ as a graded $\mathbb{S}_n$-module, it helps to study the cohomology of $\overline{\mathcal{M}}_{0,n+1}$ as well, on which $\mathbb{S}_n$ is acting by permuting the first n marked points with the last one fixed.

We write the $\mathbb{S}_n$-equivariant Poincaré polynomial of $\overline{\mathcal{M}}_{0,n}$ and $\overline{\mathcal{M}}_{0,n+1}$ as

$$P_n = \sum_{k=0}^{n-3} P_{n,k}t^k \in \Lambda_n[t], \quad \text{where } P_{n,k} = \operatorname{ch}_{\mathbb{S}_n} H^{2k}(\overline{\mathcal{M}}_{0,n}) \in \Lambda_n;$$

$$Q_n = \sum_{k=0}^{n-2} Q_{n,k}t^k \in \Lambda_n[t], \quad \text{where } Q_{n,k} = \operatorname{ch}_{\mathbb{S}_n} H^{2k}(\overline{\mathcal{M}}_{0,n+1}) \in \Lambda_n.$$

Moreover, we consider their generating series as

$$P = 1 + \mathsf{h}_1 + \mathsf{h}_2 + \sum_{n\geqslant 3} P_n \quad \in \Lambda[\![t]\!] = \prod_n \Lambda_n[\![t]\!];$$

$$Q = 1 + \mathsf{h}_1 + \sum_{n\geqslant 2} Q_n \quad \in \Lambda[\![t]\!],$$

where we treat $\overline{\mathcal{M}}_{0,n}$ as a point when $0 \leqslant n \leqslant 2$.

In [10, 11], we established formulas for P and Q that were obtained in two stages: (1) comparison of P and Q and (2) computation of Q.

Stage 1. The relation between P_n and Q_n is obtained in [10, §4] by the theory of δ-stable quasi-maps, developed in [9], in a very special case that provides an $\mathbb{S}_n$-equivariant factorization of the canonical forgetful morphism $\overline{\mathcal{M}}_{0,n+1} \to \overline{\mathcal{M}}_{0,n}$ by blowups. By the blowup formula and geometric invariant theory, we have the following.

Theorem 5.1 [10, Theorem 4.8]. *For $n \geqslant 3$,*

$$(1+t)P_n = Q_n - t\left(\sum_{2\leqslant h<\frac{n}{2}} Q_h Q_{n-h} + \mathsf{s}_{(1,1)}\circ Q_{\frac{n}{2}}\right),$$

where we set $Q_{\frac{n}{2}} = 0$ for n odd. Equivalently, the generating series P and Q satisfy (cf. [11, Corollary 3.3])

$$(1+t)P = (1 + t + \mathsf{h}_1 t)Q - t\mathsf{s}_{(1,1)}\circ Q.$$

Stage 2. The combinatorial formula for Q_n in [10] is derived from the well-known $\mathbb{S}_n$-equivariant factorization of the Kapranov map

$$\overline{\mathcal{M}}_{0,n+1} \longrightarrow \mathbb{P}^{n-2}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

by a sequence of blowups. This corresponds to a realization of $\overline{\mathcal{M}}_{0,n+1}$ as the wonderful compactification of the complement of the braid hyperplane arrangement [18]. The challenging task of tracking all the blowup centers is accomplished by the combinatorics of weighted rooted trees.

Definition 5.2 [11, Definition 3.5]. A *labeled rooted tree* is a tree with one distinguished leg called, the *output*, and additional n legs, called the *inputs*, labeled by $1, \dots, n$, for some $n \in \mathbb{Z}_{\geqslant 0}$. The vertex to which the output is attached is called the *root*.

More concretely, a labeled rooted tree is a triple of a tree T, the output map $\mathrm{pt} \to V(T)$ and the input map $\{1, \dots, n\} \to V(T)$, where $V(T)$ is the set of vertices of T. In particular, permuting the labels of the inputs attached to the same vertex does not distinguish the labeled rooted tree.

A *rooted tree* is a tree obtained from a labeled rooted tree by forgetting the labels on the inputs.

Definition 5.3. Let T be a (resp. labeled) rooted tree. A *weight function* on T is a function $w : V(T) \to \mathbb{Z}_{\geqslant 0}$ on the vertex set $V(T)$. We denote the total weight by $|w| = \sum_{v \in V(T)} w(v)$.

We call the pair (T, w) a (resp. *labeled*) *weighted rooted tree*.

Definition 5.4. For n and k in $\mathbb{Z}_{\geqslant 0}$, we define $\mathcal{T}_{n,k}$ (resp. $\mathcal{T}^{\mathrm{lab}}_{n,k}$) to be the set of (resp. labeled) weighted rooted trees (T, w) with n inputs and total weight $|w| = k$, which satisfy the following *valency conditions*:

(1) $0 \leqslant w(v) \leqslant \mathrm{val}(v) - 3$ if v is the root;
(2) $0 < w(v) \leqslant \mathrm{val}(v) - 3$ if v is not the root.

There is a natural $\mathbb{S}_n$-action on $\mathcal{T}^{\mathrm{lab}}_{n,k}$ reordering labels and $\mathcal{T}_{n,k} = \mathcal{T}^{\mathrm{lab}}_{n,k}/\mathbb{S}_n$.

Proposition 5.5 [10, Proposition 5.12]. *The $\mathbb{S}_n$-representation $H^{2k}(\overline{\mathcal{M}}_{0,n+1})$ is a permutation representation with an $\mathbb{S}_n$-basis isomorphic to $\mathcal{T}^{\mathrm{lab}}_{n,k}$.*

We decompose this permutation representation into transitive permutation representations U_H in Example 3.14 as follows.

Definition 5.6. For $(T, w) \in \mathcal{T}^{\mathrm{lab}}_{n,k}$, we denote by $\mathrm{Stab}(T, w)$ the stabilizer subgroup of (T, w) in $\mathbb{S}_n$, which is the subgroup consisting of permutations of labels that preserves the isomorphism class of (T, w) as a labeled weighted rooted tree. We define

$$U_{(T,w)} = U_{\mathrm{Stab}(T,w)} \tag{5.1}$$

using the notation of (3.18). For $(T, w) \in \mathcal{T}_{n,k} = \mathcal{T}^{\mathrm{lab}}_{n,k}/\mathbb{S}_n$, we define $\mathrm{Stab}(T, w)$ and $U_{(T,w)}$ similarly by taking any lift in $\mathcal{T}^{\mathrm{lab}}_{n,k}$.

Note that the subgroup $\mathrm{Stab}(T, w) \subset \mathbb{S}_n$ is well defined up to conjugation and $U_{(T,w)}$ is well defined as an $\mathbb{S}_n$-representation.

For convenience, we will sometimes omit w in our notation, and write T, $\mathrm{Stab}(T)$ and U_T for (T, w), $\mathrm{Stab}(T, w)$ and $U_{(T,w)}$, respectively.

Example 5.7. Suppose that T has only one vertex, namely, the root, and all the n inputs are attached to the root. Then, $\mathrm{Stab}(T) \cong \mathbb{S}_n$.

Example 5.8. Suppose that T has only two vertices with a inputs attached to the root and the remaining $n-a$ inputs attached to the other nonroot vertex. Then, $\mathrm{Stab}(T) \cong \mathbb{S}_a \times \mathbb{S}_{n-a}$.

Example 5.9. Suppose that T has three vertices where both nonroot vertices are adjacent to the root and have weight one. Suppose further that to each nonroot vertex, a inputs are attached and the remaining $n-2a$ vertices are attached to the root. Then, $\mathrm{Stab}(T) \cong \mathbb{S}_{n-2a} \times ((\mathbb{S}_a \times \mathbb{S}_a) \rtimes \mathbb{S}_2)$.

Proposition 5.5 now reads as

$$Q_{n,k} = \sum_{T \in \mathcal{T}_{n,k}} \mathrm{ch}_{\mathbb{S}_n} U_T.$$

Combining Theorem 5.1 and Proposition 5.5, we obtain formulas computing P and Q in terms of weighted rooted trees.

For actual computations, however, the combinatorial complexity of weighted rooted trees grows quickly as n increases. As a remedy, in [11], a natural recursive structure on $\mathcal{T}_{n,k}$ is observed, which leads to recursive formulas for computing Q and P.

Let $\mathcal{T}^+_{n,k} \subset \mathcal{T}_{n,k}$ be the subset consisting of weighted rooted trees that have positive weights at the root. Let $Q^+_1 = h_1$ and for $n \geqslant 2$,

$$Q^+_n = \sum_{k=0}^{n-2} Q^+_{n,k} t^k, \quad \text{where } Q^+_{n,k} = \sum_{T \in \mathcal{T}^+_{n,k}} \mathrm{ch}_{\mathbb{S}_n} U_T.$$

Consider its generating series

$$Q^+ = Q^+_1 + \sum_{n \geqslant 2} Q^+_n \quad \in \Lambda[\![t]\!].$$

Theorem 5.10 [11, Theorem 4.7]. *Q is the plethystic exponential of Q^+:*

$$Q = \mathrm{Exp}(Q^+) = \sum_{r \geqslant 0} \mathsf{h}_r \circ Q^+.$$

Consequently, Q^+ is determined by the formulas (cf. [11, Corollary 4.9])

$$\text{(Recursive)} \quad Q^+ = \mathsf{h}_1 + \sum_{r \geqslant 3} \left(\sum_{i=1}^{r-2} t^i \right) (\mathsf{h}_r \circ Q^+),$$

$$\text{(Exponential)} \quad \mathrm{Exp}(tQ^+) = t^2 \mathrm{Exp}(Q^+) + (1-t)(1 + t + \mathsf{h}_1 t).$$

More explicitly, Q_n and Q^+_n are determined by

$$Q^+_n = \sum_{\lambda \vdash n} \left(\sum_{i=1}^{\ell(\lambda)-2} t^i \right) \prod_{j=1}^{m} \left(\mathsf{h}_{r_j} \circ Q^+_{n_j} \right) \quad \textit{and} \quad Q_n = \sum_{\lambda \vdash n} \prod_{j=1}^{m} \left(\mathsf{h}_{r_j} \circ Q^+_{n_j} \right)$$

for $n \geqslant 2$, where n_j denote the parts of λ with multiplicities r_j so that $\lambda = (n_1^{r_1}, \dots, n_m^{r_m})$ with $n_1 > \cdots > n_m > 0$ and $\ell(\lambda) = \sum_{j=1}^m r_j$.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In the next subsection, we will apply the Stanley map $\widehat{\mathcal{S}}$ to the formulas in Theorems 5.1 and 5.10 to deduce corresponding formulas for the characteristic polynomials of $\overline{\mathcal{M}}_{0,n}$ and $\overline{\mathcal{M}}_{0,n+1}$. Since every formula appearing above involves only arithmetic operations and plethysm, our formulas for the characteristic polynomials involve only arithmetic operations and composition.

The proof of Theorem 5.10 uses the recursive structure on $\mathcal{T}_{n,k}$ that will be briefly explained in Lemma 5.16 in §5.3 below.

5.2 | Recursive formulas for characteristic polynomials

Let us apply the Stanley map to $P, P_n, P_{n,k}, Q, Q_n, Q_{n,k}$ to obtain the characteristic polynomials of $\overline{\mathcal{M}}_{0,n}$ and $\overline{\mathcal{M}}_{0,n+1}$ as

$$(\mathscr{P}, \mathscr{P}_n, \mathscr{P}_{n,k}) := (\widehat{\mathcal{S}}_P, \mathcal{S}_{P_n}, \mathcal{S}_{P_{n,k}});$$

$$(\mathscr{Q}, \mathscr{Q}_n, \mathscr{Q}_{n,k}) := (\widehat{\mathcal{S}}_Q, \mathcal{S}_{Q_n}, \mathcal{S}_{Q_{n,k}});$$

$$(\mathscr{Q}^+, \mathscr{Q}^+_n, \mathscr{Q}^+_{n,k}) := (\widehat{\mathcal{S}}_{Q^+}, \mathcal{S}_{Q^+_n}, \mathcal{S}_{Q^+_{n,k}}).$$

In particular, the generating series are written as

$$\mathscr{P}(m,q,t) := 1 + mq + \frac{m(m+1)}{2}q^2 + \sum_{n\geqslant 2} \mathscr{P}_n(m,t)q^n;$$

$$\mathscr{Q}(m,q,t) := 1 + mq + \sum_{n\geqslant 2} \mathscr{Q}_n(m,t)q^n;$$

$$\mathscr{Q}^+(m,q,t) := mq + \sum_{n\geqslant 2} \mathscr{Q}^+_n(m,t)q^n.$$

The following are direct consequences of Theorems 5.1 and 5.10, thanks to the properties of $\mathcal{S}$ and $\widehat{\mathcal{S}}$ in (3.8), (3.10), and (3.13).

Let Par be the set of all the partitions of nonnegative integers as in Remark 3.8. We write $|\lambda| = \sum_{i=1}^{\ell} \lambda_i$ for $\lambda = (\lambda_1, \dots, \lambda_\ell)$. Then, $\mathscr{P}$ and $\mathscr{Q}$ are determined by the following.

Theorem 5.11.

(1) *$\mathscr{P}$ and $\mathscr{Q}$ are related by the formula*

$$(1+t)\mathscr{P} = (1 + t + mqt)\mathscr{Q} - \frac{t}{2}(\mathscr{Q}^2 - \mathscr{Q}^{[2]}).$$

(2) *$\mathscr{Q}$ is the plethystic exponential of $\mathscr{Q}^+$:*

$$\mathscr{Q} = \mathrm{Exp}(\mathscr{Q}^+) = \sum_{\lambda\in \mathrm{Par}} z_\lambda^{-1} (\mathscr{Q}^+)^{[\lambda]}.$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(3) *Moreover, $\mathcal{Q}^+$ satisfies the following two equivalent formulas.*

$$\text{(Recursive)} \quad \mathcal{Q}^+ = mq + \sum_{\lambda\in\mathrm{Par},|\lambda|\geqslant 3}\left(\sum_{i=1}^{|\lambda|-2} t^i\right) z_\lambda^{-1}(\mathcal{Q}^+)^{[\lambda]}$$

$$\text{(Exponential)} \quad \mathrm{Exp}(t\mathcal{Q}^+) = t^2\,\mathrm{Exp}(\mathcal{Q}^+) + (1-t)(1+t+mqt).$$

Proof. For (1), it suffices to show that $\mathsf{s}_{(1,1)}\circ\mathcal{Q} = \frac{1}{2}(\mathcal{Q}^2 - \mathcal{Q}^{[2]})$. This follows from the identity $\mathsf{s}_{(1,1)} = \frac{1}{2}(\mathsf{p}_1^2 - \mathsf{p}_2)$. For (2) and (3), see Remark 3.8. □

Corollary 5.12. *For $n \geqslant 3$, $\mathcal{P}_n$ and $\mathcal{Q}_n$ are related by*

$$(1+t)\mathcal{P}_n = \mathcal{Q}_n - \frac{1}{2}t\left(\sum_{h=2}^{n-2}\mathcal{Q}_h\mathcal{Q}_{n-h} - \mathcal{Q}_{\frac{n}{2}}^{[2]}\right), \tag{5.2}$$

where $\mathcal{Q}_{\frac{n}{2}} = 0$ for n odd. Set $\mathcal{Q}_1^+ = m$. For $n \geqslant 2$, $\mathcal{Q}_n^+$ and $\mathcal{Q}_n$ satisfy

$$\mathcal{Q}_n^+ = \sum_{\lambda\vdash n}\left(\sum_{i=1}^{\ell(\lambda)-2} t^i\right)\prod_{j=1}^{m}\left(\mathsf{h}_{r_j}\circ\mathcal{Q}_{n_j}^+\right) \quad \text{and} \quad \mathcal{Q}_n = \sum_{\lambda\vdash n}\prod_{j=1}^{m}\left(\mathsf{h}_{r_j}\circ\mathcal{Q}_{n_j}^+\right),$$

where n_j denote the parts of λ with multiplicities r_j so that $\lambda = (n_1^{r_1},\dots,n_m^{r_m})$ with $n_1 > \cdots > n_m > 0$ and $\ell(\lambda) = \sum_{j=1}^m r_j$.

Remark 5.13. Note that $\mathsf{h}_r\circ\mathcal{Q}_n^+ = \sum_{\lambda\vdash r} z_\lambda^{-1}(\mathcal{Q}_n^+)^{[\lambda]}$ for $r, n \geqslant 1$, by the same argument as in Remark 3.8.

For later use, we record a different but equivalent form of (5.2). For any integer k, reading the coefficients of t^k in (5.2), we obtain

$$\mathcal{P}_{n,k} + \mathcal{P}_{n,k-1} = \mathcal{Q}_{n,k} - \frac{1}{2}\sum_{2\leqslant h\leqslant n-2}\sum_{j=0}^{k-1}\mathcal{Q}_{h,j}\mathcal{Q}_{n-h,k-j-1} + \frac{1}{2}\mathcal{Q}_{\frac{n}{2},\frac{k-1}{2}}, \tag{5.3}$$

where we set $\mathcal{Q}_{h,j} = 0$ whenever h or j is not an integer.

Using the algorithm in this section, we can study the log-concavities of $\mathcal{P}_n(m,t)$ and $\mathcal{Q}_n(m,t)$ (cf. Definition 2.1). Refining Definition 2.1(2), we say that a polynomial $A(m,t)$ in (2.2) is length log-concave in degree k *at length* ℓ if $A_k(m) = \sum_j A_k^j m^j$ is log-concave at $j = \ell$, that is, $(A_k^\ell)^2 \geqslant A_k^{\ell-1}A_k^{\ell+1}$.

Based on numerical computations, we propose the following.

Conjecture 5.14. *The polynomials $\mathcal{P}_n(m,t)$ and $\mathcal{Q}_n(m,t)$ are*

(1) *length log-concave at $t = 1$,*
(2) *length log-concave in degree k (for any given k) at length $\ell > 2$,*
(3) *degree log-concave at $m = 1$, and*

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(4) *degree log-concave in length ℓ for $\ell \geqslant \sqrt{n}+2$.*

We verified this conjecture up to $n \leqslant 72$. In §6, we will prove (3) and (4) for n sufficiently large as well as (2) partially.

Remark 5.15.

(1) Based on explicit computations, we observe that Conjecture 5.14(2) does *not* hold in general for $\ell = 2$. The following is the complete list of pairs (n,k) with $n \leqslant 72$ for which the polynomials $\mathscr{P}_n(m,t)$ and $\mathscr{Q}_n(m,t)$ fail to be length log-concave in degree k at length $\ell = 2$.
- For $\mathscr{P}_n(m,t)$:
  - $n = 35$ and $k = 7, 26$;
  - $n = 60$ and $k = 25, 32$;
  - $n = 63$ and $k = 9, 11, 49, 51$;
  - $n = 72$ and $k = 27, 30, 32, 33, 36, 37, 39, 42$.
- For $\mathscr{Q}_n(m,t)$:
  - $n = 72$ and $k = 28, 30, 40, 42$.

(2) It is also observed that, regarding Conjecture 5.14(4), $\mathscr{P}_n(m,t)$ and $\mathscr{Q}_n(m,t)$ are *not* degree log-concave in all lengths in general. For example, since $P_5 = \mathsf{h}_5 + \mathsf{h}_4\mathsf{h}_1 t + \mathsf{h}_5 t^2$, we have

$$\begin{aligned}\mathscr{P}_5(m,t) = {} & \frac{m^5}{120} + \frac{m^4}{12} + \frac{7m^3}{24} + \frac{5m^2}{12} + \frac{m}{5} \\ & + \left(\frac{m^5}{24} + \frac{m^4}{4} + \frac{11m^3}{24} + \frac{m^2}{4}\right)t \\ & + \left(\frac{m^5}{120} + \frac{m^4}{12} + \frac{7m^3}{24} + \frac{5m^2}{12} + \frac{m}{5}\right)t^2,\end{aligned}$$

which is not degree log-concave in length 1 and 2.

Furthermore, when $n = 6$, applying (6.1) yields the sequence

$$\{\mathscr{Q}^1_{6,k}\}_{k=0}^4 = \left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right\}.$$

Hence, $\mathscr{Q}_6$ is not degree log-concave in length 1. We note that the bound $\sqrt{n}+2$ in Conjecture 5.14(4) is not sharp.

In §6, we will use Theorem 5.11 and Corollary 5.12 to deduce asymptotic formulas and asymptotic log-concavities for the characteristic polynomials of $\overline{\mathcal{M}}_{0,n}$ and $\overline{\mathcal{M}}_{0,n+1}$.

5.3 | Interpretation as the count of colorings

In this subsection, we present an interpretation of the characteristic polynomial $\mathcal{S}_{U_T}(m)$ of (5.1) and its chromatic polynomial $\mathcal{S}^{\omega}_{U_T}(m)$ in terms of colorings, which may be of independent interest.

We first recall an inductive decomposition of weighted rooted trees, based on which the recursion in Theorem 5.10 was established in [11].

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Lemma 5.16 [11, Lemma 4.4]. *Let $T \in \mathscr{T}_{n,k}$. Let v_0 denote the root vertex of T. Let n_0 and k_0 be the number of the inputs attached to v_0 and the weight given at v_0, respectively. Let $v_1, \dots, v_p$ be the nonroot vertices of T adjacent to v_0. For $1 \leqslant i \leqslant p$, let T_i' denote the weighted rooted subtree of T such that*

(1) *its underlying rooted subtree has v_i as its root;*
(2) *$V(T_i')$ consists of $u \in V(T)$ such that u and v_i are connected by a path in T that does not pass through v_0;*
(3) *the edges (resp. inputs) of T_i' are precisely those of T connecting (resp. attached to) the vertices in $V(T_i') \subset V(T)$;*
(4) *the weight function on $V(T_i)$ is the restriction of that of T.*

Let $T_0 \in \mathscr{T}_{n_0,k_0}$ be the weighted rooted tree with only one vertex. Now let

$$\{T_1', \dots, T_p'\} = \{T_1, \dots, T_1, \dots, T_l, \dots, T_l\},$$

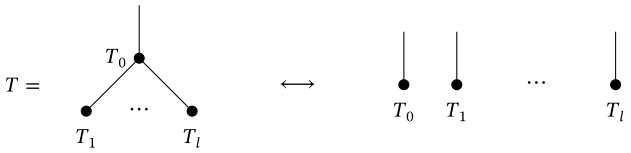
where T_j are mutually nonisomorphic and there are r_j copies of T_j. Assume that $T_j \in \mathscr{T}^+_{n_j,k_j}$ so that $n = n_0 + \sum_{j=1}^{l} r_j n_j$ and $k = k_0 + \sum_{j=1}^{l} r_j k_j$. Then,

$$\mathrm{ch}_{\mathbb{S}_n} U_T = \mathsf{h}_{n_0} \cdot \prod_{j=1}^{l} \mathsf{h}_{r_j} \circ \mathrm{ch}_{\mathbb{S}_{n_j}}(U_{T_j}). \tag{5.4}$$

Furthermore, by taking the involution ω, we have

$$\omega(\mathrm{ch}_{\mathbb{S}_n} U_T) = \mathsf{e}_{n_0} \prod_{r_j:\ \text{even}} \mathsf{h}_{r_j} \circ \omega(\mathrm{ch}_{\mathbb{S}_{n_j}} U_{T_j}) \prod_{r_j:\ \text{odd}} \mathsf{e}_{r_j} \circ \omega \left(\mathrm{ch}_{\mathbb{S}_{n_j}} U_{T_j}\right). \tag{5.5}$$

Pictorially, T decomposes as follows.



Proof. Since

$$\mathrm{Stab}(T) \cong \mathbb{S}_{n_0} \times \prod_{j=1}^{l} \left(\mathrm{Stab}(T_j)^{r_j} \rtimes \mathbb{S}_{r_j}\right) \tag{5.6}$$

by definition, (5.4) is immediate. See also [11, Lemmas 4.3]. Now, (5.5) follows from (5.4) by [21, I.8.Example 1]. □

The following is immediate from Lemma 5.16.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Lemma 5.17. *In the above notations,*

$$\mathcal{S}_{U_T}(m) = \left(\!\!\binom{m}{n_0}\!\!\right) \cdot \prod_{j=1}^{l} \left(\!\!\binom{\mathcal{S}_{U_{T_j}}(m)}{r_j}\!\!\right), \text{ and}$$

$$\mathcal{S}^{\omega}_{U_T}(m) = \binom{m}{n_0} \cdot \prod_{r_j:\,\text{even}} \left(\!\!\binom{\mathcal{S}_{U_{T_j}}(m)}{r_j}\!\!\right) \cdot \prod_{r_j:\,\text{odd}} \binom{\mathcal{S}_{U_{T_j}}(m)}{r_j}$$

as polynomials in m. □

Now one can interpret the symmetric function $\mathrm{ch}_{\mathbb{S}_n} U_T$ as a symmetric function encoding colorings of the inputs of T as follows. For a labeled weighted rooted tree T, let I_T be the set of the (labeled) inputs of T and let

$$\mathscr{K}_T := \{\kappa : I_T \to \mathbb{Z}_{\geqslant 1}\}$$

be the set of *colorings* of the inputs of T (by the colors indexed by positive integers). We consider two colorings to be the same if they are in the same orbit of $\mathrm{Stab}(T) \subset \mathbb{S}_{I_T}$. Given a coloring κ of T, we denote

$$x^{\kappa} := \prod_{i \in I_T} x_{\kappa(i)}.$$

Proposition 5.18. *Let $T \in \mathscr{T}^{\mathrm{lab}}_{n,k}$. Then,*

$$\mathrm{ch}_{\mathbb{S}_n} U_T = \sum_{\kappa \in \mathscr{K}_T} x^{\kappa}. \tag{5.7}$$

Proof. This is a consequence of Pólya theory. See Theorem 7.24.4 and (7.120) in [29]. However, for the reader's convenience, we provide a proof using Lemma 5.16 and the induction on the number of vertices of T.

When T has only one vertex, the root, then $\mathrm{ch}_{\mathbb{S}_n} U_T = \mathsf{h}_n$ and (5.7) is immediate. Suppose that $|V(T)| > 1$. The identity

$$\mathsf{h}_r \circ \sum_{\kappa \in \mathscr{K}_T} x^{\kappa} = \sum_{\kappa_1, \kappa_2, \ldots, \kappa_r \in \mathscr{K}_T} x^{\kappa_1} x^{\kappa_2} \cdots x^{\kappa_r}$$

is immediate from the definition of the plethysm product (cf. [21, I.8]). Combining this with the decomposition in Lemma 5.16, we have

$$\mathrm{ch}_{\mathbb{S}_n} U_T = \left(\sum_{\kappa_0 \in \mathscr{K}_{T_0}} x^{\kappa_0} \right) \prod_{j=1}^{l} \sum_{\kappa_{j1}, \kappa_{j2}, \ldots, \kappa_{jr_j} \in \mathscr{K}_{T_j}} x^{\kappa_{j1}} x^{\kappa_{j2}} \cdots x^{\kappa_{jr_j}} = \sum_{\kappa \in \mathscr{K}_T} x^{\kappa},$$

where the first equality holds by the induction hypothesis applied to each T_j and the second equality holds since the product maps

$$\kappa_0 \times \prod_{1 \leqslant j \leqslant l, 1 \leqslant i \leqslant r_j} \kappa_{ji} \;:\; I_{T_0} \times \prod_{1 \leqslant j \leqslant l, 1 \leqslant i \leqslant r_j} I_{T_j} \to \mathbb{Z}_{\geqslant 1}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

are precisely the colorings of T. □

An analogous statement holds for $\omega(\mathrm{ch}_{\mathbb{S}_n} T)$ if we imposes a "properness" condition on colorings of the inputs as follows. For each vertex $v \in V(T)$, there exists a unique weighted rooted subtree T_v of T that satisfies all the conditions (1)–(4) in Lemma 5.16 with v in the place of v_i. In particular, T_v has v as its root vertex.

Definition 5.19. Let $T \in \mathcal{T}^{\mathrm{lab}}_{n,k}$. We call a coloring of the inputs of T *proper* if it satisfies the following two conditions:

(1) if two different inputs are attached to the same vertex, then they are colored by different colors;
(2) if two disjoint weighted rooted subtrees of the form T_v are isomorphic by the action of $\mathrm{Stab}(T)$ and each subtree has an *odd* number of inputs, then they are nonisomorphic as colored weighted rooted trees.

We denote by $\mathscr{K}'_T$ the set of all proper colorings of the inputs of T.

Proposition 5.20. *Let $T \in \mathcal{T}^{\mathrm{lab}}_{n,k}$. Then,*

$$\omega(\mathrm{ch}_{\mathbb{S}_n} U_T) = \sum_{\kappa \in \mathscr{K}'_T} x^{\kappa}. \tag{5.8}$$

Proof. The proof is essentially the same as that of Proposition 5.18, except that we need an additional identity

$$\mathrm{e}_r \circ \sum_{\kappa \in \mathscr{K}'_T} x^{\kappa} = \sum_{\substack{\kappa_1, \kappa_2, \dots, \kappa_r \in \mathscr{K}'_T \\ \text{distinct}}} x^{\kappa_1} x^{\kappa_2} \cdots x^{\kappa_r}$$

for any $r \geqslant 1$ and weighted rooted trees T, which is also immediate from the definition of the plethysm product [21, I.8]. We omit the details. □

We will call the right-hand side of (5.8) *the chromatic symmetric function* of a labeled weighted rooted tree T, which is clearly a symmetric function by definition. Proposition 5.20 shows that the Frobenius characteristic of the ω-dual of the representation U_T is an analog for T of the chromatic symmetric function of a graph. However, in our case, the coloring is on inputs not on vertices, the properness condition is different from the usual one, and two colorings in the same orbit of $\mathrm{Stab}(T)$ are identified.

The following is now immediate.

Corollary 5.21. *Let $T \in \mathcal{T}^{\mathrm{lab}}_{n,k}$ and $m_0 \geqslant 1$. Then*

(1) *$\mathcal{S}_{U_T}(m_0)$ is the number of colorings of the inputs of T by m_0 colors.*
(2) *$\mathcal{S}^{\omega}_{U_T}(m_0)$ is the number of proper colorings of the inputs of T by m_0 colors.*

Moreover, $\mathcal{S}_{U_T}(m) = (-1)^n \mathcal{S}^{\omega}_{U_T}(-m)$ and it has nonnegative coefficients. □

Proof. (1) and (2) are immediate from Propositions 5.18 and 5.20 respectively. The last assertion follows from Lemma 3.5, (3.19) and the fact that U_T is a permutation representation. □

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Remark 5.22. Although Corollary 5.21(1) can be viewed as a special case of Example 3.14 with $H = \mathrm{Stab}(T)$, we present an independent proof above using Lemma 5.16, to provide a unified proof of (1) and (2) in Corollary 5.21.

Example 5.23. Suppose that T has only one vertex, namely, the root, and all the n inputs are attached to the root, as in Example 5.7. Then,

$$\mathrm{ch}_{\mathbb{S}_n} U_T = \mathsf{h}_n \quad \text{so that} \quad \mathcal{S}_{U_T}(m) = \left(\!\!\binom{m}{n}\!\!\right).$$

This counts the number of colorings of the n inputs of T by m colors.

Example 5.24. Suppose that T has only two vertices with a inputs attached to the root and the remaining $n-a$ inputs attached to the other nonroot vertex, as in Example 5.8. Then,

$$\mathrm{ch}_{\mathbb{S}_n} U_T = \mathsf{h}_a \mathsf{h}_{n-a} \quad \text{so that} \quad \mathcal{S}_{U_T}(m) = \left(\!\!\binom{m}{a}\!\!\right)\left(\!\!\binom{m}{n-a}\!\!\right).$$

This counts the number of colorings of the inputs of T by m colors.

Example 5.25. Suppose that T be the weighted rooted tree with three vertices, given in Example 5.9. Then, $\mathrm{ch}_{\mathbb{S}_n}(U_T) = \mathsf{h}_{n-2a} \cdot (\mathsf{h}_2 \circ \mathsf{h}_a)$ so that

$$\mathcal{S}_{U_T}(m) = \left(\!\!\binom{m}{n-2a}\!\!\right)\left(\!\!\binom{\left(\!\binom{m}{a}\!\right)}{2}\!\!\right).$$

This counts the number of colorings of the inputs of T by m colors.

Remark 5.26 (ω-twisted version). In the above examples, one can easily see that taking ω corresponds to restricting to proper colorings. Indeed,

$$\mathcal{S}^{\omega}_{\mathsf{h}_n}(m) = \binom{m}{n}$$

counts the number of colorings of n inputs not allowing repetition of colors. Similarly, from the formula of $\mathcal{S}_{U_T}(m)$ in Example 5.25, it follows that

$$\mathcal{S}^{\omega}_{U_T}(m) = (-1)^n \mathcal{S}_{U_T}(-m) = \begin{cases} \binom{m}{n-2a} \cdot \left(\!\!\binom{\binom{m}{a}}{2}\!\!\right) & \text{if } a \text{ is even,} \\ \binom{m}{n-2a} \cdot \binom{\binom{m}{a}}{2} & \text{if } a \text{ is odd.} \end{cases}$$

This counts the number of proper colorings of the inputs of T by m colors.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

6 | ASYMPTOTIC BEHAVIOR AND LOG-CONCAVITY

In this section, we study the asymptotic behavior and log-concavity of the polynomials $\mathcal{Q}_n$ and $\mathscr{P}_n$ as n grows. Since they are polynomials in two variables m and t, we may consider several notions of log-concavity in Definition 2.1.

We write

$$\mathscr{P}_{n,k}(m) = \sum_{k\geqslant 0} \mathscr{P}_{n,k}^{j} m^j, \quad \mathcal{Q}_{n,k}(m) = \sum_{k\geqslant 0} \mathcal{Q}_{n,k}^{j} m^j$$

and provide asymptotic formulas for

- (Theorem 6.1) $\mathscr{P}_{n,k}(m_0)$ and $\mathcal{Q}_{n,k}(m_0)$ with $m_0 \geqslant 1$;
- (Theorem 6.6) $\mathscr{P}_{n,k}^{n-j}$ and $\mathcal{Q}_{n,k}^{n-j}$

as n grows with k, j fixed, based on the combinatorial description in the previous section.

We then establish the asymptotic degree log-concavity at $m = m_0$ (Corollary 6.9), the asymptotic degree ultra log-concavity in length $n-j$ (Corollary 6.10) and the asymptotic length log-concavity in degree k (Corollary 6.11) for fixed j, k, by using the asymptotic formulas. The asymptotic formulas and the log-concavity results here generalize those in [3] and [11].

In the remainder, we will use the notation

$$c_k = \frac{(k+1)^{k-1}}{k!} \quad \text{and} \quad d_k = \frac{(k+1)^{k-2}}{k!}.$$

These are the Cayley number $(k+1)^{k-1}$ of trees on $k+1$ labeled vertices, divided by $k!$ and $(k+1)!$, respectively.

6.1 | Asymptotic formulas

Let us first find asymptotic formulas for $\mathscr{P}_{n,k}(m_0)$ and $\mathcal{Q}_{n,k}(m_0)$.

Theorem 6.1. *Fix $m_0 \geqslant 1$ and $k \geqslant 0$. Then, as n grows,*

$$\mathcal{Q}_{n,k}(m_0) = \frac{c_k}{((k+1)m_0-1)!} n^{(k+1)m_0-1} + O(n^{(k+1)m_0-2}),$$

$$\mathscr{P}_{n,k}(m_0) = \frac{d_k}{((k+1)m_0-1)!} n^{(k+1)m_0-1} + o(n^{(k+1)m_0-1}).$$

Remark 6.2. By Lemma 3.10 (2),

$$\mathscr{P}_{n,k}(1) = \dim H^{2k}(\overline{\mathcal{M}}_{0,n})^{\mathbb{S}_n} \quad \text{and} \quad \mathcal{Q}_{n,k}(1) = \dim H^{2k}(\overline{\mathcal{M}}_{0,n+1})^{\mathbb{S}_n}.$$

Hence, if we let $m_0 = 1$, Theorem 6.1 recovers the asymptotic formula [11, Theorem 1.9]. Likewise, Corollary 6.9 below with $m_0 = 1$ recovers [11, Corollary 6.10].

See §7.1 for a proof of Theorem 6.1. Here, we derive the formulas for $\mathcal{Q}_{n,k}(m_0)$ when $k \leqslant 2$ as illustrative examples.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Example 6.3 ($k = 0$). There exists only one tree in $\mathcal{T}_{n,0}$, which is the one described in Example 5.23 with weight 0 at the root. Thus,

$$\mathcal{Q}_{n,0}(m_0) = \left(\!\!\binom{m_0}{n}\!\!\right) = \binom{m_0+n-1}{m_0-1} = \frac{n^{m_0-1}}{(m_0-1)!} + O(n^{m_0-2}).$$

Example 6.4 ($k = 1$). We have two types of weighted rooted trees in $\mathcal{T}_{n,1}$, described in Examples 5.23 and 5.24, respectively. Hence,

$$\mathcal{Q}_{n,1}(m_0) = \left(\!\!\binom{m_0}{n}\!\!\right) + \sum_{a=1}^{n-3} \left(\!\!\binom{m_0}{a}\!\!\right) \cdot \left(\!\!\binom{m_0}{n-a}\!\!\right).$$

To obtain the asymptotic formula, notice that for $m_0 \geqslant 1$,

$$\sum_{a=0}^{n} \left(\!\!\binom{m_0}{a}\!\!\right) \cdot \left(\!\!\binom{m_0}{n-a}\!\!\right) = \left(\!\!\binom{2m_0}{n}\!\!\right) = \frac{n^{2m_0-1}}{(2m_0-1)!} + O(n^{2m_0-2}),$$

while $\left(\!\!\binom{m_0}{n}\!\!\right) = \frac{n^{m_0-1}}{(m_0-1)!} + O(n^{m_0-2})$. Since $2m_0 - 1 > m_0 - 1$ for $m_0 \geqslant 1$,

$$\mathcal{Q}_{n,1}(m_0) = \frac{n^{2m_0-1}}{(2m_0-1)!} + O(n^{2m_0-2}).$$

Example 6.5 ($k = 2$). The contribution of weighted rooted trees T in $\mathcal{T}_{n,2}$ with at most two vertices, to $\mathcal{Q}_{n,2}(m_0)$, is

$$\left(\!\!\binom{m_0}{n}\!\!\right) + \sum_{a=4}^{n-1} \left(\!\!\binom{m_0}{a}\!\!\right) \cdot \left(\!\!\binom{m_0}{n-a}\!\!\right) + \sum_{a=3}^{n-2} \left(\!\!\binom{m_0}{a}\!\!\right) \cdot \left(\!\!\binom{m_0}{n-a}\!\!\right)$$

$$\leqslant \left(\!\!\binom{m_0}{n}\!\!\right) + 2\sum_{a=0}^{n} \left(\!\!\binom{m_0}{a}\!\!\right) \cdot \left(\!\!\binom{m_0}{n-a}\!\!\right) = \frac{2n^{2m_0-1}}{(2m_0-1)!} + O(n^{2m_0-2}).$$

Now we compute the contribution of $T \in \mathcal{T}_{n,2}$ with three vertices. There are three types depending on the shape of trees without inputs and $\mathrm{Stab}(T)$ so that

$$\sum_{a\geqslant 3,\ b\geqslant 2} \left(\!\!\binom{m_0}{a}\!\!\right) \cdot \left(\!\!\binom{m_0}{b}\!\!\right) \cdot \left(\!\!\binom{m_0}{n-a-b}\!\!\right)$$

$$+ \sum_{\substack{3\leqslant a<b\leqslant n,\\ a+b\leqslant n}} \left(\!\!\binom{m_0}{a}\!\!\right) \cdot \left(\!\!\binom{m_0}{b}\!\!\right) \cdot \left(\!\!\binom{m_0}{n-a-b}\!\!\right) + \sum_{3\leqslant a<\frac{n}{2}} \left(\!\!\binom{\left(\!\!\binom{m_0}{a}\!\!\right)}{2}\!\!\right) \cdot \left(\!\!\binom{m_0}{n-2a}\!\!\right),$$

where the last term is computed in Example 5.25, and it is dominated by the second term. Hence, we find that the sum is equal to

$$(1+\frac{1}{2}) \cdot \sum_{\substack{a,b,c\geqslant 0,\\ a+b+c=n}} \left(\!\!\binom{m_0}{a}\!\!\right) \cdot \left(\!\!\binom{m_0}{b}\!\!\right) \cdot \left(\!\!\binom{m_0}{c}\!\!\right) = \frac{3}{2} \cdot \left(\!\!\binom{3m_0}{n}\!\!\right) = \frac{3}{2} \cdot \frac{n^{3m_0-1}}{(3m_0-1)!}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

modulo $O(n^{3m_0-2})$, as a function of n. In particular,

$$\mathcal{Q}_{n,2}(m_0) = \frac{3}{2} \cdot \frac{n^{3m_0-1}}{(3m_0-1)!} + O(n^{3m_0-2}).$$

We next present asymptotic formulas for the coefficients $\mathcal{P}_{n,k}^{n-j}$ and $\mathcal{Q}_{n,k}^{n-j}$ of $\mathcal{P}_{n,k}(m)$ and $\mathcal{Q}_{n,k}(m)$ as n grows. Recall that the *signless Stirling number of the first kind*, denoted by $c(n,j)$, is the number of elements $\sigma \in \mathbb{S}_n$ with j disjoint cycles.

Theorem 6.6. *Let $k, j \in \mathbb{Z}_{\geqslant 0}$. Then, as a function of n,*

$$\mathcal{Q}_{n,k}^{n-j} = c_k \cdot \frac{(k+1)^{n-j} c(n,n-j)}{n!} + O\left(\frac{k^n n^{2j}}{n!}\right) \quad \text{and}$$

$$\mathcal{P}_{n,k}^{n-j} = d_k \cdot \frac{(k+1)^{n-j} c(n,n-j)}{n!} + o\left(\frac{(k+1)^n n^{2j}}{n!}\right).$$

We remark that $O(\frac{(k+1)^{n-j}c(n,n-j)}{n!}) = O(\frac{(k+1)^n n^{2j}}{n!})$, as $c(n,n-j)$ and $\frac{n^{2j}}{2^j j!}$ are asymptotically equal, that is, their ratio converges to 1 as $n \to \infty$.

See §7.4 for a proof of Theorem 6.6.

Remark 6.7 (Leading coefficients). The first formula in Theorem 6.6 with $j = 0$ and Lemma 3.10 (1) imply that

$$\frac{h^{2k}(\overline{\mathcal{M}}_{0,n+1})}{n!} = \mathcal{Q}_{n,k}^{n} = c_k \cdot \frac{(k+1)^n}{n!} + O\left(\frac{k^n}{n!}\right).$$

We thus obtain

$$\lim_{n\to\infty} \frac{h^{2k}(\overline{\mathcal{M}}_{0,n+1})}{(k+1)^n} = c_k,$$

which is a result of Aluffi–Chen–Marcolli (cf. [3, Theorem 1.3] with n replaced by $n+1$).

Remark 6.8 (Linear terms). An analogous statement for $\mathcal{Q}_{n,k}^{j}$ is not true. This failure can be observed even for the linear terms. Note that $\mathcal{Q}_{n,k}$ is always divisible by m. One can see from (3.19) that $\mathcal{Q}_{n,k}^{1} = \frac{1}{n}$ whenever n is a prime number. Indeed, when n is prime, $\mathrm{Stab}(T) \subset \mathbb{S}_n$ contains an n-cycle only if T has only one vertex and $\mathrm{Stab}(T) = \mathbb{S}_n$. Since the number of n-cycles in $\mathrm{Stab}(T) = \mathbb{S}_n$ is precisely $(n-1)!$, the coefficient of m is $\frac{(n-1)!}{n!} = \frac{1}{n}$.

On the other hand, when $n = 2p \geqslant 6$ for a prime p, the weighted rooted tree T consisting of two nonroot vertices attached to the root, each with p inputs, contributes an additional $\frac{1}{n}$, provided that T satisfies the valency conditions (cf. Definition 5.4).

More generally, by the same argument using (3.19), one can check that

$$\mathcal{Q}_{n,k}^{1} = \frac{1}{n} + \frac{1}{n} \sum_{k_0=0}^{k-2} N_{n,k}(k_0), \tag{6.1}$$

where for $0 \leqslant k_0 \leqslant k-2$, the integers $N_{n,k}(k_0)$ are defined as

$$N_{n,k}(k_0) := \left| \left\{ r \in \mathbb{Z}_{>1} \,:\, k_0 + 2 \leqslant r \leqslant \frac{n-k+k_0}{2} \text{ and } r \mid gcd(n, k-k_0) \right\} \right|.$$

The number $N_{n,k}(k_0)$ is the contribution from the weighted rooted trees having r nonroot vertices attached to the root, each with $\frac{n}{r}$ inputs, where the root vertex has weight k_0 and nonroot vertices have weight $\frac{k-k_0}{r}$.

6.2 | Asymptotic log-concavity

From Theorem 6.1, we obtain the asymptotic degree log-concavity at $m = m_0 \geqslant 1$ of $\mathscr{P}_n(m,t)$ and $\mathscr{Q}_n(m,t)$.

Corollary 6.9. *Fix $m_0, k \geqslant 1$. Then, we have*

$$\mathscr{P}_{n,k}(m_0)^2 \geqslant \mathscr{P}_{n,k-1}(m_0)\mathscr{P}_{n,k+1}(m_0) \text{ and}$$
$$\mathscr{Q}_{n,k}(m_0)^2 \geqslant \mathscr{Q}_{n,k-1}(m_0)\mathscr{Q}_{n,k+1}(m_0)$$

for sufficiently large n. Moreover, when $m_0 > 1$, the inequalities are strict.

Proof. By Theorem 6.1, it suffices to show that the sequences

$$\left\{ \frac{c_k}{((k+1)m_0-1)!} \right\}_{k\geqslant 0} \quad \text{and} \quad \left\{ \frac{d_k}{((k+1)m_0-1)!} \right\}_{k\geqslant 0}$$

are log-concave. Since the case of $m_0 = 1$ was already proved in [11], we may assume $m_0 > 1$. Then, one can check that

$$\frac{(km_0-1)!((k+2)m_0-1)!}{(((k+1)m_0-1)!)^2} = \frac{(km_0+m_0)}{km_0}\frac{(km_0+m_0+1)}{km_0+1}\cdots\frac{(km_0+2m_0-1)}{(km_0+m_0-1)}$$
$$= \prod_{i=0}^{m_0-1}\left(1+\frac{m_0}{km_0+i}\right) > 1+\frac{1}{k} = \frac{k+1}{k},$$

while $\frac{c_k^2}{c_{k-1}c_{k+1}} = \frac{(k+1)^{2k-1}}{k^{k-1}(k+2)^k}$. Combining these, we have

$$\frac{c_k^2}{c_{k-1}c_{k+1}} \cdot \frac{(km_0-1)!((k+2)m_0-1)!}{(((k+1)m_0-1)!)^2} > \left(1+\frac{1}{k^2+2k}\right)^k > 1.$$

This shows the second (strict) inequality in the assertion. Similarly,

$$\frac{d_k^2}{d_{k-1}d_{k+1}} \cdot \frac{(km_0-1)!((k+2)m_0-1)!}{(((k+1)m-1)!)^2} > \left(1+\frac{1}{k^2+2k}\right)^{k-1} \geqslant 1.$$

This shows the first (strict) inequality in the assertion. □

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

From Theorem 6.6, we establish two asymptotic log-concavities of $\mathscr{P}_n(m,t)$ and $\mathscr{Q}_n(m,t)$. First, we show the asymptotic degree *ultra* log-concavity in length $n-j$ as follows.

Corollary 6.10. *Fix $j \geqslant 0$ and $k \geqslant 1$. Then, for sufficiently large n,*

$$\left(\frac{\mathscr{P}_{n,k}^{n-j}}{\binom{n-3}{k}}\right)^2 \geqslant \left(\frac{\mathscr{P}_{n,k-1}^{n-j}}{\binom{n-3}{k-1}}\right)\left(\frac{\mathscr{P}_{n,k+1}^{n-j}}{\binom{n-3}{k+1}}\right) \text{ and}$$

$$\left(\frac{\mathscr{Q}_{n,k}^{n-j}}{\binom{n-2}{k}}\right)^2 \geqslant \left(\frac{\mathscr{Q}_{n,k-1}^{n-j}}{\binom{n-2}{k-1}}\right)\left(\frac{\mathscr{Q}_{n,k+1}^{n-j}}{\binom{n-2}{k+1}}\right).$$

Proof. When n is sufficiently large, these are by Theorem 6.6 equivalent to log-concavity of $\left\{d_k(k+1)^n/\binom{n-3}{k}\right\}_k$ and $\left\{c_k(k+1)^n/\binom{n-2}{k}\right\}_k$ at given k, respectively. One can easily check that the latter are true for all $k \geqslant 1$. □

This result generalizes the asymptotic ultra log-concavity of the Poincaré polynomial of $\overline{\mathcal{M}}_{0,n}$, previously discovered in [3], which corresponds to the case of $j = 0$.

Next, we examine the length log-concavity of the characteristic polynomials $\mathscr{P}_n(m,t)$ and $\mathscr{Q}_n(m,t)$ in degree k.

Corollary 6.11. *Fix $j \geqslant 1$ and $k \geqslant 0$. Then for sufficiently large n,*

$$\left(\mathscr{P}_{n,k}^{n-j}\right)^2 \geqslant \left(\mathscr{P}_{n,k}^{n-j-1}\right)\cdot\left(\mathscr{P}_{n,k}^{n-j+1}\right) \text{ and}$$

$$\left(\mathscr{Q}_{n,k}^{n-j}\right)^2 \geqslant \left(\mathscr{Q}_{n,k}^{n-j-1}\right)\cdot\left(\mathscr{Q}_{n,k}^{n-j+1}\right).$$

Proof. By Theorem 6.6, when n is sufficiently large, this reduces to the log-concavity of $\{c(n,n-j)\}_j$, which follows immediately from the identity $m(m+1)\cdots(m+n-1) = \sum_{j=0}^{n} c(n,j)m^j$ and Lemma 2.5. □

7 | PROOFS OF ASYMPTOTIC FORMULAS

In this section, we provide proofs of Theorem 6.1 and Theorem 6.6. The proofs are based on the combinatorial description involving the weighted rooted trees. A central observation is that the dominant terms in both asymptotic formulas arise from the rooted tree with the maximum possible number of vertices, which accounts for the appearance of the Cayley numbers in the formula.

7.1 | Proof of Theorem 6.1

Let $\overline{\mathcal{T}}_k$ be the set of isomorphism classes of weighted rooted trees with total weight k and no inputs. Let

$$F_n : \mathcal{T}_{n,k} \longrightarrow \overline{\mathcal{T}}_k$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

be the forgetful map that forgets all the inputs of a given $T \in \mathcal{T}_{n,k}$. In other words, the fiber of $\overline{T} \in \overline{\mathcal{T}}_k$ consists of all possible ways of attaching n inputs to the vertices of $\overline{T}$, subject to the valency conditions (cf. Definition 5.4).

Decompose $\mathcal{Q}_{n,k}$ with respect to F_n:

$$\mathcal{Q}_{n,k} = \sum_{\overline{T} \in \overline{\mathcal{T}}_k} \mathcal{Q}_{n,\overline{T}}, \quad \text{where } \mathcal{Q}_{n,\overline{T}} := \sum_{T \in F_n^{-1}(\overline{T})} \mathcal{S}_{U_T}.$$

Suppose that $j := |V(\overline{T})| \leqslant k$. Then, by Corollary 5.21, $\mathcal{Q}_{n,\overline{T}}(m_0)$ is bounded by the number of all possible ways of attaching n inputs to the vertices of $\overline{T}$ and coloring them using m_0 colors, which is given by

$$\sum_{\substack{n_1+\cdots+n_j=n,\\ n_i \geqslant 0}} \prod_{i=1}^{j} \left(\!\!\binom{m_0}{n_i}\!\!\right) = \left(\!\!\binom{jm_0}{n}\!\!\right) = \frac{n^{jm_0-1}}{(jm_0-1)!} + O(n^{jm_0-2}).$$

In particular, $\mathcal{Q}_{n,\overline{T}}(m) = O(n^{(k+1)m_0-2})$ for all $\overline{T}$ with at most k vertices.

Suppose that $|V(\overline{T})| = k+1$. By a similar argument, we have

$$\mathcal{Q}_{n,\overline{T}}(m_0) = \frac{1}{|\operatorname{Aut}(\overline{T})|} \cdot \frac{n^{(k+1)m_0-1}}{((k+1)m_0-1)!} + O(n^{(k+1)m_0-2}).$$

Furthermore, since $k+1$ is the maximal possible number of vertices that $\overline{T} \in \overline{\mathcal{T}}_k$ can have, every nonroot vertex of $\overline{T}$ has weight one while the root vertex has weight zero. In particular, automorphisms of $\overline{T}$ as weighted rooted trees are precisely those as rooted trees. Therefore, by the same argument as in the proof of [11, Theorem 6.3], we have

$$\sum_{\overline{T} \in \overline{\mathcal{T}}_k,\ |V(\overline{T})|=k+1} \frac{1}{|\operatorname{Aut}(\overline{T})|} = \frac{(k+1)^{k-1}}{|\mathbb{S}_k|} = c_k, \tag{7.1}$$

where $\overline{T}$ are considered as trees with $k+1$ vertices, and $(k+1)^{k-1}$ is the number of trees on $k+1$ labeled vertices by Cayley's tree formula. Consequently,

$$\mathcal{Q}_{n,k}(m_0) = \sum_{\overline{T} \in \overline{T}_k} \mathcal{Q}_{n,\overline{T}}(m_0) = c_k \cdot \frac{n^{(k+1)m_0-1}}{((k+1)m_0-1)!} + O(n^{(k+1)m_0-2}), \tag{7.2}$$

which implies the first formula.

For the second formula, divide both sides of (5.3) by $n^{(k+1)m_0-1}$ and take their limits as $n \to \infty$. Then,

$$\lim_{n\to\infty} \frac{\mathcal{P}_{n,k}(m_0) + \mathcal{P}_{n,k-1}(m_0)}{n^{(k+1)m_0-1}} = \frac{c_k}{((k+1)m_0-1)!}$$

$$- \frac{1}{2} \sum_{j=0}^{k-1} \frac{c_j c_{k-1-j}}{((j+1)m_0-1)!((k-j)m_0-1)!} \cdot \lim_{n\to\infty} \sum_{h=2}^{n-2} \frac{h^{(j+1)m-1}(n-h)^{(k-j)m_0-1}}{n^{(k+1)m_0-1}},$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

due to (7.2). Indeed, for integers $a, b, c \geqslant 0$,

$$\lim_{n\to\infty} \sum_{h=2}^{n-2} h^a (n-h)^b \frac{1}{n^{c+1}} = \int_0^1 x^a (1-x)^b dx = \frac{a!b!}{(a+b+1)!}$$

if $a + b = c$, and hence, it vanishes if $a + b < c$. This also implies that

$$\lim_{n\to\infty} \frac{\mathscr{P}_{n,k}(m_0) + \mathscr{P}_{n,k-1}(m_0)}{n^{(k+1)m_0-1}} = \frac{1}{((k+1)m_0-1)!}\left(c_k - \frac{1}{2}\sum_{j=0}^{k-1} c_j c_{k-1-j}\right),$$

which is equal to $d_k/((k+1)m_0-1)!$ as checked in [11, Theorem 6.9]. Now the second formula in Theorem 6.1 follows by an induction on k. □

The rest of this section is devoted to a proof of Theorem 6.6. It requires certain facts about $\mathrm{Stab}(T)$, as described in the following subsections.

7.2 | Stratifications of Stab(T)

For a labeled weighted rooted tree T, let $F(T)$ denote the weighted rooted tree without inputs obtained from T by forgetting the inputs. Recall that $\mathrm{Aut}(F(T)) \subset \mathbb{S}_{V(F(T))}$ consists of elements whose actions on $V(T) = V(F(T))$ preserve the isomorphism class of $F(T)$ as a weighted rooted tree. Then, the natural group homomorphism $\mathrm{Stab}(T) \to \mathbb{S}_{V(T)}$ factors through $\mathrm{Aut}(F(T))$.

Let $\lambda(T)$ denote the partition of the set of the inputs according to the vertices of T they are attached to. This defines a subgroup $\mathbb{S}_{\lambda(T)} \subset \mathrm{Stab}(T)$.

Lemma 7.1. *Let T be a labeled rooted tree. Let $\bar{T} = F(T)$. Then, $\mathbb{S}_{\lambda(T)}$ is the kernel of the group homomorphism* $\mathrm{Stab}(T) \to \mathrm{Aut}(\bar{T})$.

Proof. This follows from the inductive construction of T and $\mathrm{Stab}(T)$ in Lemma 5.16 and (5.6), and that of $\bar{T}$ and $\mathrm{Aut}(\bar{T})$. □

Example 7.2. Let $\bar{T}$ be the weighted rooted tree without inputs and with $V(\bar{T}) = \{v_0, v_1, v_2, v_3\}$, where v_0 is the root with weight 0 and the other vertices are adjacent to v_0 and have weight 1.

We attach labeled inputs to vertices of $\bar{T}$ to obtain a labeled weighted rooted tree T. Assume that $I = \{1, \dots, 9\}$ is the set of all inputs and I_{v_j} is the subset of the inputs attached to v_j defined as follows:

$$I = I_{v_0} \sqcup I_{v_1} \sqcup I_{v_2} \sqcup I_{v_3}$$

with $I_{v_0} = \emptyset$, $I_{v_1} = \{1, 4, 7\}$, $I_{v_2} = \{2, 5, 8\}$ and $I_{v_3} = \{3, 6, 9\}$. In particular,

$$\mathrm{Stab}(T) = (\mathbb{S}_{I_{v_1}} \times \mathbb{S}_{I_{v_2}} \times \mathbb{S}_{I_{v_3}}) \rtimes \mathbb{S}_3 \subset \mathbb{S}_I = \mathbb{S}_9,$$

with a splitting given by

$$\mathbb{S}_3 \longrightarrow \mathrm{Stab}(T), \quad \tau \mapsto (3i + j \mapsto 3i + \tau(j) \in I)_{0\leqslant i\leqslant 2, 1\leqslant j\leqslant 3}.$$

The splitting identifies $\mathrm{Aut}(\bar{T}) = \mathbb{S}_3$, as a natural quotient of $\mathrm{Stab}(T)$.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

More concretely, an element of $\mathrm{Stab}(T)$ is the permutation σ in $\mathbb{S}_I$ that preserves the partition $I_{v_1} \sqcup I_{v_2} \sqcup I_{v_3}$, that is, $\sigma(I_{v_j}) = I_{\tau(v_j)}$ for each j for some permutation $\tau \in \mathbb{S}_{\{v_1,v_2,v_3\}}$. In this case, τ is the image of σ by the map $\mathrm{Stab}(T) \to \mathrm{Aut}(\bar{T})$.

Definition 7.3. Let T be a labeled weighted rooted tree. Let $\bar{T} = F(T)$. Let n be the number of the inputs of T so that $\mathrm{Stab}(T) \subset \mathbb{S}_n$.

(1) Define the *length* of an element σ in $\mathrm{Stab}(T)$, denoted by $\ell(\sigma)$, as the number of disjoint cycles in σ as an element of $\mathbb{S}_n$, or equivalently, the number of its orbits in the set of labels. This is also the length of the partition of n associated to $\sigma \in \mathbb{S}_n$.
(2) Let $\mathrm{Stab}^{\ell}(T) \subset \mathrm{Stab}(T)$ be the subset of elements of length $\ell > 0$.
(3) Let $\mathrm{Stab}_\tau(T) \subset \mathrm{Stab}(T)$ be the preimage of $\tau \in \mathrm{Aut}(\bar{T})$ under the standard map $\mathrm{Stab}(T) \to \mathrm{Aut}(\bar{T})$.
(4) Let $\mathrm{Stab}_\tau^{\ell}(T) = \mathrm{Stab}^{\ell} \cap \mathrm{Stab}_\tau(T)$ for each ℓ and τ.

In particular, $\mathrm{Stab}(T) = \bigsqcup_{\ell>0,\tau\in\mathrm{Aut}(\bar{T})} \mathrm{Stab}_\tau^{\ell}(T)$.

We define similarly for unlabeled T, in which case every object above is well defined only (as a subset of $\mathbb{S}_n$) up to conjugation.

Example 7.4. Let T, $\bar{T}$, I, and I_{v_j} be as in Example 7.2. Then, the permutation $\sigma = (1,5,9,4,2,6,7,8,3) \in \mathbb{S}_I = \mathbb{S}_9$ lies in $\mathrm{Stab}(T)$. Moreover, its image under the map $\mathrm{Stab}(T) \to \mathrm{Aut}(\bar{T}) = \mathbb{S}_3$ is $\tau := (1,2,3)$, since σ cyclically permutes the subsets $I_{v_1} \to I_{v_2} \to I_{v_3} \to I_{v_1}$. In particular, $\sigma \in \mathrm{Stab}_\tau(T)$.

Example 7.5. Let T be a labeled weighted rooted tree with n inputs. Then, applying (3.19) in Example 3.14 to $H = \mathrm{Stab}(T)$, we obtain

$$\mathcal{S}_{U_T}(m) = \sum_{j=1}^{n} \frac{|\mathrm{Stab}^j(T)|}{|\mathrm{Stab}(T)|} m^j = \sum_{j=1}^{n} \sum_{\tau\in\mathrm{Aut}(\bar{T})} \frac{|\mathrm{Stab}_\tau^j(T)|}{|\mathrm{Stab}(T)|} m^j.$$

This is the same for unlabeled T.

The first step in the proof of Theorem 6.6 is to decompose the coefficient $\mathcal{Q}_{n,k}^{n-j}$. To explain this, we first consider the decomposition of $\mathcal{Q}_{n,k}$ with respect to the map $F_n : \mathcal{T}_{n,k} \to \bar{\mathcal{T}}_k$ as follows:

$$\mathcal{Q}_{n,k} = \sum_{\bar{T}\in\bar{\mathcal{T}}_k} \mathcal{Q}_{n,\bar{T}}, \quad \text{where } \mathcal{Q}_{n,\bar{T}} := \sum_{T\in F_n^{-1}(\bar{T})} \mathcal{S}_{U_T}$$

for $\bar{T} \in \overline{\mathcal{T}}_k$. By definition, $F_n^{-1}(\bar{T}) \subset \mathcal{T}_{n,k}$ consists of weighted rooted trees obtained by attaching n inputs to the vertices of $\bar{T}$.

This decomposes further by Example 7.5: taking the coefficients of m^{n-j},

$$\mathcal{Q}_{n,\bar{T}}^{n-j} = \sum_{\tau\in\mathrm{Aut}(\bar{T})} \sum_{T\in F_n^{-1}(\bar{T})} \frac{|\mathrm{Stab}_\tau^{n-j}(T)|}{|\mathrm{Stab}(T)|}. \tag{7.3}$$

As we will see later, the summand in (7.3) for $\tau = e$ dominates the others as n grows. So, we first obtain an asymptotic formula for the case of $\tau = e$.

When $\tau = e$, the numerator $|\mathrm{Stab}_e^{n-j}(T)|$ counts elements in $\mathrm{Stab}_e(T) \cong \mathbb{S}_{\lambda(T)}$ with $n-j$ disjoint cycles, where the number of inputs attached to each vertex of $\bar{T}$ can vary. Since $\bar{T}$ is fixed, so is $\ell(\lambda(T)) = |V(\bar{T})|$. Thus, the summand for $\tau = e$ is controlled by the $|V(\bar{T})|$th power of the generating function of the signless Stirling numbers of the first kind defined as follows.

Recall that $c(n, j)$ satisfy

$$m(m+1)\cdots(m+n-1) = \sum_{j=0}^{n} c(n,j)m^j \quad \text{and} \quad c(0,0) = 1. \tag{7.4}$$

The (double) generating function of these numbers is known as follows:

$$\mathfrak{S} := \sum_{n\geqslant 1}\sum_{j=1}^{n} \frac{c(n,j)}{n!}u^n v^j = (1-u)^{-v} \quad (= \exp(-v\log(1-u))),$$

where u and v are formal variables.

For notational convenience, we write for given $j, k \geqslant 0$,

$$A_{j,k}(n) = \mathrm{coeff}_{m^{n-j}}\left(\binom{m(k+1)}{n}\right) = \frac{(k+1)^{n-j}c(n,n-j)}{n!}.$$

Note that $\lim_{n\to\infty} c(n,n-j)/(\frac{n^{2j}}{2^j j!}) = 1$. In particular, $O(A_{j,k-1}) = O(\frac{k^n n^{2j}}{n!})$.

Proposition 7.6. *Fix j and $k \geqslant 0$. Then, as n grows,*

$$\sum_{\bar{T}\in\bar{\mathcal{T}}_k}\sum_{T\in F_n^{-1}(\bar{T})} \frac{|\mathrm{Stab}_e^{n-j}(T)|}{|\mathrm{Stab}(T)|} = c_k \cdot A_{j,k}(n) + O\left(\frac{k^n n^{2j}}{n!}\right).$$

Proof. The proposition immediately follows from (7.1) and the following: for $\bar{T} \in \bar{\mathcal{T}}$ with N vertices,

$$\sum_{T\in F_n^{-1}(\bar{T})} \frac{|\mathrm{Stab}_e^{n-j}(T)|}{|\mathrm{Stab}(T)|} = \frac{1}{|\mathrm{Aut}(\bar{T})|}\cdot A_{j,N-1}(n) + O\left(\frac{(N-1)^n n^{2j}}{n!}\right). \tag{7.5}$$

The left-hand side of (7.5) is the summation over all possible N-tuples of partitions of $n_1, \ldots, n_N$ with $\sum n_j = n$ subject to the valency conditions, where each summand is multiplied by $|\mathrm{Aut}(\bar{T})|$. Since the valency conditions are given in the form $n_j \geqslant m_j$ for some m_j depending of the shape of $\bar{T}$ and the weights on the vertices of $\bar{T}$, the left-hand side is asymptotically equal to the summations over such N-tuples with any $n_j \geqslant 0$, upon multiplication by $|\mathrm{Aut}(\bar{T})|$. Therefore,

$$\sum_{T\in F_n^{-1}(\bar{T})} \frac{|\mathrm{Stab}_e^{n-j}(T)|}{|\mathrm{Stab}(T)|} = \frac{1}{|\mathrm{Aut}(\bar{T})|}\sum_{\substack{n_1+\cdots+n_N=n,\\ j_1+\cdots+j_N=j,\\ n_i,j_i\geqslant 0 \text{ for all } i}} \prod_{i=1}^{N} \frac{c(n_i, n_i-j_i)}{n_i!}$$

14602446, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

modulo $O(\frac{(N-1)^n n^{2j}}{n!})$. The summation on the right-hand side of (7.5) is equal to the coefficient of $u^n v^{n-j}$ in

$$\mathfrak{S}^N = (1-u)^{-Nv} = \exp(-Nv\log(1-u))$$

when expanded to a power series. Since

$$\frac{d^n}{du^n}(1-u)^{-Nv} = \prod_{i=0}^{n-1}(Nv+i)\cdot(1-u)^{-Nv-n},$$

the coefficient of u^n in $\mathfrak{S}^N$ is

$$\frac{1}{n!}\frac{d^n}{du^n}(1-u)^{-Nv}|_{u=0} = \frac{1}{n!}\prod_{i=0}^{n-1}(Nv+i).$$

Hence, by (7.4), the coefficient of $u^n v^{n-j}$ in $\mathfrak{S}^N$ is equal to $c(n,n-j)N^{n-j}$ divided by $n!$. □

It remains to show that the summands in (7.3) corresponding to $\tau \neq e$ are asymptotically dominated by the summand corresponding to $\tau = e$. To this end, we introduce the notion of quotient trees by τ in the following subsection. We then compare the summands corresponding to $\tau \neq e$ with the summations associated to the quotient trees by τ. Roughly speaking, the fact that these quotient trees are strictly smaller than the original tree ensures that the associated summations are asymptotically dominated by the one corresponding to $\tau = e$.

7.3 | Quotient trees

Let $\bar{T}$ be a weighted rooted tree without inputs, where its vertices are ordered. Let $\tau \in \mathrm{Aut}(\bar{T}) \subset \mathbb{S}_{V(\bar{T})}$ be an automorphism of $\bar{T}$. Then, there is a natural quotient $\bar{T}/\tau$ of $\bar{T}$, regarded as an element of $\bar{\mathcal{T}}$, by the τ-actions on $V(\bar{T})$ and $E(\bar{T})$ so that

$$V(\bar{T}/\tau) = V(\bar{T})/\tau \quad \text{and} \quad E(\bar{T}/\tau) = E(\bar{T})/\tau,$$

respectively. The weight function $w : V(\bar{T}) \to \mathbb{Z}_{\geqslant 0}$ of $\bar{T}$ naturally factors through $V(\bar{T}/\tau) \to \mathbb{Z}_{\geqslant 0}$. We assign this as the weight function of $\bar{T}/\tau$.

Now let T be a labeled weighted rooted tree. Let $\bar{T} = F(T)$. Suppose that τ is in the image of $\mathrm{Stab}(T) \to \mathrm{Aut}(\bar{T})$. Let n_v denote the number of inputs attached to each vertex v of T. Then, we have $n_u = n_v$ whenever two vertices u and v are in the same τ-orbit. Indeed, any $\sigma \in \mathrm{Stab}_\tau(T)$ induces an isomorphism between two rooted subtrees T_u and T_v in T that have u and v as their root vertices, respectively.

Definition 7.7. We define T/τ to be the weighted rooted tree obtained from $\bar{T}/\tau$ by attaching n_v inputs to the image of each v in $V(\bar{T}/\tau) = V(T)/\tau$.

Roughly speaking, T/τ is obtained from T by identifying all the weighted rooted subtrees T_v for v having the same images in $V(T)/\tau$.

It is natural to expect that $\sigma \in \mathrm{Stab}_\tau(T)$ induces an element in $\mathrm{Stab}_e(T/\tau)$, since its appropriate power (depending on v) will stabilize each vertex v and permute the n_v inputs attached to v. In

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

what follows, we construct a map

$$\mathrm{Stab}_\tau(T) \longrightarrow \mathrm{Stab}_e(T/\tau)$$

and show that it has some useful properties.

Let I be the set of labels on the inputs of T, and for each $v \in V(T)$, let I_v be the set of the labels on the inputs attached to v. Hence, $I := \bigsqcup_{v\in V(T)} I_v$, and $\mathrm{Stab}_\tau(T)$ is a subset of the symmetric group $\mathbb{S}_I$ on I.

Let $d_v > 0$ be the minimal positive integer such that $\tau^{d_v}(v) = v$ for each $v \in V(T)$. Then, there exists a map

$$\varphi_v : \mathrm{Stab}_\tau(T) \longrightarrow \mathbb{S}_{I_v} \subset \mathrm{Stab}(T_v)$$

that sends σ to σ^{d_v} restricted to the set I_v of the inputs attached to v. As $\mathrm{Stab}_e(T/\tau) \cong \mathbb{S}_{\lambda(T/\tau)}$, the desired map can be obtained as the product of the maps φ_v where v runs over representatives of $V(T)/\tau$.

Choose representatives $v_1, \dots, v_p \in V(T)$, one for each τ-orbit, so that $V(T)/\tau \cong \{v_1, \dots, v_p\}$. We consider the product map

$$\prod_{j=1}^{p} \varphi_{v_j} : \mathrm{Stab}_\tau(T) \longrightarrow \prod_{j=1}^{p} \mathbb{S}_{I_{v_j}} = \mathbb{S}_{\lambda(T/\tau)}, \tag{7.6}$$

where I decomposes as $I = \bigsqcup_{j=1}^{p} I_j$ with

$$I_j := I_{v_j} \sqcup I_{\tau(v_j)} \sqcup \cdots \sqcup I_{\tau^{d_j-1}(v_j)},$$

for each j, where $d_j := d_{v_j}$. This map has the following properties.

Lemma 7.8. *The map (7.6) is length-preserving. Moreover, it is surjective and its fibers have the same cardinalities. In particular, for every $\ell > 0$,*

$$\frac{|\mathrm{Stab}_\tau^\ell(T)|}{|\mathrm{Stab}(T)|} \leqslant \frac{|\mathrm{Stab}_e^\ell(T/\tau)|}{|\mathrm{Stab}(T/\tau)|}.$$

The inequality will be useful when we prove Theorem 6.6 in the next subsection, and easily follows from the first two properties, as we will see.

Lemma 7.8 will easily follow from a better understanding of the image of $\mathrm{Stab}_\tau(T) \subset \mathbb{S}_I$. Note that every $\sigma \in \mathrm{Stab}_\tau(T)$ acts on I_j compatibly with the τ-action on the orbit of v_j. In particular, $\mathrm{Stab}_\tau(T) \subset \mathbb{S}_I$ factors through

$$\prod_{j=1}^{p} \left(\mathbb{S}_{I_{v_j}} \times \mathbb{S}_{I_{\tau(v_j)}} \times \cdots \times \mathbb{S}_{I_{\tau^{d_j-1}(v_j)}} \right) \rtimes \langle\tau\rangle \cong \prod_{j=1}^{p} \left(\mathbb{S}_{n_j}^{d_j} \rtimes \mathbb{Z}/d_j\mathbb{Z} \right), \tag{7.7}$$

which is a subgroup of $\prod_{j=1}^{p} \mathbb{S}_{I_j} \subset \mathbb{S}_I$, and even further, it factors through

$$\prod_{j=1}^{p} \left(\mathbb{S}_{I_{v_j}} \times \mathbb{S}_{I_{\tau(v_j)}} \times \cdots \times \mathbb{S}_{I_{\tau^{d_j-1}(v_j)}} \right) \rtimes \{\tau\} \tag{7.8}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

as a subset of (7.7). Here, by abuse of notation, we denote any lift in $\mathbb{S}_{I_j}$ of τ (restricted to the τ-orbit of v_j) again by τ. Note that we may pick any such lift, since any two lifts differ only by reordering labels in each $I_{\tau^i(v_j)}$. It is now clear that $\mathrm{Stab}_\tau(T)$ is equal to (7.8), since both have the same cardinalities $|\mathrm{Stab}_\tau(T)| = |\mathbb{S}_{\lambda(T)}| = \prod_{j=1}^{p} |\mathbb{S}_{n_j}^{d_j}|$.

The above discussion is illustrated in the following example.

Example 7.9. Let T, $\bar{T}$, I, and I_{v_j} with $0 \leqslant j \leqslant 3$ be as in Examples 7.2 and 7.4. The element $\tau = (1,2,3) \in \mathrm{Aut}(\bar{T}) = \mathbb{S}_3$ acts on v_j by $v_1 \mapsto v_2 \mapsto v_3$. The lift of τ in $\mathrm{Stab}(T) \subset \mathbb{S}_I$ via the above splitting is then

$$\tau = (1,2,3)(4,5,6)(7,8,9) \ \in \ \mathbb{S}_I = \mathbb{S}_9, \tag{7.9}$$

which we denote also by τ by abuse of notation. It follows from this that

$$\mathrm{Stab}_\tau(T) = \left(\mathbb{S}_{I_{v_1}} \times \mathbb{S}_{I_{v_2}} \times \mathbb{S}_{I_{v_3}}\right) \cdot \tau \ \subset \ \mathrm{Stab}(T) \ \subset \ \mathbb{S}_I = \mathbb{S}_9$$

as described in (7.8), where $p = 1$ in this case.

Moreover, the map (7.6) is described as follows. Let $\sigma \in \mathrm{Stab}_\tau(T)$. Express $\sigma := (\sigma_1, \sigma_2, \sigma_3; \tau)$ as an element in (7.8). Note that $\sigma_j := \sigma_{v_j} \in \mathbb{S}_{I_{v_j}}$ is chosen so that $\sigma(i) = \sigma_{\tau(v_j)}(\tau(i))$ if $i \in I_{v_j}$. By taking its third power, we obtain

$$\begin{aligned}
\sigma^3 &= (\sigma_1,\ \sigma_2,\ \sigma_3;\ \tau) \cdot (\sigma_1,\ \sigma_2,\ \sigma_3;\ \tau) \cdot (\sigma_1,\ \sigma_2,\ \sigma_3;\ \tau) \\
&= \left(\sigma_1\varphi_\tau(\sigma_3),\ \sigma_2\varphi_\tau(\sigma_1),\ \sigma_3\varphi_\tau(\sigma_2);\ \tau^2\right) \cdot (\sigma_1,\ \sigma_2,\ \sigma_3;\ \tau) \\
&= (\sigma_1\varphi_\tau(\sigma_3)\varphi_{\tau^2}(\sigma_2),\ \sigma_2\varphi_\tau(\sigma_1)\varphi_{\tau^2}(\sigma_3),\ \sigma_3\varphi_\tau(\sigma_2)\varphi_{\tau^2}(\sigma_1);\ e) \\
&= (\sigma_1\tau\sigma_3\tau\sigma_2\tau,\ \sigma_2\tau\sigma_1\tau\sigma_3\tau,\ \sigma_3\tau\sigma_2\tau\sigma_1\tau;\ e)
\end{aligned}$$

in $\mathrm{Stab}(T)$, where $\varphi_\tau(\sigma_j) := \tau\sigma_j\tau^{-1} \in \mathbb{S}_{I_{v_{\tau(j)}}}$. Thus, $\sigma^3 \in \mathbb{S}_{I_{v_1}} \times \mathbb{S}_{I_{v_2}} \times \mathbb{S}_{I_{v_3}}$. The map (7.6) is obtained by composing this with the first projection map

$$\mathrm{Stab}_\tau(T) \xrightarrow{(-)^3} \mathbb{S}_{I_{v_1}} \times \mathbb{S}_{I_{v_2}} \times \mathbb{S}_{I_{v_3}} \xrightarrow{pr_1} \mathbb{S}_{I_{v_1}} \tag{7.10}$$

that sends $(\sigma_1, \sigma_2, \sigma_3; \tau)$ to $\sigma_1\tau\sigma_3\tau\sigma_2\tau \in \mathbb{S}_{I_{v_1}}$.

For example, when $\sigma = (1,5,9,4,2,6,7,8,3) \in \mathrm{Stab}_\tau(T) \subset \mathbb{S}_I = \mathbb{S}_9$, it can be written as $\sigma = (\sigma_1, \sigma_2, \sigma_3; \tau)$ with $\sigma_1 = (4,7)$, $\sigma_2 = (2,5)$, $\sigma_3 = (3,6,9)$ and τ as in (7.9). The image of σ under (7.10) is the first projection of $\sigma^3 = ((1,4,7),(2,8,5),(3,9,6))$, which is $(1,4,7) = \sigma_1\tau\sigma_3\tau\sigma_2\tau \in \mathbb{S}_{I_{v_1}}$.

One can also see that composing the third power map with the second or the third projection, rather than the first projection, yields maps, which are conjugate to each other.

Now we prove Lemma 7.8.

Proof of Lemma 7.8. For each $\sigma \in \mathrm{Stab}_\tau(T)$, the length of σ is equal to the number of its orbits in I while the length of its image under (7.6) is equal to the number of the orbits in $I_{v_1} \sqcup \cdots \sqcup I_{v_p}$.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Since the σ-orbits in I_j are in a bijective correspondence with the σ^{d_j}-orbits in I_{v_j} for every j, (7.6) is length-preserving.

For the second property, we will endow an action of $\mathbb{S}_{\lambda(T/\tau)} = \prod_{j=1}^{p} \mathbb{S}_{I_{v_j}}$ on $\mathrm{Stab}_\tau(T)$ so that (7.6) is $\mathbb{S}_{\lambda(T/\tau)}$-equivariant, with respect to the left multiplication on the target $\mathbb{S}_{\lambda(T/\tau)}$. Since the left multiplication is transitive, this will imply that all the fibers are isomorphic to each other.

By letting $n_j = |I_{v_j}|$, the map φ_{v_j} in (7.6) is of the form

$$\mathbb{S}_{n_j}^{d_j} \rtimes \{\tau\} \hookrightarrow \mathbb{S}_{n_j}^{d_j} \rtimes \mathbb{Z}/d\mathbb{Z} \longrightarrow \mathbb{S}_{n_j}^{d_j} \xrightarrow{pr_1} \mathbb{S}_{n_j}$$

that sends $\sigma := (\sigma_1, \dots, \sigma_d; \tau)$ with $\sigma_i \in S_{n_j}$ and $\tau = (1, \dots, d) \in \mathbb{Z}/d\mathbb{Z}$ to

$$\begin{aligned}\sigma^d &= \left(\sigma_1\varphi_\tau(\sigma_d)\cdots\varphi_{\tau^{d-1}}(\sigma_2),\ \dots,\ \sigma_d\varphi_{\tau(\sigma_{d-1})}\cdots\varphi_{\tau^{d-1}}(\sigma_1)\,;\ e\right)\\ &= (\sigma_1\tau\sigma_d\tau\cdots\tau\sigma_2\tau,\ \dots, \sigma_d\tau\sigma_{d-1}\tau\cdots\tau\sigma_1\tau\,;\ e) \quad \in \mathbb{S}_{n_j}^{d_j}\end{aligned}$$

and then to the first entry $\sigma_1\tau\sigma_d\tau\cdots\tau\sigma_2\tau$. Here, the jth entry of σ^d is

$$\sigma_j\varphi_\tau(\sigma_{\tau^{-1}(j)})\cdots\varphi_{\tau^{d-1}}(\sigma_{\tau^{-(d-1)}(j)}) = \sigma_j\tau\sigma_{\tau^{-1}(j)}\tau\cdots\tau\sigma_{\tau^{-(d-1)}(j)}\tau \quad \in \mathbb{S}_{I_{v_{\tau(j)}}},$$

and we set $\varphi_\tau(a_j) := \tau a_j \tau^{-1} \in \mathbb{S}_{I_{v_{\tau(j)}}}$. Thus, this map is $\mathbb{S}_{n_j}$-equivariant with respect to the left multiplication (only) on the first entry σ_1. In particular, (7.6) is $\prod_{j=1}^{p} \mathbb{S}_{I_{v_j}}$-equivariant.

Finally, the last inequality in the assertion follows from

$$\frac{|\mathrm{Stab}_\tau^\ell(T)|}{|\mathrm{Stab}_e^\ell(T/\tau)|} = \frac{|\mathrm{Stab}_\tau(T)|}{|\mathrm{Stab}_e(T/\tau)|} = \frac{|\mathbb{S}_{\lambda(T)}|}{|\mathbb{S}_{\lambda(T/\tau)}|} \leqslant \frac{|\mathrm{Stab}(T)|}{|\mathrm{Stab}(T/\tau)|},$$

where the last inequality follows from Lemma 7.1. □

7.4 | Proof of Theorem 6.6

Let us complete the proof. Recall that

$$\mathcal{Q}_{n,\bar{T}}^{n-j} = \sum_{T\in F_n^{-1}(\bar{T})} \frac{|\mathrm{Stab}_e^{n-j}(T)|}{|\mathrm{Stab}(T)|} + \sum_{\tau\neq e\in \mathrm{Aut}(\bar{T})} \sum_{T\in F_n^{-1}(\bar{T})} \frac{|\mathrm{Stab}_\tau^{n-j}(T)|}{|\mathrm{Stab}(T)|},$$

where the asymptotic formula for the first summand in the right-hand side is given in Proposition 7.6. When $\tau \neq e$, by Lemma 7.8, we have

$$0 \leqslant \sum_{T\in F_n^{-1}(\bar{T})} \frac{|\mathrm{Stab}_\tau^{n-j}(T)|}{|\mathrm{Stab}(T)|} \leqslant \sum_{T\in F_n^{-1}(\bar{T})} \frac{|\mathrm{Stab}_e^{n-j}(T/\tau)|}{|\mathrm{Stab}(T/\tau)|}.$$

Moreover, whenever the summand is nonzero on the right-hand side, the number n' of inputs of T/τ is $|\lambda(T/\tau)|$, thus greater than or equal to $n-j$. This shows, in particular, that the right-hand side is bounded above by

$$\sum_{n-j\leqslant n'<n}\sum_{T'}\frac{|\mathrm{Stab}_e^{n'-j'}(T')|}{|\mathrm{Stab}(T')|},$$

where T' runs over all the trees obtained by attaching n' inputs to the vertices of $\bar{T}/\tau$ in the way that they do not necessarily satisfy the valency conditions, and $j' := j-(n-n')$ so that $n'-j' = n-j$.

Since $\tau \neq e$, we have the strict inequalities $j' < j$ and $|V(\bar{T}/\tau)| < |V(\bar{T})|$. Hence, this is dominated by the summand corresponding to $\tau = e$ studied in Proposition 7.6. Combining all these results, we have

$$\mathcal{Q}_{n,k}^{n-j} = \sum_{\bar{T}\in\mathcal{T}_k}\mathcal{Q}_{n,\bar{T}}^{n-j} = c_k\cdot A_{j,k}(n) + O\left(\frac{k^n n^{2j}}{n!}\right). \tag{7.11}$$

This completes our proof of the asymptotic formula for $\mathcal{Q}_{n,k}$ in Theorem 6.6.

We prove the case of $\mathcal{P}_{n,k}$. First note that by definition,

$$\sum_{\substack{n_1+n_2=n,\\ j_1+j_2=j}} A_{j_1,k_1}(n_1)\cdot A_{j_2,k_2}(n_2) = A_{j,k_1+k_2+1}(n)$$

for any k_1, k_2 and $j \geqslant 0$. Therefore, for $k \geqslant 0$,

$$\begin{aligned}
&\lim_{n\to\infty}\frac{n!}{(k+1)^{n-j}c(n,n-j)}\sum_{\substack{n_1+n_2=n,\\ j_1+j_2=j}} A_{j_1,k_1}(n_1)\cdot A_{j_2,k_2}(n_2)\\
&= \lim_{n\to\infty}\frac{n!}{(k+1)^{n-j}c(n,n-j)}\cdot A_{j,k_1+k_2+1}(n)\\
&= \lim_{n\to\infty}\left(\frac{k_1+k_2+2}{k+1}\right)^{n-j} = \begin{cases}1 & \text{if } k_1+k_2 = k-1\\ 0 & \text{if } k_1+k_2 < k-1.\end{cases}
\end{aligned}$$

Now read the coefficients of m^{n-j} off (5.3), multiply them by $\frac{n!}{(k+1)^{n-j}c(n,n-j)}$ and take their limits as $n\to\infty$. Then, by (7.11) and the above observation,

$$\lim_{n\to\infty}\frac{n!\cdot\left(\mathcal{P}_{n,k}^{n-j}+\mathcal{P}_{n,k-1}^{n-j}\right)}{(k+1)^{n-j}c(n,n-j)}$$

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is equal to

$$c_k - \frac{1}{2}\lim_{n\to\infty} \sum_{\substack{j_1+j_2=j,\\ n_1+n_2=n,\\ k_1+k_2=k-1}} \frac{n!\cdot \mathcal{Q}_{n_1,k_1}^{n_1-j_1}\cdot \mathcal{Q}_{n_2,k_2}^{n_2-j_2}}{(k+1)^{n-j}c(n,n-j)}$$

$$= c_k - \frac{1}{2}\lim_{n\to\infty} \sum_{\substack{j_1+j_2=j,\\ n_1+n_2=n,\\ k_1+k_2=k-1}} \frac{n!\cdot c_{k_1}\cdot A_{j_1,k_1}(n_1)\cdot c_{k_2}\cdot A_{j_2,k_2}(n_2)}{(k+1)^{n-j}c(n,n-j)}$$

$$= c_k - \frac{1}{2}\lim_{n\to\infty} \sum_{k_1+k_2=k-1} \frac{c_{k_1}c_{k_2}\cdot n!\cdot A_{j,k}(n)}{(k+1)^{n-j}c(n,n-j)} = c_k - \frac{1}{2}\sum_{k_1+k_2=k-1} c_{k_1}c_{k_2},$$

which is equal to d_k. This completes our proof of Theorem 6.6. □

ACKNOWLEDGMENTS

Part of this paper was written during the third author's visit to IAS in the fall of 2024. He is grateful to June Huh for the invitation and to IAS for its hospitality.

J.C. was partially supported by NRF grant 2018R1C1B6005600. D.L. was supported by the Institute for Basic Science IBS-R032-D1 and Korea NRF grant 2021R1A2C1093787.

JOURNAL INFORMATION

The *Proceedings of the London Mathematical Society* is wholly owned and managed by the London Mathematical Society, a not-for-profit Charity registered with the UK Charity Commission. All surplus income from its publishing programme is used to support mathematicians and mathematics research in the form of research grants, conference grants, prizes, initiatives for early career researchers and the promotion of mathematics.

ORCID

Donggun Lee https://orcid.org/0000-0001-9610-2180

REFERENCES

1. H. Abe, M. Harada, T. Horiguchi, and M. Masuda, *The cohomology rings of regular nilpotent Hessenberg varieties in Lie type A*, Int. Math. Res. Not. IMRN **2019** (2019), no. 17, 5316–5388.
2. H. Abe and T. Horiguchi, *A survey of recent developments on Hessenberg varieties*, International Conference on the Trends in Schubert Calculus, vol. 332, 2020, pp. 251–279.
3. P. Aluffi, S. Chen, and M. Marcolli, *Log concavity of the Grothendieck class of* $\overline{\mathcal{M}}_{0,n}$, Preprint, arXiv:2402.02646.
4. P. Aluffi, M. Marcolli, and E. Nascimento, *Explicit formulas for the Grothendieck class of* $\overline{\mathcal{M}}_{0,n}$, Preprint, arXiv:2406.13095.
5. C. A. Athanasiadis, *Power sum expansion of chromatic quasisymmetric functions*, Electron. J. Combin. **22** (2015), no. 2, 2.7, 9.
6. J. Bergström and S. Minabe, *On the cohomology of moduli spaces of (weighted) stable rational curves*, Math. Z. **275** (2013), 1095–1108.
7. P. Brosnan and T. Chow, *Unit interval orders and the dot action on the cohomology of regular semisimple Hessenberg varieties*, Adv. Math. **329** (2018), 955–1001.
8. P. Brosnan, L. Escobar, J. Hong, D. Lee, E. Lee, A. Mellit, and E. Sommers, *Automorphisms and deformations of regular semisimple Hessenberg varieties*, Preprint, arXiv:2405.18313.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms.70096 by Sookmyung Womens University , Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

9. J. Choi and Y.-H. Kiem, *A new method toward the Landau-Ginzburg/Calabi-Yau correspondence via quasimaps*, Math. Z. **294** (2020), 161–199.
10. J. Choi, Y.-H. Kiem, and D. Lee, *Representations on the cohomology of* $\overline{\mathcal{M}}_{0,n}$, Adv. Math. **435** (2023), 109364.
11. J. Choi, Y.-H. Kiem, and D. Lee, *Recursive algorithm and log-concavity of representations on the cohomology of* $\overline{\mathcal{M}}_{0,n}$, Preprint, arXiv:2408.10728.
12. F. De Mari, C. Procesi, and M. A. Shayman, *Hessenberg varieties*, Trans. Amer. Math. Soc. **332** (1992), no. 2, 529–534.
13. C. Eur, L. Ferroni, J. P. Matherne, R. Pagaria, and L. Vecchi, *Building sets, Chow rings, and their Hilbert series*, Preprint, arXiv:2504.16776.
14. M. Goresky, R. Kottwitz, and R. MacPherson, *Equivariant cohomology, Koszul duality, and the localization theorem*, Invent. Math. **131** (1998), no. 1, 25–83.
15. M. Guay-Paquet, *A second proof of the Shareshian-Wachs conjecture, by way of a new Hopf algebra*, Preprint, arXiv:1601.05498.
16. B. Hassett, *Moduli spaces of weighted pointed stable curves*, Adv. Math. **173** (2003), no. 2, 316–352.
17. J. Huh, *Milnor numbers of projective hypersurfaces and the chromatic polynomial of graphs*, J. Amer. Math. Soc. **25** (2012), no. 3, 907–927.
18. M. Kapranov, *Chow quotients of Grassmannians, I*, I. M. Gel'fand seminar, Adv. Soviet Math., vol. 16, Part 2, Amer. Math. Soc., Providence, RI, 1993, pp. 29–110.
19. Y.-H. Kiem and D. Lee, *Birational geometry of generalized Hessenberg varieties and the generalized Shareshian-Wachs conjecture*, J. Combin. Theory Ser. A **206** (2024), 105884.
20. F. C. Kirwan, Cohomology of quotients in symplectic and algebraic geometry, Princeton University Press, Princeton, 1984.
21. I. G. Macdonald, *Symmetric functions and Hall polynomials*, 2nd ed, The Clarendon Press, Oxford University Press, New York, 1995.
22. D. Mumford, J. Fogarty, and F. Kirwan, *Geometric invariant theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete (2) [Results in Mathematics and Related Areas (2)], vol. 34, 3rd ed., Springer, Berlin, 1994.
23. J. Shareshian and M. L. Wachs, *Chromatic quasisymmetric functions*, Adv. Math. **295** (2016), 497–551.
24. E. Sommers and J. Tymoczko, *Exponents of B-stable ideals*, Trans. Amer. Math. Soc. **358** (2006), no. 8, 3493–3509.
25. R. P. Stanley, *Log-concave and unimodal sequences in algebra, combinatorics, and geometry*, Graph theory and its applications: East and West (Jinan, 1986), Ann. New York Acad. Sci., vol. 576, New York Acad. Sci., New York, 1989, pp. 500–535.
26. R. P. Stanley, *A symmetric function generalization of the chromatic polynomial of a graph*, Adv. Math. **111** (1995), no. 1, 166–194.
27. R. P. Stanley, *Enumerative combinatorics. Vol. 1*, 2nd ed., Cambridge Studies in Advanced Mathematics, vol. 49, Cambridge University Press, Cambridge, 2011.
28. R. P. Stanley, *Supplementary exercises for Chapter 7 of Enumerative Combinatorics, Vol. 2*, Exercises available at https://math.mit.edu/rstan/ec/ch7supp.pdf and solutions available at https://math.mit.edu/rstan/ec/ch7suppsol.pdf, 2023.
29. R. P. Stanley, *Enumerative combinatorics. Vol. 2*, 2nd edition, Cambridge Studies in Advanced Mathematics, vol. 208, Cambridge University Press, Cambridge, 2024.
30. J. Tymoczko, *Permutation actions on equivariant cohomology of flag varieties*, Toric topology, Contemp. Math., vol. 460, Amer. Math. Soc., Providence, RI, 2008, pp. 365–384.

1460244x, 2025, 5, Downloaded from https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/jlms.70096 by Sookmyung Womens University, Wiley Online Library on [03/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RESEARCH INSTITUTE OF NATURAL SCIENCES

2025 자연과학연구소 연구결과백서

통계학과

김별 교수

논문

- M. Hellstern, B. Kim, Z. Harchaoui, A. Shojaie (2025). Spectral differential network analysis for high-Dimensional time series. Proceedings of The 28th International Conference on Artificial Intelligence and Statistics. PMLR 258:2512-2520.

Spectral Differential Network Analysis for High-Dimensional Time Series

Michael Hellstern[1] **Byol Kim**[2] **Zaid Harchaoui**[1] **Ali Shojaie**[1]
[1]University of Washington [2]Sookmyung Women's University

Abstract

Spectral networks derived from multivariate time series data arise in many domains, from brain science to Earth science. Often, it is of interest to study how these networks change under different conditions. For instance, to better understand epilepsy, it would be interesting to capture the changes in the brain connectivity network as a patient experiences a seizure, using electroencephalography data. A common approach relies on estimating the networks in each condition and calculating their difference. Such estimates may behave poorly in high dimensions as the networks themselves may not be sparse in structure while their difference may be. We build upon this observation to develop an estimator of the difference in inverse spectral densities across two conditions. Using an ℓ_1 penalty on the difference, consistency is established by only requiring the difference to be sparse. We illustrate the method on synthetic data experiments and on experiments with electroencephalography data.

1 INTRODUCTION

Spectral network analysis of multivariate time series plays a key role in fields ranging from oceanography and seismology to neuroscience (Laurindo et al., 2019; James et al., 2017; Bloch et al., 2022). We use the inverse spectral density matrix as our choice of network as the i, j entry more directly encodes the dependence between nodes i and j compared to other networks in the spectral domain such as coherence Dahlhaus (2000). Studying individual networks can provide insights into how features interact; however, it is often of interest to study how networks change across conditions or in response to an external intervention (Shojaie, 2021). In neuroscience, for example, many neurodegenerative disorders are associated with abnormal brain connectivity networks (Bloch et al., 2022). Spectral features are regularly used to interpret many types of neuroscientific data, from electroencephalography to magnetoencephalography data (Gnassounou et al., 2023; Richard et al., 2020; Dupré la Tour et al., 2018).

Coherence, the frequency domain analog to correlation, is a common choice for investigating interactions in multivariate time series analysis. This notion is especially appealing in neuroscience applications, where activities captured at different frequencies better reveal brain oscillations in sensory-cognitive processes. Therefore, despite the availability of nonlinear association measures, such as mutual information (Belghazi et al., 2018) and transfer entropy (Ursino et al., 2020), coherence is commonly used by neuroscientists to define brain functional connectivity networks.

Similar to correlation, coherence includes the indirect effects of other nodes in the network. Thus, coherence may not be an informative measure of the dependence between nodes. The inverse spectral density is a more direct measure of dependence between nodes as the inverse spectral density between nodes i and j is a rescaling of the coherence after removing the linear effects of all other nodes $\{k \neq i, j\}$ (Dahlhaus, 2000). Therefore, the inverse spectral density better resembles the effective connectivity between brain regions (Friston, 2011), providing an initial understanding of how two regions may be causally related.

Modern data collection methods have facilitated the collection of data, where the dimensionality is much greater than the number of data points. This setting is often referred to as high-dimensional ($p \gg n$). When analyzing high-dimensional time series data, regularization techniques, such as the LASSO, are needed to make the problem computationally and statistically tractable (Banerjee et al., 2008). In the frequency domain, regularization can allow for better numerical sta-

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

bility and better performance (Böhm and von Sachs, 2009).

While several methods exist to directly estimate the difference in inverse covariance matrices (see Tsai et al. (2022) and references therein), no such methods exist for the difference in inverse spectral densities. Naïvely, to estimate the difference in inverse spectral densities, one could first estimate the inverse spectral density in each condition and then take their difference. In high-dimensions, consistency of the resulting estimate requires both networks to be sparse (Deb et al., 2024). This assumption may be unrealistic due to the presence of hub nodes (Wang et al., 2021) and thereby lead to degraded statistical performance.

In this paper, we propose Spectral D-trace Difference (SDD), a direct estimator of the difference in inverse spectral densities, which, to our knowledge, is the first method to directly target a differential network in the spectral domain. By leveraging advances in analysis of time series data, our direct estimator only requires sparsity of the difference, which is a more realistic assumption in many settings (Wang et al., 2021). For instance, despite identifying many nonzero coherence values in resting state and stimulation states, Bloch et al. (2022) found many coherence changes to be near zero. These small differences are likely due to noise in the estimation procedure, corresponding to no underlying change in coherence. Assuming sparsity of the difference, and only under a geometrically decaying time dependence condition, we show that our direct difference estimator consistently estimates the true difference. Performance of our estimator is studied in a variety of simulation settings, as well as a real data application.

Notation. We use $\|A\|_1$, $\|A\|_F$, $\|A\|_\infty$, $\|A\|_0$ to denote the ℓ_1, Frobenius, infinity, and ℓ_0 norms of a matrix A, respectively. The element-wise absolute value of a matrix A is denoted as $|A|$. The conjugate transpose of a matrix or vector will be denoted as A^H. We denote the inner product between two matrices A and B as $\langle A, B\rangle = \mathrm{Tr}(AB^T)$ and use $A \otimes B$ to represent their Kronecker product.

2 METHOD

Figure 1 provides a visual explanation of our method, SDD, and each step (A)–(E) is elaborated upon in this section.

(A) Inputs. Suppose we have mean-zero piecewise stationary data in two conditions, denoted as 1 and 2. In condition 1, there are n_1 observations of p covariates, while condition 2 has n_2 observations. It is worth noting that the assumption of piecewise stationarity only requires stationarity within each condition and not stationarity across conditions. This may be satisfied in many neuroscience applications where experimental factors affecting the brain state are known from the experimental design. For example, in Bloch et al. (2022), conditions 1 and 2 might represent pre- and post-stimulation, respectively and stimulation times are known.

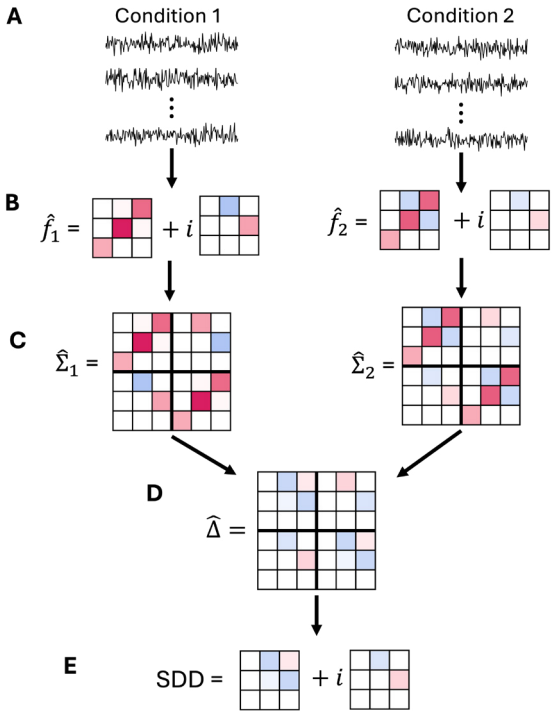


Figure 1: SDD Method. **(A) Inputs:** Time series datasets are observed for two conditions. **(B) Computing the Spectral Densities:** We then compute the spectral densities in each condition using the smoothed periodogram. **(C) Expanding to Real Space:** These spectral densities are expanded to the real space. **(D) Direct Estimation of the Difference in Inverse Spectral Densities:** Using the expanded spectral densities we directly compute the difference estimator. **(E) Output:** We form the difference in inverse spectral densitites by taking the (1,1) and (2,1) blocks of $\hat{\Delta}$.

(B) Computing the Spectral Densities. The spectral density in condition $l \in \{1, 2\}$ at frequency λ, $f_l(\lambda)$, is defined as the Fourier transform of the autocovariance matrix. Intuitively, the spectral density represents the portion of the covariance of the signal that can be attributed to the specific frequency λ (Brockwell and Davis, 1991, pp. 331–332). Mathematically we can write our data as $\{\mathbf{x}_{l,t}\}$ for $t = 1, \ldots, n_l$

where $\mathbf{x}_{l,t}$ is a p-dimensional vector in condition $l \in \{1,2\}$. For each condition l, the autocovariance at lag $h \in \mathbb{Z}$ is denoted as $\Gamma_l(h) = \mathbb{E}\left(\mathbf{x}_{l,t}\mathbf{x}_{l,t+h}^T\right)$. If $\sum_{h=-\infty}^{\infty} |\Gamma_l(h)| < \infty$ then the spectral density in condition l at frequency λ exists and is defined as

$$f_l(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h}\Gamma_l(h), \qquad -\pi \le \lambda \le \pi.$$

To estimate the difference of *inverse* spectral densities, we will require estimates of the spectral densities themselves. We use the smoothed periodogram to estimate the spectral density matrix and refer to it as $\hat{f}_l$. To define the smoothed periodogram, we must first define the periodogram. For condition l, the periodogram at Fourier frequency $\{\lambda_j = 2\pi j/n_l, -\lfloor (n_l-1)/2 \rfloor \le j \le \lfloor n_l/2 \rfloor\}$ is defined as

$$P_l(\lambda_j) = \frac{\left(\sum_{t=1}^{n_l} \mathbf{x}_{l,t} e^{-i\lambda_j t}\right)\left(\sum_{t=1}^{n_l} \mathbf{x}_{l,t} e^{-i\lambda_j t}\right)^H}{2\pi n_l}.$$

Recall $\mathbf{x}^H$ is the conjugate transpose of the vector $\mathbf{x}$. Using the Fast Fourier transform (FFT) algorithm, the periodogram can be computed quickly. The periodogram is known to be an inconsistent estimator of the spectral density so a smoothed version is used where the periodograms of $2M_{n_l}$ nearby frequencies are averaged (Brockwell and Davis, 1991, pp. 347–350). M_{n_l} is referred to as the bandwidth or smoothing span. That is, the smoothed periodogram at fourier frequency λ_j is generated as

$$\hat{f}_l(\lambda_j) = \frac{1}{2M_{n_l}+1} \sum_{k=j-M_{n_l}}^{j+M_{n_l}} P_l(\lambda_k).$$

(C) Expanding to the Real Space. In practice, many different λ values are of interest, but for notational simplicity we assume λ is fixed throughout and suppress the dependence on λ. The spectral density, f_l, is complex-valued in general which can complicate the estimation. To transform this problem to the real space, we note that any complex matrix can be written as the sum of a real matrix and i times another real matrix (for $i^2 = -1$). Specifically we write the spectral density and its inverse as $f_l = A_l + iB_l$, and $f_l^{-1} = \tilde{A}_l + i\tilde{B}_l$, respectively, where A_l, B_l, $\tilde{A}_l$, and $\tilde{B}_l$ are real matrices. With these representations, we can now work in the real space by studying $\Sigma_l = \begin{bmatrix} A_l & -B_l \\ B_l & A_l \end{bmatrix} \in \mathbb{R}^{2p\times 2p}$ instead of f_l. By Lemma A.1 of Fiecas et al. (2019),

$$\begin{bmatrix} A_l & -B_l \\ B_l & A_l \end{bmatrix}\begin{bmatrix} \tilde{A}_l & -\tilde{B}_l \\ \tilde{B}_l & \tilde{A}_l \end{bmatrix} = \Sigma_l\Sigma_l^{-1} = I_{2p},$$

where I_{2p} represents the $2p \times 2p$ identity matrix. In this representation, Σ_l^{-1} is the expansion of f_l^{-1} to the real space. Moreover, f_l and f_l^{-1} can easily be recovered by taking the (1,1) and (2,1) blocks of Σ_l and Σ_l^{-1}, respectively.

(D) Direct Estimation of the Difference in Inverse Spectral Densities. Having defined the problem in the real space, we use the ℓ_1 penalized D-trace loss function (Yuan et al., 2017) to directly estimate $\Delta = \Sigma_1^{-1} - \Sigma_2^{-1}$. To justify this estimation strategy, in Appendix A.1 we show that this loss is convex and the population loss is minimized at the true $\Delta = \Sigma_1^{-1} - \Sigma_2^{-1}$. The inputs to this estimator are $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ which are formed by expanding $\hat{f}_1$ and $\hat{f}_2$ to the real space respectively. Our estimator is

$$\begin{aligned}\hat{\Delta} = \arg\min_{\Delta}\ & \frac{1}{4}\left(\langle \hat{\Sigma}_2\Delta, \Delta\hat{\Sigma}_1\rangle + \langle \hat{\Sigma}_1\Delta, \Delta, \hat{\Sigma}_2\rangle\right) - \\ & \langle \Delta, \hat{\Sigma}_2 - \hat{\Sigma}_1\rangle + \tau_{n_1,n_2}\|\Delta\|_1\end{aligned} \tag{1}$$

where τ_{n_1,n_2} is a penalty parameter depending on sample sizes n_1, n_2 and $\|\Delta\|_1$ is the ℓ_1 penalty which returns the sum of the absolute values of all entries of a matrix. We solve for $\hat{\Delta}$ using the alternating direction method of multipliers (ADMM) algorithm from Yuan et al. (2017). Since we are inputting matrices $\hat{\Sigma}_i \in \mathbb{R}^{2p\times 2p}$, the computational complexity of each iteration of this algorithm is $O(8p^3)$ while the memory requirements are $O(4p^2)$.

(E) Output. Recall that the inverse spectral density, f_l^{-1}, represents the coherence between two nodes after removing the linear effects of all other nodes. Our goal is to estimate how this effective connectivity differs between the two conditions. That is, we wish to estimate $f_1^{-1} - f_2^{-1}$. Using our representation in the real space, the true difference $f_1^{-1} - f_2^{-1}$ is recovered by taking the (1,1) and (2,1) blocks of

$$\Sigma_1^{-1} - \Sigma_2^{-1} = \begin{bmatrix} \tilde{A}_1 & -\tilde{B}_1 \\ \tilde{B}_1 & \tilde{A}_1 \end{bmatrix} - \begin{bmatrix} \tilde{A}_2 & -\tilde{B}_2 \\ \tilde{B}_2 & \tilde{A}_2 \end{bmatrix} := \Delta.$$

Similarly, using our estimate $\hat{\Delta}$ of Δ we can estimate $f_1^{-1} - f_2^{-1}$ using the (1,1) and (2,1) blocks of $\hat{\Delta}$.

In the following section, we study the consistency of our estimator. Intuitively, if $\hat{f}_l$ estimates the spectral density well, we should be able to estimate Δ well. Our theoretical results show that this is indeed the case and the rate at which the smoothed periodogram $\hat{f}_l$ converges to the true periodogram is also the rate at which $\hat{\Delta}$ converges to Δ. Due to the dependence in the data, the smoothed periodogram and $\hat{\Delta}$ are not $\sqrt{n}$ consistent as one might expect in the i.i.d case. Rather, we show that for an optimal choice of smoothing span, $\hat{f}_l$ and $\hat{\Delta}$ converge to f_l and Δ respectively at a rate of $n^{-1/3}$.

3 THEORY

In this section, we study the theoretical properties of our SDD estimator, $\hat{\Delta}$. Specifically, we establish the convergence rate of $\hat{\Delta}$ to the population quantity Δ. From Equation 1, our estimator $\hat{\Delta}$ relies on expanded estimates of the spectral density $\hat{\Sigma}_1, \hat{\Sigma}_2$. Thus, it is natural that the rate of convergence of $\hat{\Delta}$ to Δ relies on the rates of convergence of $\hat{f}_1$ and $\hat{f}_2$ to f_1 and f_2.

Theorem 1. *Suppose Assumption 1 in Appendix B holds and that* $\tau_{n_1,n_2} \geq O(\min(n_1, n_2)^{-1/3})$ *and* $M_{n_l} = O(n_l^{2/3})$. *For* n_1 *and* n_2 *large enough and for any* $H > 0$, *there exists some* $C > 0$ *such that with probability greater than* $1 - Cp^2 \min(n_1, n_2)^{-H}$,

$$\|\hat{\Delta} - \Delta\|_F \leq O\left(s_\Delta \min(n_1, n_2)^{-1/3}\right),$$

where s_Δ *is the number of non-zero elements in the true difference matrix* Δ.

The full proof of Theorem 1 is given in Appendix C. To establish convergence rates of $\hat{f}_1$ and $\hat{f}_2$ in high-dimensions, we rely in part on Theorem 3.1 of Fiecas et al. (2019), which uses the functional dependency framework of Wu (2005) and Wu and Zaffaroni (2018). Specifically, our convergence rates hinge on the assumption of a geometrically decaying dependence in the data. An overview of the functional dependence framework is given in Appendix B and a formal statement of the geometric decay assumption is given in Assumption 1 in Appendix B. The geometric decay assumption is mild and is satisfied by many processes including ARMA, ARMA-ARCH, and ARMA-GARCH processes, as well as other nonlinear autoregressive processes (Shao and Wu, 2007; Liu and Wu, 2010).

While our method uses a similar setup to Yuan et al. (2017), the analysis of our estimator is fundamentally different in two ways. First, we establish Theorem 1 only assuming sparsity of the difference and we explicitly avoid the irrepresentability assumption of Yuan et al. (2017) which has been shown not to hold in practice (Zhao and Yu, 2006). For example, Wang et al. (2021) show through simulations that for graphs of moderate or larger dimensions, this condition almost never holds. This is true even for very sparse differences and weak data dependence. Second, in contrast to the proof of Yuan et al. (2017), which assumes i.i.d. observations, our proofs also leverage advances in theoretical analysis of high-dimensional time series to handle the temporal dependence among observations over time. It is also worth noting that Theorem 1 also applies to differences in inverse spectral densities across frequencies.

We now discuss interesting intermediate results from the proof. In Lemma 1, we show that, with high probability and a smoothing span of $M_{n_l} = O(n_l^{2/3})$, the deviations of the smoothed periodogram from the true spectral density in condition l are $O(n_l^{-1/3})$. Lemma 2 shows that with high probability, the estimated Hessian of the D-trace loss converges to the true Hessian at a rate of $O(\min(n_1, n_2)^{1/3})$. That is, $\|\hat{\Gamma} - \Gamma\|_\infty = O(\min(n_1, n_2)^{1/3})$ where $\hat{\Gamma} = 0.5\left(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \hat{\Sigma}_2 \otimes \hat{\Sigma}_1\right)$ and Γ is the same but without the hats. Lastly, Equation 6 establishes that, with high probability, the maximum norm of deviations of the derivative of the D-trace loss, $\left\|0.5\left(\hat{\Sigma}_1\Delta\hat{\Sigma}_2 + \hat{\Sigma}_2\Delta\hat{\Sigma}_1\right) - \left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)\right\|_\infty$, is $O(\min(n_1, n_2)^{-1/3})$.

Due to the dependence in the data, the convergence rate in Theorem 1 is slower than the $n^{-1/2}$ rates for i.i.d data. This is because, compared with parametric models that amount to fast decaying dependence among observations, our dependence assumption is very mild and general. In fact, as we choose M_n to minimize the deviations in Theorem 3.1 of Fiecas et al. (2019), our rate is the best we can expect to achieve using this general dependence framework when only assuming a geometrically decaying time dependence.

4 SIMULATIONS

In this section, we evaluate the performance of the proposed estimator $\hat{\Delta}$ using simulation studies. The code to reproduce these results as well as those in Section 5 are available at https://github.com/mikehellstern/spectral-differential-networks. All simulations use VAR(1) processes, with $p = 54$, as this allows us to compute the true spectral density using results from Sun et al. (2018). We control the sparsity of the difference in inverse spectral densities by using a block diagonal structure. In all settings, the transition matrix in condition 1 is the same as that in condition 2 except the last 3×3 block is multiplied by -1. The first setting uses similar coefficients as in Sun et al. (2018). In the second and third simulation settings, the transition matrix consists of one 51×51 block and one 3×3 block. The larger block was generated with 60% and 95% sparsity in the second and third settings, respectively. All simulations were performed using $n = 100, 200, 500, 1000, 2000$ observations for both conditions. Smoothed periodograms were computed from data for each condition using a smoothing window of $M_n = \lceil n^{2/3} \rceil$ and then converted to the real space to generate $\hat{\Sigma}_l$. All computations were performed on a MacBook with a 2.3 GHz Quad-Core Intel Core i7 processor and 16 GB of memory. More information on the simulation setup, optimization techniques, and tuning parameter selection can be

found in Appendix D.

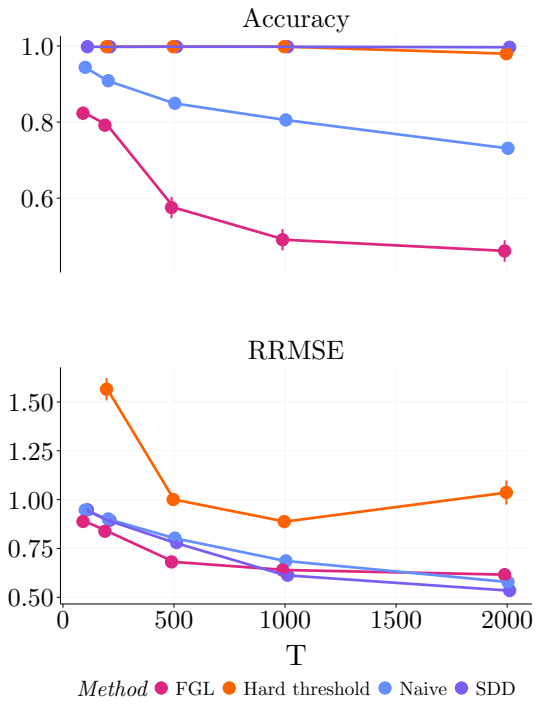


Figure 2: Simulation 1. Results are reported as mean (dots) and SE (vertical lines) where the mean and SE are taken across all frequencies for a given sample size n.

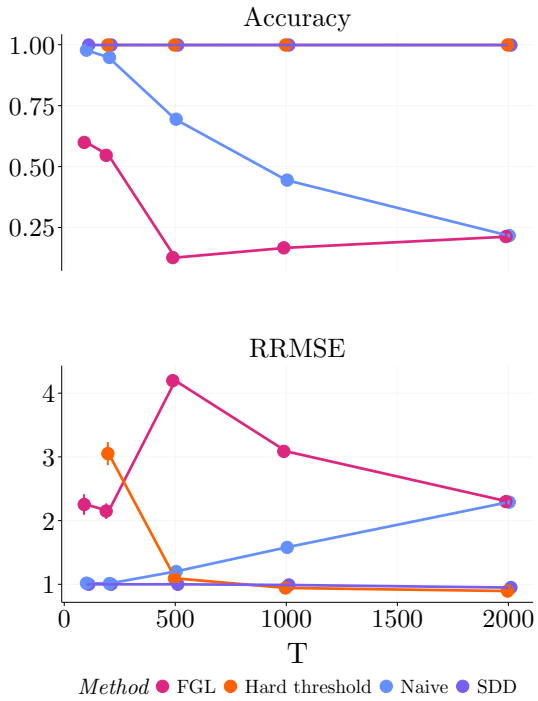


Figure 3: Simulation 2. Results are reported as mean (dots) and SE (vertical lines) where the mean and SE are taken across all frequencies for a given sample size n.

As previously noted, our method is the first to directly estimate the difference in inverse spectral densities and no direct competitors or state-of-the-art methods exist. However, in this section we will discuss and compare several alternative methods. While not focused on directly estimating the difference, a related method is the joint graphical lasso with a fusion penalty (FGL, Danaher et al., 2014). This method aims to simultaneously estimate inverses in each condition that are believed to share structural similarities. To encourage similarities in the inverse estimates, FGL uses an ℓ_1 penalty on their difference. In simulations, FGL was tuned using AIC from Danaher et al. (2014). We have also conceptualized two additional methods to compare to SDD. The first is a Naïve method, which estimates Δ by taking the difference after estimating individual inverse spectral densities separately, $\hat{\Delta}_{\text{N}} = \hat{\Sigma}_1^{-1} - \hat{\Sigma}_2^{-1}$. For the Naïve method, the graphical LASSO (GLASSO, Friedman et al., 2008) was used to estimate a sparse inverse spectral density in each condition. Specifically we used the fast implementation from Sustik and Calderhead (2012) available in the `glassoFast` package in `R` (GPL ($\geq$ 3.0)). To induce sparsity in individual estimates and (potentially) their difference, GLASSO was tuned for each condition separately using eBIC with $\gamma = 0.5$ (Foygel and Drton, 2010). The next method directly thresholds the difference in inverse spectral densities to induce sparse estimates. Specifically, we estimate the inverse spectral density in each condition and take the difference and then use Hard thresholding. The thresholds were tuned using the same eBIC as SDD given in (7). Similar to Deb et al. (2024), all methods are compared based on the following metrics: Accuracy, Precision, Recall, and Relative Root Mean Square Error (RRMSE). The full definitions can be found in Appendix D.

Simulation results for Accuracy and RRMSE in simulations 1 and 2 are reported in Figures 2 and 3. Results for Precision and Recall for simulations 1 and 2 can be found in Appendix D Figures 8 and 9. Results from simulation 3 are in Appendix D, Figures 10 and 11. The full set of results for each of the simulations are in Appendix D Tables 1, 2, 3. In general, SDD has better accuracy than both FGL and the Naïve method with the difference in accuracy increasing as the sam-

ple size increases. This is likely due to the fact that, in general, SDD estimates much fewer edges than either method. This also results in fewer false positive edges but more false negatives, resulting in better precision for SDD at the expense of reduced recall; see Table 1 in Appendix D. Across all settings and sample sizes we also see that SDD generally has either lower or similar RRMSE compared to other methods. Overall, compared to FGL and the Naïve method, SDD identifies a similar number of true edges and achieves similar or lower RRMSE using a much smaller edge set. Compared to Hard thresholding, SDD achieves much better RRMSE and recall across all settings and sample sizes except setting 2 for large n where Hard thresholding slightly outperforms SDD in terms of RRMSE. To further support our analysis, we have also summarized the difference between SDD and each of the competing methods for each metric in Appendix D, Tables 4, 5, 6, 7, 8, 9, 10, 11, 12.

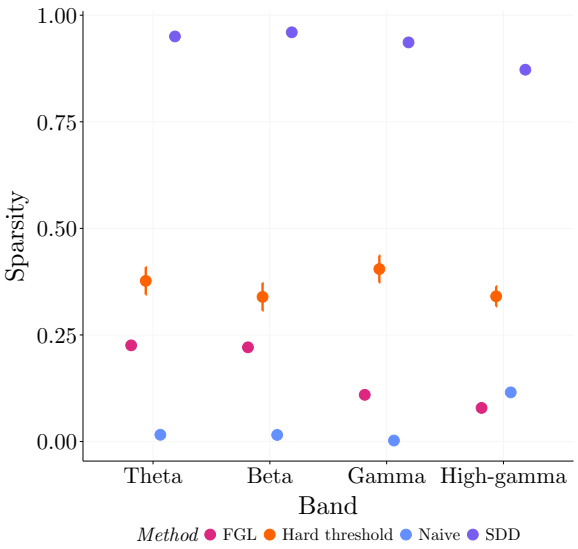


Figure 4: EEG Within Session Analysis. Vertical lines indicate SE.

5 APPLICATION TO EEG DATA

We apply our direct difference estimator (SSD) and competing estimators to electroencephalograms (EEG) data from Hatlestad-Hall et al. (2022). Data were recorded with a 64 channel EEG array for 111 healthy subjects at a sampling frequency of 1024 Hz. Four minutes of brain activity was recorded while subjects were resting with their eyes closed. For 42 subjects, a second session was recorded 2–3 months after the initial session. For our analysis, we used the pre-cleaned data provided in OpenNeuro Dataset ds003775 (Markiewicz et al., 2021). Specific cleaning steps can be found in Hatlestad-Hall et al. (2022). We also downsampled the data to 512 Hz.

To validate our method, we analyzed the subjects with follow-up and without follow-up separately. For those with follow-up, we estimated the difference in networks from the first to the second session (*across session analysis*). For those without, we estimated the difference in networks from 0-60s to 120-180s (*within session analysis*). The 60-120s block was used as a rest. It has been shown that brain networks are temporally dynamic (Zalesky et al., 2014; Nobukawa et al., 2019). Therefore, *a priori*, we expect the differential networks to be sparser in the within session analysis compared to the across session analysis and finding such results would provide validation of our method.

We compare the sparsity of the estimated differential network for SDD, FGL, the Naïve, and Hard threshold methods across the commonly-used Theta, Beta, Gamma and High-Gamma bands. Sparsity was defined as the proportion of entries in the estimated difference that were non-zero. Five or six evenly spaced frequencies were considered for each band. They were (in Hz): Theta: (4,5,6,7,8), Beta: (12, 16, 20, 24, 28), Gamma: (30, 40, 50, 60, 70), High-gamma: (80, 95, 110, 125, 140, 150). Results averaged over subjects and frequencies within each band are presented in Figures 4 and 5.

Figures 4 and 5 shows that SDD is the only method that estimates a sparser difference in the within session analysis compared to the across session analysis across all frequency bands. On the other hand, competing methods have similar or less sparsity in the within session analysis compared to the across-session analysis. These results clearly indicate that SDD estimates are more consistent with the experimental setting in this application.

6 APPLICATION TO SIMULATED μECoG DATA

We next apply our SDD estimator to simulated micro-electrocorticography (μECoG) data. The simulation design is based on the optogenetic stimulation μECoG experiments from Yazdan-Shahmorad et al. (2016, 2018); Bloch et al. (2022). Studying how stimulation changes the brain network using such experiments can help design protocols to change the brain network in a targeted way. This can lead to treatments for conditions, such as schizophrenia or epilepsy, which are known to stem from abnormal brain connectivity networks (Bloch et al., 2022). Our simulations used a sim-

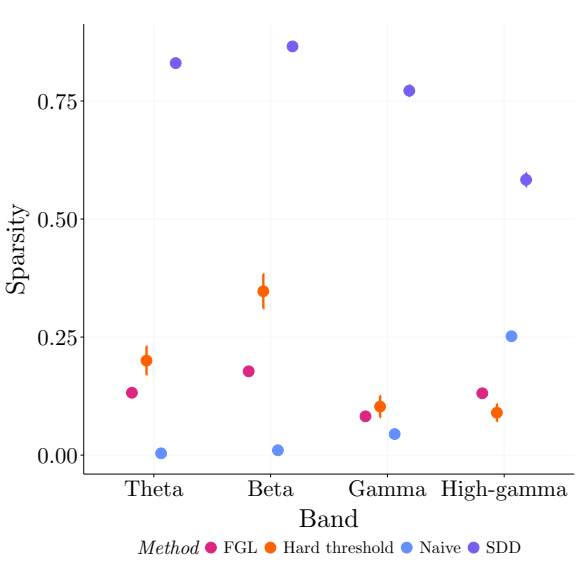


Figure 5: EEG Across Session Analysis. Vertical lines indicate SE.

ilar setup to Yazdan-Shahmorad et al. (2016). Specifically, a total of 32 experimental and 4 control sessions were simulated. Using the same session types as Bloch et al. (2022), of the 32 experimental sessions, 23 were generated using a simulated 10ms stimulation delay and 9 were generated with a 100ms delay. Each experimental session consisted of five 10-minute stimulation blocks (stimulation state) with a 5-minute recording block (resting state) before and after. Thus, each experimental session alternated between resting state and stimulation state blocks consisting of 6 total resting state blocks and 5 total stimulation state blocks. We simulated data using a sampling frequency of 500Hz.

Data in all sessions was simulated using a VAR(1) process with i.i.d standard normal errors. It is worth noting that due to the autoregressive nature of the process, autocovariance matrices between variables will not be diagonal in general despite using a diagonal covariance matrix for the errors (Lütkepohl, 2005). Since the experiments in Yazdan-Shahmorad et al. (2016) included a varying number of good electrodes out of the 96 total electrodes, we randomly selected between $p = 56$ and $p = 94$ good electrodes for each session. Coefficients in a base transition matrix ($A_1 \in \mathbb{R}^{p \times p}$) were generated by randomly sampling from a standard normal distribution. To obtain a stable time series, the matrix A_1 was modified to have maximum eigenvalue < 1. Similar to Yazdan-Shahmorad et al. (2016), two electrodes in each session were randomly chosen to be stimulated. The effect of stimulation was modeled by changing the coefficient in A_1 between simulation electrodes which is denoted as A_{stim}. For control sessions, there was no stimulation so $A_1 = A_{\text{stim}}$. The effect of a 10ms and 100ms stimulation delay was simulated by changing the coefficient between stim sites to be $2\max(|A_1|)$ and $0.5\max(|A_1|)$, respectively. In this way, we simulate a weaker effect of the 100ms stimulation delay (Bloch et al., 2022). It is worth noting that, in contrast to the simulations in Section 4, the transition matrices are dense with no block structure. Thus, it is not immediately clear how the true spectral density changes as the transition matrices change.

During stimulation blocks, data were simulated using the corresponding A_{stim} matrix. To simulate a lasting, but weakened effect of stimulation on the following resting state, resting state blocks were simulated using a progressively increasing effect of stimulation. Specifically, resting state block 1 uses A_1 since no stimulation has occurred. The effect of stimulation for resting state blocks 2 through 6 linearly increases from ≈ 0.05 to ≈ 0.6 times the stimulation effect. For example, for a 10ms delay session, resting state block 2 was simulated using a VAR(1) process with transition matrix equal to A_1 where the coefficient between stimulation sites was replaced with $\approx 0.05 * 2\max(|A_1|)$.

Using our simulated data, we study how stimulation changes the connectivity between stimulation sites. However, instead of using a naïve estimator of the connectivity changes as in Bloch et al. (2022), we use SDD. Using SDD, we estimate the change in inverse spectral densities from resting state to the following resting state (RS to RS) and from resting state to the following stimulation state (RS to SS). For each session there are 5 RS to RS differences (RS1 to RS2, RS2 to RS3, ..., RS5 to RS6) and 5 RS to SS difference (RS1 to SS1, ..., RS5 to SS5). Note that an estimated SDD coefficient of zero indicates no change while non-zero coefficients indicate a change. This could be either a change in magnitude of the connection, a new connection, or a lost connection. The frequencies we study in each band are the same as in Section 5. Results are averaged over all sessions, frequencies within each band, delay (control, 10ms, or 100ms) and RS to RS or SS to SS.

To study how stimulation changes the edges between stimulation sites, we compute the probability of edge change between the stimulation sites for each band, delay, and RS to RS or RS to SS. The RS to RS and RS to SS results are displayed in Figures 6 and 7 respectively. In the RS to SS analysis, stimulation with a 10ms or 100ms delay is more likely to induce a change in connectivity between the stimulation sites across all frequency bands compared to control sessions. After stimulation ends (RS to RS analysis), this connectiv-

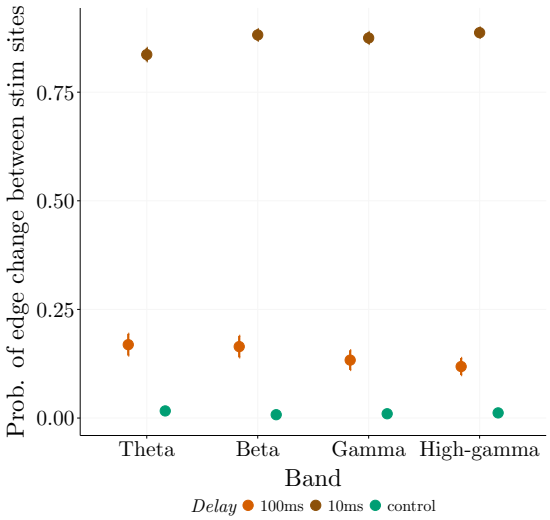


Figure 6: RS to RS analysis. Probability of Change in Edges Between Stimulation Sites by Frequency Band and Delay. Vertical bars indicate standard errors.

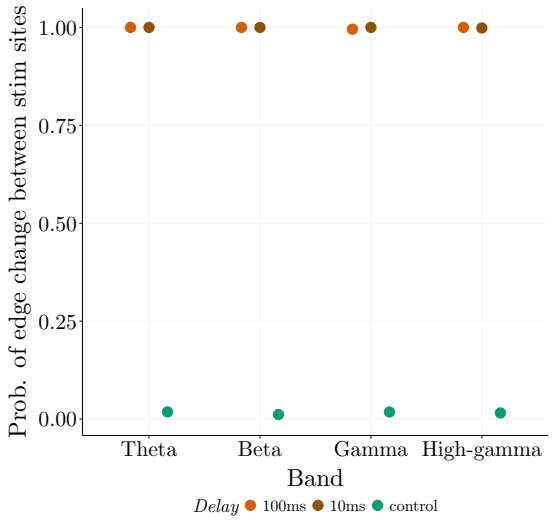


Figure 7: RS to SS analysis. Probability of Change in Edges Between Stimulation Sites by Frequency Band and Delay. Vertical bars indicate standard errors.

ity change is attenuated, but still persists. This is in line with our simulation framework and prior work which showed a change in functional connectivity between stimulation sites in the stimulation state which persisted into the following resting state (Bloch et al., 2022).

7 DISCUSSION

The SDD estimator was compared to the joint graphical lasso with a fusion penalty (FGL, Danaher et al., 2014), the the Naïve method of estimating sparse inverse the spectral densities and taking their difference, and directly hard thresholding the difference. The four methods were compared across three different simulation settings, where the true difference in inverse spectral densities was sparse. Comparison metrics included precision, recall, accuracy, and relative root mean square error (RRMSE). When compared to the other methods, SDD performed better in accurately identifying edges and non-edges and the SDD estimates had a lower relative root mean square error. Using EEG data we further validated SDD by showing it estimated sparser differences between networks that were closer in time, in line with expected behavior.

While consistency of SDD is established in Theorem 1, this result does not quantify the degree of uncertainty around these estimates. A particularly interesting future direction is developing methods for uncertainty quantification, such as confidence intervals and hypothesis testing. These may be important when the relationship between two brain regions or nodes in the graph is of primary interest or when the sample size is small and noise may be large relative to the signal.

8 CONCLUSION

We proposed a direct estimate of the differences in inverse spectral densities between two conditions, termed the Spectral D-trace Difference (SDD) estimator. By using the direct difference estimator, we only need to assume sparsity in the difference of inverse spectral densities. Compared to the usual assumptions of sparsity of each inverse spectral density, the sparsity of the difference is more realistic if, for example, the inverse spectral densities do not change much between conditions. This is indeed what we expect in many biological settings, especially neurodegenerative disorders that are associated with changes in brain connectivity. Convergence rates of our estimator to the true difference were derived using only an assumption on decaying dependence within the time series.

Acknowledgements

This work was partially supported by U.S. NIH grants R01-MH125429 and R01-GM133848.

References

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse

maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

Julien Bloch, Alexander Greaves-Tunnell, Eric Shea-Brown, Zaid Harchaoui, Ali Shojaie, and Azadeh Yazdan-Shahmorad. Network structure mediates functional reorganization induced by optogenetic stimulation of non-human primate sensorimotor cortex. *Iscience*, 25(5), 2022.

Hilmar Böhm and Rainer von Sachs. Shrinkage estimation in the frequency domain of multivariate time series. *Journal of Multivariate Analysis*, 100(5):913–935, 2009.

Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer science & business media, 1991.

Rainer Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51:157–172, 2000.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(2):373–397, 2014.

Navonil Deb, Amy Kuceyeski, and Sumanta Basu. Regularized estimation of sparse spectral precision matrices. *arXiv preprint arXiv:2401.11128*, 2024.

Tom Dupré la Tour, Thomas Moreau, Mainak Jas, and Alexandre Gramfort. Multivariate convolutional sparse coding for electromagnetic brain signals. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Mark Fiecas, Chenlei Leng, Weidong Liu, and Yi Yu. Spectral analysis of high-dimensional time series. 2019.

Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. *Advances in neural information processing systems*, 23, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Karl J Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

Théo Gnassounou, Rémi Flamary, and Alexandre Gramfort. Convolution monge mapping normalization for learning on sleep data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10457–10476. Curran Associates, Inc., 2023.

Christoffer Hatlestad-Hall, Trine Waage Rygvold, and Stein Andersson. "srm resting-state eeg", 2022.

SR James, HA Knox, RE Abbott, and EJ Screaton. Improved moving window cross-spectral analysis for resolving large temporal seismic velocity changes in permafrost. *Geophysical Research Letters*, 44(9): 4018–4026, 2017.

Lucas C Laurindo, Leo Siqueira, Arthur J Mariano, and Ben P Kirtman. Cross-spectral analysis of the sst/10-m wind speed coupling resolved by satellite products and climate model simulations. *Climate dynamics*, 52(9):5071–5098, 2019.

Weidong Liu and Wei Biao Wu. Asymptotics of spectral density estimates. *Econometric Theory*, 26(4): 1218–1245, 2010.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncavles, et al. The openneuro resource for sharing of neuroscience data. *Elife*, 10:e71774, 2021.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. 2012.

Sou Nobukawa, Mitsuru Kikuchi, and Tetsuya Takahashi. Changes in functional connectivity dynamics with aging: a dynamical phase synchronization approach. *Neuroimage*, 188:357–368, 2019.

Hugo Richard, Luigi Gresele, Aapo Hyvarinen, Bertrand Thirion, Alexandre Gramfort, and Pierre Ablin. Modeling shared responses in neuroimaging studies through multiview ica. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19149–19162. Curran Associates, Inc., 2020.

Xiaofeng Shao and Wei Biao Wu. Asymptotic spectral theory for nonlinear time series. 2007.

Ali Shojaie. Differential network analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(2):e1508, 2021.

Yiming Sun, Yige Li, Amy Kuceyeski, and Sumanta Basu. Large spectral density matrix estimation by thresholding. *arXiv preprint arXiv:1812.00532*, 2018.

Mátyás A Sustik and Ben Calderhead. Glassofast: an efficient glasso implementation. *UTCS Technical Report TR-12-29 2012*, 2012.

Katherine Tsai, Oluwasanmi Koyejo, and Mladen Kolar. Joint gaussian graphical model estimation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(6):e1582, 2022.

Mauro Ursino, Giulia Ricci, and Elisa Magosso. Transfer entropy as a measure of brain connectivity: A critical analysis with the help of neural mass models. *Frontiers in computational neuroscience*, 14:45, 2020.

Yue Wang, Jing Ma, and Ali Shojaie. Direct estimation of differential granger causality between two high-dimensional time series. *arXiv preprint arXiv:2109.07609*, 2021.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005.

Wei Biao Wu and Paolo Zaffaroni. Asymptotic theory for spectral density estimates of general multivariate time series. *Econometric Theory*, 34(1):1–22, 2018.

Azadeh Yazdan-Shahmorad, Camilo Diaz-Botia, Timothy L Hanson, Viktor Kharazia, Peter Ledochowitsch, Michel M Maharbiz, and Philip N Sabes. A large-scale interface for optogenetic stimulation and recording in nonhuman primates. *Neuron*, 89(5): 927–939, 2016.

Azadeh Yazdan-Shahmorad, Daniel B Silversmith, Viktor Kharazia, and Philip N Sabes. Targeted cortical reorganization using optogenetics in non-human primates. *Elife*, 7:e31034, 2018.

Huili Yuan, Ruibin Xi, Chong Chen, and Minghua Deng. Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4):755–770, 2017.

Andrew Zalesky, Alex Fornito, Luca Cocchi, Leonardo L Gollo, and Michael Breakspear. Time-resolved resting-state brain networks. *Proceedings of the National Academy of Sciences*, 111 (28):10341–10346, 2014.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7 (Nov):2541–2563, 2006.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Not Applicable. No new models released.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Not Applicable. No new models released.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] Not Applicable. No new models released.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Yes, see Section 3 as well as Appendix C and Appendix B.
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Yes, see Appendix C.
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes, see Section 3 and Appendix B.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] Yes, all code to reproduce experimental and application results are provided in the supplemental material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Yes, training details are contained in the paper and in the supplemental material.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] Yes, these are included in the main paper as well as the supplemental material.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Yes computing infrastructure is discussed in Section 4.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] Yes, existing assets are cited in the paper as well as in Appendix D.

(b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Yes, licenses are included in the paper as well as in Appendix D.

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] Yes, new assests are available in the supplemental material.

(d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Not Applicable.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable.

Supplementary Materials

Contents

A ADDITIONAL DETAILS FOR METHODS **12**

A.1 D-trace Loss 12

B FUNCTIONAL DEPENDENCE FRAMEWORK **13**

C THEORETICAL RESULTS **13**

D SIMULATION SPECIFICS AND RESULTS **17**

A ADDITIONAL DETAILS FOR METHODS

A.1 D-trace Loss

To estimate Δ, we use the ℓ_1 penalized D-trace loss function from Yuan et al. (2017):

$$L_D\left(\Delta, \Sigma_2, \Sigma_1\right) = \frac{1}{4}\left(\langle \Sigma_2\Delta, \Delta\Sigma_1\rangle + \langle \Sigma_1\Delta, \Delta\Sigma_2\rangle\right) - \langle \Delta, \Sigma_2 - \Sigma_1\rangle$$

where $\langle A, B\rangle = \mathrm{Tr}(AB^T)$. Similar to Yuan et al. (2017), we can take the derivative with respect to Δ to see that $\Sigma_1^{-1} - \Sigma_2^{-1}$ minimizes the D-trace loss. More specifically, let

$$\frac{\partial L_D}{\partial \Delta} = \left(\Sigma_2\Delta\Sigma_1 + \Sigma_1\Delta\Sigma_2\right)/2 - \left(\Sigma_2 - \Sigma_1\right) . \tag{2}$$

Evaluating $\frac{\partial L_D}{\partial \Delta}$ at $\Delta = \Sigma_1^{-1} - \Sigma_2^{-1}$ yields 0, establishing that $\Delta = \Sigma_1^{-1} - \Sigma_2^{-1}$ minimizes L_D. Furthermore, the Hessian of L_D with respect to Δ is $\frac{\partial^2 L_D}{\partial \Delta^2} = (\Sigma_1 \otimes \Sigma_2 + \Sigma_2 \otimes \Sigma_1)/2$ where $\otimes$ is the kronecker product. Assuming both Σ_1 and Σ_2 are positive definite, L_D is convex in Δ and $\Delta = \Sigma_1^{-1} - \Sigma_2^{-1}$ is the unique minimizer.

In practice, the population quantites Σ_1 and Σ_2 are not available. Instead, we use the estimates $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$. We can additionally incorporate an ℓ_1 penalty to estimate a sparse Δ which gives us our estimator in (1).

B FUNCTIONAL DEPENDENCE FRAMEWORK

Let ϵ_0^* and $\{\epsilon_t\}_{t\in\mathbb{Z}}$ be i.i.d. vectors in $\mathbb{R}^b$ and $\mathcal{F}_t = (\ldots, \epsilon_{t-1}, \epsilon_t)$. We further let $\mathbf{x}_{l,t} = (x_{l,1t}, \ldots x_{l,pt})^T$ be a p-dimensional process in condition l where

$$x_{l,jt} = R_{l,j}(\mathcal{F}_t).$$

Note that l indexes the condition, in this case 1 or 2, and j indexes the variable, $1, \ldots, p$. To measure the dependence of $\mathbf{x}_{l,t}$ on ϵ_0 we can replace ϵ_0 in $\mathcal{F}_t$ with an i.i.d. copy ϵ_0^*. This gives $\mathcal{F}_t' = (\ldots, \epsilon_{-1}, \epsilon_0^*, \epsilon_1, \ldots, \epsilon_t)$ and

$$x_{l,jt}' = R_{l,j}(\mathcal{F}_t').$$

The dependence on ϵ_0 can then be measured by

$$\theta_{l,jt} = \left(\mathbb{E}\left|x_{l,jt} - x_{l,jt}'\right|^2\right)^{1/2}$$

where $\mathbb{E}\left|x\right|^2$ is the expected value of $|x|^2$. Controlling this dependence measure is essential for establishing the convergence rates of $\hat{f}_l$. The needed conditions on $\theta_{l,jt}$ are laid out in Assumption 1.

Assumption 1. *Assume for some constant* $0 < \rho < 1$,

$$\max_{l=\{1,2\}} \max_{j=1,\ldots,p} \theta_{l,jt} = O(\rho^t),$$

and for some constant $\kappa > 0$, $C_0 > 0$,

$$\max_{l=\{1,2\}} \max_{j=1,\ldots,p} \mathbb{E}\left(\exp\left(\kappa|\mathbf{x}_{l,j0}|\right)\right) \leq C_0.$$

Assumption 1 is essentially a restatement of Assumption 1 in Fiecas et al. (2019), except that we slightly generalize it to account for the fact that we have data from two separate conditions. This condition requires that the dependence on ϵ_0 decreases geometrically as time increases. A similar condition is used in Wu and Zaffaroni (2018) and Liu and Wu (2010) and essentially requires that the dependence in the data on prior time points cannot be too strong. As noted in Shao and Wu (2007) and examples 1 and 2 of Liu and Wu (2010), many processes satisfy this constraint, including ARMA, ARMA-ARCH, and ARMA-GARCH processes as well as other nonlinear autoregressive processes.

Assumption 1 is satisfied if for example Assumption 1 from Fiecas et al. (2019) is satisfied for both conditions. For example if $\{\rho_1, \kappa_1, C_{0,1}\}$ and $\{\rho_2, \kappa_2, C_{0,2}\}$ are the parameters that satisfy Assumption 1 of Fiecas et al. (2019) for the data in conditions 1 and 2 respectively, then $\{\max(\rho_1, \rho_2), \min(\kappa_1, \kappa_2), \max(C_{0,1}, C_{0,2})\}$ satisfy our Assumption 1. We are now ready to state the consistency of our estimator, $\hat{\Delta}$.

C THEORETICAL RESULTS

To prove Theorem 1 we will use the framework of Negahban et al. (2012), specifically Corollary 1. This requires establishing that the D-trace loss satisfies the Restricted Strong Convexity (RSC) condition and that the regularizing penalty is decomposable. An essential ingredient to all these results is a concentration bound on $\hat{f}_l$. We begin by establishing this bound.

Lemma 1 (Concentration inequality on $\hat{f}_l$)**.** *Under Assumption 1, for any* $H > 0$, *and for an optimal bandwidth choice of* $M_{n_l}^* = O(n_l^{2/3})$, *there exists a constant* $C_1 > 0$ *that depends only on* κ *and* C_0 *such that*

$$\mathbb{P}\left(\sup_{\{\lambda_j = 2\pi j/n_l, \lfloor (n_l-1)/2 \rfloor \leq j \leq \lfloor n_l/2 \rfloor\}} \|\hat{f}_l(\lambda_j) - f_l(\lambda_j)\|_\infty > C_1 8^{2/3} 2 n_l^{-\frac{1}{3}}\right) \leq C_2 p^2 n_l^{-H}, \tag{3}$$

where $l \in \{1, 2\}$ *indicates condition and the constant* $C_2 > 0$ *is constant in* n_l *but depends on* C_1 *and* H.

Proof. Under Assumption 1 we have that Theorem 3.1 from (Fiecas et al., 2019) holds so that for any $\delta > 0$ and $H > 0$ and for $C_1 > 0$ depending only on κ and C_0

$$\mathbb{P}\left(\sup_{\{\lambda_j = 2\pi j/n_l, \lfloor (n_l-1)/2 \rfloor \leq j \leq \lfloor n_l/2 \rfloor\}} \|\hat{f}_l(\lambda_j) - f_l(\lambda_j)\|_\infty > \frac{C_1 M_{n_l}}{n_l} + \frac{8n_l^\delta}{M_{n_l}^{1/2+\delta}}\right) \leq C_2 p^2 n_l^{-H},$$

where l indexes condition 1 or 2 and C_2 is constant in n_l but depends on C_1 and H. Note that this same inequality holds for both conditions $l = 1, 2$ because Assumption 1 holds for both conditions. We allow n_l to differ between conditions for full generality. We can then select M_{n_l} to minimize the deviations. That is we select M_{n_l} to minimize $\frac{C_1 M_{n_l}}{n_l} + \frac{8n_l^\delta}{M_{n_l}^{1/2+\delta}}$. The algebra is omitted but setting $\frac{\partial}{\partial M_{n_l}}\left[\frac{C_1 M_{n_l}}{n_l} + \frac{8n_l^\delta}{M_{n_l}^{1/2+\delta}}\right] = 0$ and solving we get $M^*_{n_l} = \left(\frac{8}{C_1}\left(\frac{1}{2}+\delta\right)\right)^{2/(3+2\delta)} n_l^{\frac{2+2\delta}{3+2\delta}}$. It can be shown that the second derivative is positive on $\delta > 0$ so this is indeed a global minimum (on $\delta > 0$). Plugging this back into the equation yields

$$\left.\frac{C_1 M_{n_l}}{n_l} + \frac{8n_l^\delta}{M_{n_l}^{1/2+\delta}}\right|_{M_{n_l}=M^*_{n_l}} = C_1^{\frac{1+2\delta}{3+2\delta}} 8^{\frac{2}{3+2\delta}} \left(\left(\frac{1}{2}+\delta\right)^{\frac{2}{3+2\delta}} + \left(\frac{1}{2}+\delta\right)^{-\frac{1+2\delta}{3+2\delta}}\right) n_l^{-\frac{1}{3+2\delta}}.$$

One can further show that $\left(\left(\frac{1}{2}+\delta\right)^{\frac{2}{3+2\delta}} + \left(\frac{1}{2}+\delta\right)^{-\frac{1+2\delta}{3+2\delta}}\right)$ is maximized at $\delta = 1/2$ with a maximum value of 2 (algebra omitted). Furthermore $C_1^{\frac{1+2\delta}{3+2\delta}} \leq C_1$ and $8^{\frac{2}{3+2\delta}} \leq 8^{2/3}$ for all $\delta > 0$. Therefore we have that for all $\delta > 0$,

$$\frac{C_1 M^*_{n_l}}{n_l} + \frac{8n_l^\delta}{(M^*_{n_l})^{1/2+\delta}} \leq C_1 8^{2/3} 2 n_l^{-\frac{1}{3+2\delta}}.$$

Letting $\delta \to 0$ the RHS of our concentration bound becomes $C_1 8^{2/3} 2 n_l^{-\frac{1}{3}}$ establishing the lemma.

□

Remark 1. *Since* $\|\hat{\Sigma}_l(\lambda_j) - \Sigma_l(\lambda_j)\|_\infty \leq \|\hat{f}_l(\lambda_j) - f_l(\lambda_j)\|_\infty$*, it follows that Lemma 1 also holds for* $\|\hat{\Sigma}_l(\lambda_j) - \Sigma_l(\lambda_j)\|_\infty$.

To establish RSC of the D-trace loss function we will also need a concentration inequality on the second derivative of the D-trace loss which is stated in Lemma 2.

Lemma 2 (Concentration inequality on second derivative of D-trace loss)**.** *Under the conditions from Lemma 1 we have that for* $\min(n_1, n_2) \geq 8^{9/2} C_1^3 / C_f^3$*, with probability* $> 1 - 2C_2 p^2 \min(n_1, n_2)^{-H}$*,*

$$\left\| 0.5\left(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \hat{\Sigma}_2 \otimes \hat{\Sigma}_1\right) - 0.5\left(\Sigma_1 \otimes \Sigma_2 + \Sigma_2 \otimes \Sigma_1\right)\right\|_\infty \leq \frac{C_f C_1 8^{5/2}}{\min(n_1, n_2)^{1/3}},$$

where $C_f = \max(\|f_1\|_\infty, \|f_2\|_\infty)$ *and* C_1 *is the same as in Lemma 1.*

Proof. Begin by noting that both

$$\|\hat{\Sigma}_2 \otimes \hat{\Sigma}_1 - \Sigma_2 \otimes \Sigma_1\|_\infty, \|\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 - \Sigma_1 \otimes \Sigma_2\|_\infty \leq \|\Sigma_1\|_\infty \|\hat{\Sigma}_2 - \Sigma_2\|_\infty + \|\Sigma_2\|_\infty \|\hat{\Sigma}_1 - \Sigma_1\|_\infty + \|\hat{\Sigma}_1 - \Sigma_1\|_\infty \|\hat{\Sigma}_2 - \Sigma_2\|_\infty.$$

Also we have that $\|\Sigma_l\|_\infty \leq \|f_l\|_\infty$. Using Lemma 1, we get that with probability $> 1 - 2C_2 p^2 \min(n_1, n_2)^{-H}$,

$$\begin{aligned}\|\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 - \Sigma_1 \otimes \Sigma_2\|_\infty &\leq C_f\left(C_1 8^{3/2} 2 n_1^{-1/3} + C_1 8^{3/2} 2 n_2^{-1/3}\right) + C_1^2 8^{6/2} 4 n_1^{-1/3} n_2^{-1/3} \\ &\leq C_f C_1 8^{3/2} 4 \min(n_1, n_2)^{-1/3} + C_1^2 8^{6/2} 4 \min(n_1, n_2)^{-2/3}.\end{aligned}$$

We can see that for $\min(n_1, n_2) \geq 8^{9/2} C_1^3 / C_f^3$ the 2nd term on the RHS is smaller than the 1st term on the RHS so we can combine them to simplify the bound to be 2 times the first term on the RHS or

$$\|\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 - \Sigma_1 \otimes \Sigma_2\|_\infty \leq \frac{C_f C_1 8^{5/2}}{\min(n_1, n_2)^{1/3}}.$$

This also holds for $\|\hat{\Sigma}_2 \otimes \hat{\Sigma}_1 - \Sigma_2 \otimes \Sigma_1\|_\infty$. Defining

$$\text{Event } \mathcal{A} := \left\{\|\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 - \Sigma_1 \otimes \Sigma_2\|_\infty \leq \frac{C_f C_1 8^{5/2}}{\min(n_1, n_2)^{1/3}}\right\} \cap \left\{\|\hat{\Sigma}_2 \otimes \hat{\Sigma}_1 - \Sigma_2 \otimes \Sigma_1\|_\infty \leq \frac{C_f C_1 8^{5/2}}{\min(n_1, n_2)^{1/3}}\right\}$$

then from the above results, for $\min(n_1, n_2) \geq 8^{9/2} C_1^3 / C_f^3$, Event $\mathcal{A}$ occurs with probability $> 1 - 2C_2 p^2 \min(n_1, n_2)^{-H}$ and we have

$$\left\| 0.5 \left(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \hat{\Sigma}_2 \otimes \hat{\Sigma}_1\right) - 0.5 \left(\Sigma_1 \otimes \Sigma_2 + \Sigma_2 \otimes \Sigma_1\right) \right\|_\infty \leq \frac{C_f C_1 8^{5/2}}{\min(n_1, n_2)^{1/3}}\,. \tag{4}$$

□

Using Lemma 1 and Lemma 2 we can establish that the D-trace loss satisfies the restricted strong convexity condition. This is formalized in Lemma 3

Lemma 3 (RSC of D-trace loss)**.** *Under the conditions of Lemma 1, for* $\min(n_1, n_2) \geq \max\left(8^{9/2} C_1^3 / C_f^3, \left(\frac{32 s_\Delta C_f C_1 8^{5/2}}{\lambda_{\min}(f_1)\lambda_{\min}(f_2)}\right)^3\right)$ *with probability* $> 1 - 2C_2 p^2 \min(n_1, n_2)^{-H}$ *RSC holds for the D-trace loss with* $\kappa_L = \lambda_{\min}(f_1)\lambda_{\min}(f_2)/2$*. We use* $\lambda_{\min}(A)$ *to denote the minimum eigenvalue of a matrix* A *and* s_Δ *to denote the number of non-zero entries of* Δ*. The constant* C_f *is defined in Lemma 2 while* C_1, C_2 *are defined in Lemma 1.*

Proof. To show that the D-trace loss function satisfies the RSC condition, we must show that $\mathbf{m}^T \left(\nabla^2 L(\Delta, \hat{\Sigma}_1, \hat{\Sigma}_2)\right) \mathbf{m} > \kappa_L \|\mathbf{m}\|_2^2$ for all $\mathbf{m} \in \mathcal{C}(S_\Delta)$ where $\mathcal{C}(S_\Delta) = \left\{\mathbf{m} \in \mathbb{R}^{4p^2} \middle| \|\mathbf{m}_{S_\Delta^c}\|_1 \leq 3\|\mathbf{m}_{S_\Delta}\|_1\right\}$ for some κ_L. For our loss, we let S_Δ be the support of Δ, the true difference. The notation $\mathbf{m}_{S_\Delta}$ selects the entries of $\mathbf{m}$ that correspond to the non-zero indices of Δ while $\mathbf{m}_{S_\Delta^c}$ selects the entries corresponding to the zero indices of Δ.

First note that for $\mathbf{m} \in \mathcal{C}(S_\Delta)$, $\|\mathbf{m}\|_1 \leq 4\|\mathbf{m}_{S_\Delta}\|_1 \leq 4\sqrt{s_\Delta}\|\mathbf{m}_{S_\Delta}\|_2$. Recall $\nabla^2 L(\Delta, \hat{\Sigma}_1, \hat{\Sigma}_2) = 0.5(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \hat{\Sigma}_2 \otimes \hat{\Sigma}_1)$. Showing RSC proceeds the same as in Wang et al. (2021).

$$\begin{aligned}
\mathbf{m}^T(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2)\mathbf{m} &\geq \mathbf{m}^T(\Sigma_1 \otimes \Sigma_2)\mathbf{m} + \mathbf{m}^T \left(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 - \Sigma_1 \otimes \Sigma_2\right) \mathbf{m} \\
&\geq \lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)\|\mathbf{m}\|_2^2 - \left|\mathbf{m}^T \left(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 - \Sigma_1 \otimes \Sigma_2\right) \mathbf{m}\right| \\
&\geq \lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)\|\mathbf{m}\|_2^2 - \|\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 - \Sigma_1 \otimes \Sigma_2\|_\infty \|\mathbf{m}\|_1^2 \\
&\geq \lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)\|\mathbf{m}\|_2^2 - 16 s_\Delta \|\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 - \Sigma_1 \otimes \Sigma_2\|_\infty \|\mathbf{m}\|_2^2\,.
\end{aligned} \tag{5}$$

The last inequality follows from the fact that $\|\mathbf{m}\|_1 = \|\mathbf{m}_{S_\Delta^c}\|_1 + \|\mathbf{m}_{S_\Delta}\|_1 \leq 4\|\mathbf{m}_{S_\Delta}\|_1 \leq 4\sqrt{s_\Delta}\|\mathbf{m}_{S_\Delta}\|_2 \leq 4\sqrt{s_\Delta}\|\mathbf{m}\|_2$. Similarly

$$\mathbf{m}^T(\hat{\Sigma}_2 \otimes \hat{\Sigma}_1)\mathbf{m} \geq \lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)\|\mathbf{m}\|_2^2 - 16 s_\Delta \|\hat{\Sigma}_2 \otimes \hat{\Sigma}_1 - \Sigma_2 \otimes \Sigma_1\|_\infty \|\mathbf{m}\|_2^2\,.$$

Using Lemma 2 and the fact that $\lambda_{\min}(\Sigma_l) = \lambda_{\min}(f_l)$ we have that, for $\min(n_1, n_2) \geq 8^{9/2} C_1^3 / C_f^3$ and with probability $> 1 - 2C_2 p^2 \min(n_1, n_2)^{-H}$,

$$\mathbf{m}^T 0.5(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \hat{\Sigma}_2 \otimes \hat{\Sigma}_1)\mathbf{m} \geq \lambda_{\min}(f_1)\lambda_{\min}(f_2)\|\mathbf{m}\|_2^2 - 16 s_\Delta \frac{C_f C_1 8^{5/2}}{\min(n_1, n_2)^{1/3}} \|\mathbf{m}\|_2^2$$

So for $\min(n_1, n_2) \geq \left(\frac{32 s_\Delta C_f C_1 8^{5/2}}{\lambda_{\min}(f_1)\lambda_{\min}(f_2)}\right)^3$ then the 2nd term on the RHS is ≤ 0.5*(1st term on RHS). Thus we get

$$\mathbf{m}^T 0.5(\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \hat{\Sigma}_2 \otimes \hat{\Sigma}_1)\mathbf{m} \geq \frac{\lambda_{\min}(f_1)\lambda_{\min}(f_2)}{2}\|\mathbf{m}\|_2^2$$

Therefore we conclude that for $\min(n_1, n_2) \geq \max\left(8^{9/2}C_1^3/C_f^3, \left(\frac{32 s_\Delta C_f C_1 8^{5/2}}{\lambda_{\min}(f_1)\lambda_{\min}(f_2)}\right)^3\right)$ with probability $> 1 - 2C_2 p^2 \min(n_1, n_2)^{-H}$ RSC holds with $\kappa_L = \lambda_{\min}(f_1)\lambda_{\min}(f_2)/2$.

□

We are now ready to prove Theorem 1.

Proof of Theorem 1. We first proceed by establishing the decomposability of our penalty term, $\|\Delta\|_1$. Decomposability is defined in Definition 1 of Negahban et al. (2012). Let $\overline{\mathcal{M}} = \left\{\theta : \mathbb{R}^{2p\times 2p} \middle| \theta_{ij} \neq 0 \text{ for all } (i,j) \in S_\Delta\right\}$ and $\overline{\mathcal{M}}^\perp = \left\{\gamma : \mathbb{R}^{2p\times 2p} \middle| \gamma_{ij} = 0 \text{ for all } (i,j) \in S_\Delta\right\}$. Then $\|\cdot\|_1$ is decomposable since $\|\theta+\gamma\|_1 = \|\theta\|_1 + \|\gamma\|_1$ for all $\theta \in \overline{\mathcal{M}}$ and $\gamma \in \overline{\mathcal{M}}^\perp$.

Next we discuss a suitable choice of penalty scale τ_{n_1,n_2}. From Theorem 1 in Negahban et al. (2012) we require $\tau_{n_1,n_2} \geq 2\|\nabla L_D(\Delta, \hat{\Sigma}_1, \hat{\Sigma}_2)\|_\infty$. We know that $2\|\nabla L_D(\Delta, \hat{\Sigma}_1, \hat{\Sigma}_2)\|_\infty = 2\|0.5\left(\hat{\Sigma}_1\Delta\hat{\Sigma}_2 + \hat{\Sigma}_2\Delta\hat{\Sigma}_1\right) - \left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)\|_\infty$.

Let $\Gamma = 0.5\left(\Sigma_2 \otimes \Sigma_1 + \Sigma_1 \otimes \Sigma_2\right)$ and $\hat{\Gamma} = 0.5\left(\hat{\Sigma}_2 \otimes \hat{\Sigma}_1 + \hat{\Sigma}_1 \otimes \hat{\Sigma}_2\right)$. Then we write

$$\begin{aligned}
\|0.5\left(\hat{\Sigma}_1\Delta\hat{\Sigma}_2 + \hat{\Sigma}_2\Delta\hat{\Sigma}_1\right) - \left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)\|_\infty &= \|\hat{\Gamma}\text{vec}(\Delta) - \left(\text{vec}(\hat{\Sigma}_1) - \text{vec}(\hat{\Sigma}_2)\right)\|_\infty \\
&= \|(\hat{\Gamma}-\Gamma)\text{vec}(\Delta) - \left(\text{vec}(\hat{\Sigma}_1) - \text{vec}(\Sigma_1)\right) - \left(\text{vec}(\hat{\Sigma}_2) - \text{vec}(\Sigma_2)\right)\|_\infty \\
&\leq \|(\hat{\Gamma}-\Gamma)\text{vec}(\Delta)\|_\infty + \|\hat{\Sigma}_1 - \Sigma_1\|_\infty + \|\hat{\Sigma}_2 - \Sigma_2\|_\infty \\
&\leq \|\hat{\Gamma}-\Gamma\|_\infty\|\Delta\|_1 + \|\hat{\Sigma}_1 - \Sigma_1\|_\infty + \|\hat{\Sigma}_2 - \Sigma_2\|_\infty .
\end{aligned}$$

By Lemma 2 we have for $\min(n_1, n_2) \geq \frac{C_1^3}{C_f^3}8^{9/2}$, with probability $> 1 - 2C_2 p^2 \min(n_1, n_2)^{-H}$,

$$\begin{aligned}
\left\|0.5\left(\hat{\Sigma}_1\Delta\hat{\Sigma}_2 + \hat{\Sigma}_2\Delta\hat{\Sigma}_1\right) - \left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)\right\|_\infty &\leq \frac{C_1 C_f 8^{5/2}}{\min(n_1,n_2)^{1/3}}\|\Delta\|_1 + \frac{C_1 8^{3/2} 2}{n_1^{1/3}} + \frac{C_1 8^{3/2} 2}{n_2^{1/3}} \\
&\leq \frac{\left(C_1 C_f 8^{5/2}\|\Delta\|_1 + 4C_1 8^{3/2}\right)}{\min(n_1,n_2)^{1/3}} . \qquad (6)
\end{aligned}$$

Therefore if we choose

$$\tau_{n_1,n_2} \geq \frac{2\left(C_1 C_f 8^{5/2}\|\Delta\|_1 + 4C_1 8^{3/2}\right)}{\min(n_1,n_2)^{1/3}} ,$$

we will satisfy $\tau_{n_1,n_2} \geq 2\|\nabla L_D(\Delta, \hat{\Sigma}_1, \hat{\Sigma}_2)\|_\infty$ with probability $> 1 - 2C_2 p^2 \min(n_1, n_2)^{-H}$ when $\min(n_1, n_2) \geq \frac{C_1^3}{C_f^3}8^{9/2}$.

Combining the above with Lemma 3 we get that for $\tau_{n_1,n_2} \geq \frac{2}{\min(n_1,n_2)^{1/3}}\left(C_1 C_f 8^{5/2}\|\Delta\|_1 + 4C_1 8^{3/2}\right)$, $\min(n_1, n_2) \geq \max\left(8^{9/2}C_1^3/C_f^3, \left(\frac{32 s_\Delta C_f C_1 8^{5/2}}{\lambda_{\min}(f_1)\lambda_{\min}(f_2)}\right)^3\right)$, and $M_{n,l} = O(n_l^{2/3})$ we have with probability greater than $1 - Cp^2 \min(n_1, n_2)^{-H}$ for any $H > 0$,

$$\|\hat{\Delta} - \Delta\|_F \leq \frac{\sqrt{18 s_\Delta}\tau_{n_1,n_2}}{\lambda_{\min}(f_1)\lambda_{\min}(f_2)} ,$$

where we have defined $C = 2C_2$. Since τ_{n_1,n_2} is of order $\min(n_1, n_2)^{-1/3}$, convergence of our D-trace estimator is also of order $\min(n_1, n_2)^{-1/3}$.

□

D SIMULATION SPECIFICS AND RESULTS

We generate data in condition l as

$$\mathbf{x}_{l,t} = A_l \mathbf{x}_{l,t-1} + \boldsymbol{\epsilon}_{l,t}$$

where $\boldsymbol{\epsilon}_{l,t} \sim N_p(\mathbf{0}, I_p)$. From Sun et al. (2018), the spectral density at frequency λ is known to be

$$f_l(\lambda) = \frac{1}{2\pi} \left(\mathcal{A}_l(e^{-i\lambda})\right)^{-1} I_p \left(\left(\mathcal{A}_l(e^{-i\lambda})\right)^{-1}\right)^H,$$

where $\mathcal{A}_l(z) = I_p - A_l z$. When the transition matrix A_l is block diagonal, the spectral density $f_l(\lambda)$ is also block diagonal. Since the inverse of a block diagonal matrix can be computed block by block we can easily generate a sparse difference matrix by enforcing the transition matrix in conditions 1 and 2 to be the same except for a small block. For example, with $p = 54$ if we generate A_1 and A_2 as block diagonal with the same 51×51 block and only differing in the final 3×3 block, then their spectral densities will only differ in this last 3×3 block. When converting to the real space, the expanded $\Sigma_i \in (2p \times 2p)$ and thus the true difference $\Delta = \Sigma_1 - \Sigma_2$ will differ in four 3×3 blocks.

For simulation setting 1, the transition matrix is the same as in (Sun et al., 2018). Specifically, it consists of 18 blocks of dimension 3×3, where each block is $\begin{bmatrix} 0.5 & 0.9 & 0 \\ 0 & 0.5 & 0.9 \\ 0 & 0 & 0.5 \end{bmatrix}$. In both simulation setting 2 and setting 3, 60% of the coefficients in the 3×3 block were randomly drawn from either a Uniform$(-0.5, -0.2)$ or a Uniform$(0.2, 0.5)$ each with equal probability. In the second and third setting, 40% and 5% respectively of the entries of the larger block were randomly drawn from a Uniform$(-0.5, -0.2)$ and Uniform$(0.2, 0.5)$ each with equal probability. This corresponded to 60% and 95% sparsity respectively. The number of non-zero entries in the difference in expanded inverse spectral densities, which we will refer to as edges, varies by frequency but is almost always 22 for Simulation 1, 14 for Simulation 2, and 28 for Simulation 3.

For all settings 100 evenly spaced Fourier frequencies from 0 to $\pi - n^{-1}$ was used. These were used as the spectral density is conjugate symmetric around 0. In the case of $n = 100$ only 50 Fourier frequencies were used as there are only 50 Fourier frequencies from 0 to $\pi - n^{-1}$. Given the availability of results from multiple frequencies, each simulation was run once and the results are averaged across all frequencies. The standard error of each metric was also computed across frequencies and are reported in parentheses.

To solve the SDD estimation problem (1), the alternating direction method of multipliers (ADMM) algorithm from Yuan et al. (2017) was used. Specifically, we used the `L1_dts` function in the `Difdtl` package available on GitHub at SusanYuan/Difdtl (GPL ($\geq$ 2)). To generate the sequence of penalties, $\{\tau_{n_1,n_2}\}$, we used 20 values on a log-linear scale from $0.001 * \tau_{n_1,n_2,\max}$ to $\tau_{n_1,n_2,\max}$ where $\tau_{n_1,n_2,\max}$ represents the minimum value where all entries of $\hat{\Delta}$ are 0. In this case, $\tau_{n_1,n_2,\max} = 2\max(|\hat{\Sigma}_1 - \hat{\Sigma}_2|)$. The penalty $\tau^*_{n_1,n_2}$ was chosen as the τ_{n_1,n_2} that minimizes the eBIC which is computed as

$$\text{eBIC}(\hat{\Delta})_\gamma = \min(n_1, n_2) \left\| \frac{1}{2}\left(\hat{\Sigma}_1\hat{\Delta}\hat{\Sigma}_2 + \hat{\Sigma}_2\hat{\Delta}\hat{\Sigma}_1 - \hat{\Sigma}_2 + \hat{\Sigma}_1\right)\right\|_\infty + \log(\min(n_1, n_2))|E| + 4\gamma|E|\log(p), \quad (7)$$

where $|E|$ is the number of unique edges in $\hat{\Delta}$. Since $\hat{\Delta}$ represents the difference in expanded spectral densities $|E|$ is the number of non-zero entries in the upper triangular portions, including diagonals, of the submatrices $\hat{\Delta}_{1:p,1:p}$ and $\hat{\Delta}_{1:p,(p+1):2p}$. For all applications we use $\gamma = 0.5$. For FGL, Naïve, and Hard Thresholding, 20 tuning parameter values were also used. FGL requires two tuning parameter values, λ_1, λ_2, so a grid of 20 (λ_1, λ_2) pairs was considered. Since the parameter λ_1 controls the sparsity of each inverse matrix while λ_2 controls sparsity of the difference we considered only two values for λ_1 and 10 values for λ_2. The values for λ_1 were $(0.01\lambda_{1,\max}, 0.1\lambda_{1,\max})$ where $\lambda_{1,\max} = \left\|\hat{\Sigma}_{1,O} + \hat{\Sigma}_{2,O}\right\|_\infty / 2$ and $\hat{\Sigma}_{i,O}$ is the off-diagonals of $\hat{\Sigma}_i$. The

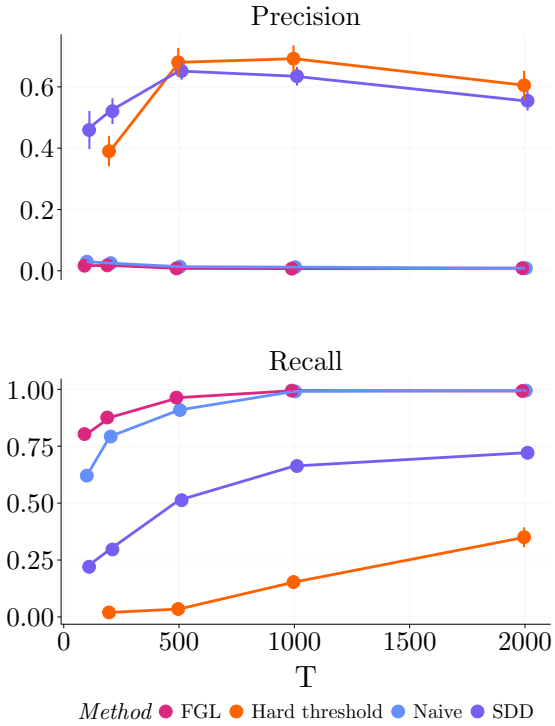


Figure 8: Simulation 1 Precision and Recall. Results are reported as mean (dots) and SE (vertical lines) where the mean and SE are taken across all frequencies for a given sample size n.

λ_2 sequence was generated using 10 values on a log-linear scale from $0.0001 * \lambda_{2,\max}$ to $\lambda_{2,\max}$ where $\lambda_{2,\max} = \max\left(\left\|\hat{\Sigma}_{1,O}\right\|_\infty - 0.01\lambda_{1,\max}, \left\|\hat{\Sigma}_{2,O}\right\|_\infty - 0.01\lambda_{1,\max}\right)$. For the Naïve method, we first computed $\lambda_{\max,i} = \left\|\hat{\Sigma}_i\right\|_\infty$ in each condition and then computed the tuning parameters as 20 evenly-spaced values on a log-linear scale from $0.0001\lambda_{\max,i}$ to $\lambda_{\max,i}$. Similarly, for the Hard Thresholding method, we used 20 evenly-spaced values on a log-linear scale from $\min(|\hat{\Delta}_H|)$ to $\max(|\hat{\Delta}_H|)$ where $\hat{\Delta}_H$ is generated by inverting $\hat{\Sigma}_i$ and forming the difference. Note that for $n = 100$, it is not possible to invert $\hat{\Sigma}_i$.

Let TP, FN, TN, FP denote the true positive, false negative, true negative, and false positive edges identified by either SDD or the Naïve method, respectively. A value greater than 1×10^{-6} in absolute value was considered an edge. The metrics are defined as follows

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{4p^2} \qquad \text{RRMSE} = \sqrt{\frac{\sum_{i,j}(\hat{\Delta}_{i,j} - \Delta_{i,j})^2}{\sum_{i,j}(\Delta_{i,j})^2}},$$

where $\hat{\Delta}$ represents the difference estimator, either SDD or the Naïve difference, Δ is the true difference. The denominator of the accuracy measure is $4p^2$ as expanding a $p \times p$ spectral density to the real space gives a $2p \times 2p$ matrix which has $4p^2$ entries. We also report the number of average number of true edges and the average number of estimated edges across frequencies.

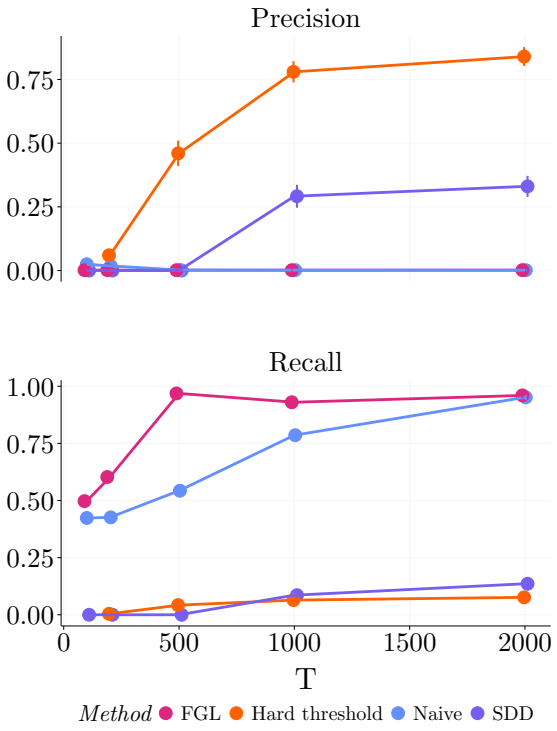


Figure 9: Simulation 2 Precision and Recall. Results are reported as mean (dots) and SE (vertical lines) where the mean and SE are taken across all frequencies for a given sample size n.

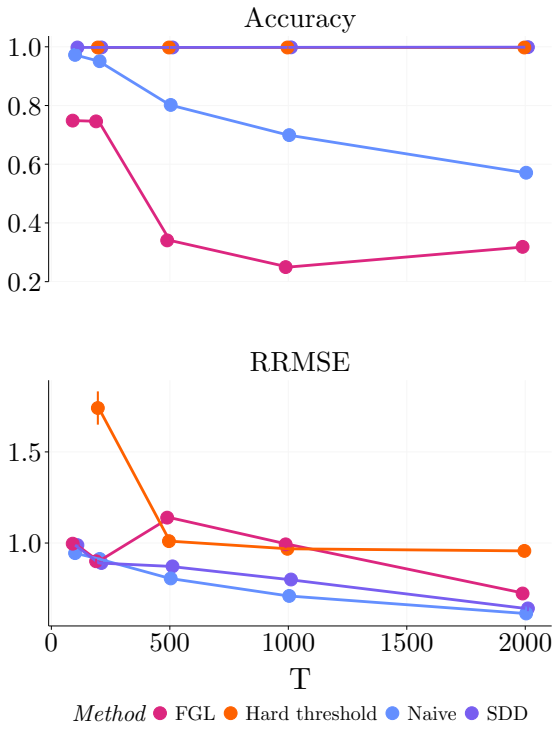


Figure 10: Simulation 3 Accuracy and RRMSE. Results are reported as mean (dots) and SE (vertical lines) where the mean and SE are taken across all frequencies for a given sample size n.

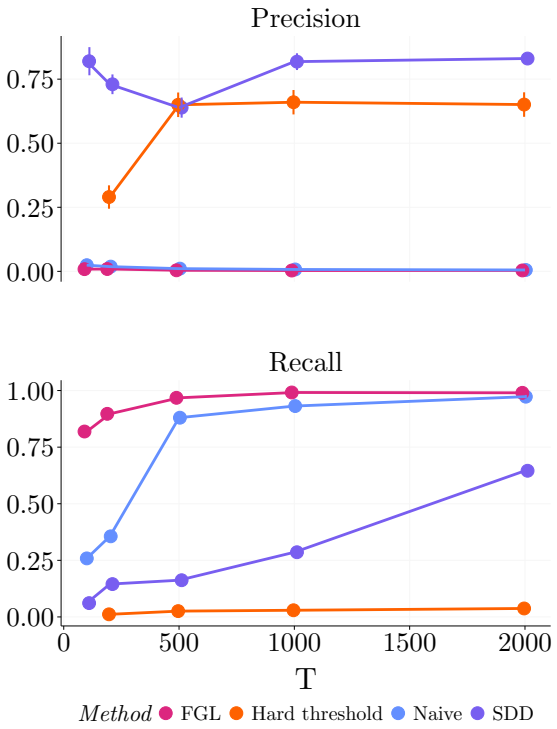


Figure 11: Simulation 3 Precision and Recall. Results are reported as mean (dots) and SE (vertical lines) where the mean and SE are taken across all frequencies for a given sample size n.

Table 1: Simulation 1. Results are reported as Mean (SE) where the mean and SE are taken across all frequencies for a given sample size n.

		SDD				
n	# True edges	# Est edges	Precision	Recall	Accuracy	RRMSE
100	21.6 (0.32)	10.6 (1.04)	0.46 (0.06)	0.22 (0.03)	1.00 (0.00)	0.95 (0.01)
200	21.8 (0.16)	16.7 (1.00)	0.52 (0.04)	0.30 (0.02)	1.00 (0.00)	0.89 (0.01)
500	21.9 (0.08)	21.1 (1.15)	0.65 (0.03)	0.51 (0.02)	1.00 (0.00)	0.78 (0.01)
1000	21.9 (0.08)	31.1 (2.03)	0.63 (0.03)	0.66 (0.02)	1.00 (0.00)	0.61 (0.01)
2000	21.9 (0.08)	48.7 (4.60)	0.55 (0.03)	0.72 (0.02)	1.00 (0.00)	0.53 (0.01)
			FGL			
		2073 (194)	0.02 (0.00)	0.80 (0.02)	0.82 (0.02)	0.89 (0.01)
		2444 (137)	0.02 (0.00)	0.88 (0.01)	0.79 (0.01)	0.84 (0.01)
		4973 (326)	0.01 (0.00)	0.96 (0.01)	0.58 (0.03)	0.68 (0.01)
		5962 (324)	0.01 (0.00)	0.99 (0.00)	0.49 (0.03)	0.64 (0.02)
		6308 (327)	0.01 (0.00)	0.99 (0.00)	0.46 (0.03)	0.62 (0.02)
			Naïve			
		659 (46.1)	0.03 (0.00)	0.62 (0.02)	0.94 (0.00)	0.95 (0.01)
		1077 (56.9)	0.03 (0.00)	0.79 (0.02)	0.91 (0.00)	0.90 (0.00)
		1775 (72.1)	0.01 (0.00)	0.91 (0.01)	0.85 (0.01)	0.80 (0.01)
		2289 (104)	0.01 (0.00)	0.99 (0.00)	0.81 (0.01)	0.69 (0.01)
		3158 (151)	0.01 (0.00)	0.99 (0.00)	0.73 (0.01)	0.58 (0.01)
			Hard threshold			
		-	-	-	-	-
		0.69 (0.05)	0.39 (0.05)	0.02 (0.00)	1.00 (0.00)	1.57 (0.06)
		0.79 (0.06)	0.68 (0.05)	0.03 (0.00)	1.00 (0.00)	1.00 (0.00)
		5.01 (1.11)	0.69 (0.04)	0.15 (0.03)	1.00 (0.00)	0.89 (0.02)
		230 (53.1)	0.60 (0.05)	0.35 (0.04)	0.98 (0.00)	1.04 (0.06)

Table 2: Simulation 2. Results are reported as Mean (SE) where the mean and SE are taken across all frequencies for a given sample size n.

		SDD				
n	# True edges	# Est edges	Precision	Recall	Accuracy	RRMSE
100	13.7 (0.21)	1.52 (0.27)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)
200	13.9 (0.11)	0.90 (0.13)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)
500	14.0 (0.04)	2.24 (0.35)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)
1000	14.0 (0.04)	3.80 (0.47)	0.29 (0.04)	0.09 (0.01)	1.00 (0.00)	0.99 (0.01)
2000	14.0 (0.04)	6.52 (0.83)	0.33 (0.04)	0.14 (0.01)	1.00 (0.00)	0.95 (0.01)
		FGL				
		4675 (227)	0.00 (0.00)	0.50 (0.02)	0.60 (0.02)	2.25 (0.16)
		5297 (185)	0.00 (0.00)	0.60 (0.02)	0.55 (0.02)	2.15 (0.12)
		10201 (118)	0.00 (0.00)	0.97 (0.01)	0.13 (0.01)	4.20 (0.07)
		9731 (155)	0.00 (0.00)	0.93 (0.02)	0.17 (0.01)	3.09 (0.06)
		9196 (176)	0.00 (0.00)	0.96 (0.01)	0.21 (0.02)	2.30 (0.05)
		Naïve				
		259 (10.9)	0.02 (0.00)	0.42 (0.01)	0.98 (0.00)	1.02 (0.01)
		609 (46.3)	0.02 (0.00)	0.43 (0.00)	0.95 (0.00)	1.01 (0.01)
		3566 (56.3)	0.00 (0.00)	0.54 (0.01)	0.69 (0.00)	1.20 (0.03)
		6487 (58.2)	0.00 (0.00)	0.79 (0.02)	0.44 (0.00)	1.58 (0.04)
		9145 (98.2)	0.00 (0.00)	0.95 (0.01)	0.22 (0.01)	2.29 (0.04)
		Hard threshold				
		-	-	-	-	-
		0.72 (0.05)	0.06 (0.02)	0.00 (0.00)	1.00 (0.00)	3.05 (0.18)
		0.93 (0.08)	0.46 (0.05)	0.04 (0.01)	1.00 (0.00)	1.09 (0.02)
		0.93 (0.06)	0.78 (0.04)	0.06 (0.00)	1.00 (0.00)	0.94 (0.01)
		1.08 (0.06)	0.84 (0.04)	0.08 (0.00)	1.00 (0.00)	0.89 (0.01)

Table 3: Simulation 3. Results are reported as Mean (SE) where the mean and SE are taken across all frequencies for a given sample size n.

		SDD				
n	# True edges	# Est edges	Precision	Recall	Accuracy	RRMSE
100	27.4 (0.39)	1.76 (0.12)	0.82 (0.05)	0.06 (0.00)	1.00 (0.00)	0.99 (0.00)
200	27.7 (0.20)	4.80 (0.31)	0.73 (0.04)	0.14 (0.01)	1.00 (0.00)	0.89 (0.01)
500	27.9 (0.12)	6.68 (0.56)	0.64 (0.04)	0.16 (0.01)	1.00 (0.00)	0.87 (0.01)
1000	27.9 (0.12)	8.86 (0.71)	0.82 (0.03)	0.29 (0.02)	1.00 (0.00)	0.80 (0.02)
2000	27.9 (0.12)	22.1 (0.77)	0.83 (0.01)	0.65 (0.02)	1.00 (0.00)	0.64 (0.01)
			FGL			
		2950 (157)	0.01 (0.00)	0.82 (0.03)	0.75 (0.01)	1.00 (0.02)
		2984 (97.0)	0.01 (0.00)	0.90 (0.02)	0.75 (0.01)	0.90 (0.01)
		7718 (254)	0.00 (0.00)	0.97 (0.01)	0.34 (0.02)	1.14 (0.03)
		8780 (137)	0.00 (0.00)	0.99 (0.00)	0.25 (0.01)	0.99 (0.02)
		7975 (156)	0.00 (0.00)	0.99 (0.01)	0.32 (0.01)	0.72 (0.01)
			Naïve			
		308 (12.7)	0.02 (0.00)	0.26 (0.01)	0.97 (0.00)	0.94 (0.01)
		563 (20.2)	0.02 (0.00)	0.36 (0.01)	0.95 (0.00)	0.91 (0.01)
		2331 (34.6)	0.01 (0.00)	0.88 (0.02)	0.80 (0.00)	0.81 (0.01)
		3532 (49.8)	0.01 (0.00)	0.93 (0.01)	0.70 (0.00)	0.71 (0.01)
		5030 (61.8)	0.01 (0.00)	0.97 (0.01)	0.57 (0.01)	0.61 (0.01)
			Hard threshold			
		-	-	-	-	-
		0.58 (0.05)	0.29 (0.05)	0.01 (0.00)	1.00 (0.00)	1.74 (0.09)
		0.72 (0.06)	0.65 (0.05)	0.03 (0.00)	1.00 (0.00)	1.01 (0.00)
		0.81 (0.07)	0.66 (0.05)	0.03 (0.00)	1.00 (0.00)	0.97 (0.00)
		2.67 (1.78)	0.65 (0.05)	0.04 (0.01)	1.00 (0.00)	0.96 (0.00)

Table 4: Simulation 1. Difference Between SDD and FGL for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-2062 (194)	0.44 (0.06)	-0.58 (0.03)	0.17 (0.02)	0.06 (0.01)
200	-2428 (137)	0.50 (0.04)	-0.58 (0.02)	0.21 (0.01)	0.06 (0.02)
500	-4952 (326)	0.64 (0.03)	-0.45 (0.02)	0.42 (0.03)	0.10 (0.02)
1000	-5931 (325)	0.63 (0.03)	-0.33 (0.02)	0.51 (0.03)	-0.03 (0.03)
2000	-6259 (329)	0.55 (0.03)	-0.27 (0.02)	0.54 (0.03)	-0.08 (0.03)

Table 5: Simulation 1. Difference Between SDD and Naïve Method for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-649 (46.0)	0.43 (0.06)	-0.40 (0.03)	0.05 (0.00)	0.00 (0.00)
200	-1060 (56.4)	0.50 (0.04)	-0.50 (0.01)	0.09 (0.00)	-0.01 (0.01)
500	-1754 (71.9)	0.64 (0.03)	-0.40 (0.02)	0.15 (0.01)	-0.02 (0.01)
1000	-2258 (103)	0.62 (0.03)	-0.33 (0.02)	0.19 (0.01)	-0.07 (0.01)
2000	-3109 (149)	0.55 (0.03)	-0.27 (0.02)	0.27 (0.01)	-0.04 (0.01)

Table 6: Simulation 1. Difference Between SDD and Hard Thresholding for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-	-	-	-	-
200	16.0 (1.00)	0.13 (0.07)	0.28 (0.02)	-0.00 (0.00)	-0.67 (0.06)
500	20.3 (1.15)	-0.03 (0.06)	0.48 (0.02)	0.00 (0.00)	-0.22 (0.01)
1000	26.1 (1.86)	-0.06 (0.05)	0.51 (0.04)	-0.00 (0.00)	-0.27 (0.03)
2000	-182 (52.5)	-0.05 (0.05)	0.37 (0.05)	0.02 (0.00)	-0.50 (0.07)

Table 7: Simulation 2. Difference Between SDD and FGL for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-4674 (227)	-0.00 (0.00)	-0.50 (0.02)	0.40 (0.02)	-1.25 (0.16)
200	-5296 (185)	-0.00 (0.00)	-0.60 (0.02)	0.45 (0.02)	-1.15 (0.12)
500	-10199 (119)	-0.00 (0.00)	-0.97 (0.01)	0.87 (0.01)	-3.20 (0.07)
1000	-9727 (155)	0.29 (0.04)	-0.84 (0.02)	0.83 (0.01)	-2.10 (0.06)
2000	-9189 (177)	0.33 (0.04)	-0.82 (0.02)	0.79 (0.02)	-1.35 (0.05)

Table 8: Simulation 2. Difference Between SDD and Naïve Method for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-257 (10.7)	-0.02 (0.00)	-0.42 (0.01)	0.02 (0.00)	-0.02 (0.01)
200	-608 (46.2)	-0.02 (0.00)	-0.43 (0.00)	0.05 (0.00)	-0.01 (0.01)
500	-3564 (56.3)	-0.00 (0.00)	-0.54 (0.01)	0.30 (0.00)	-0.20 (0.03)
1000	-6483 (58.4)	0.29 (0.04)	-0.70 (0.02)	0.55 (0.00)	-0.59 (0.04)
2000	-9139 (98.6)	0.33 (0.04)	-0.82 (0.02)	0.78 (0.01)	-1.34 (0.04)

Table 9: Simulation 2. Difference Between SDD and Hard Thresholding for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-	-	-	-	-
200	0.18 (0.14)	-0.06 (0.02)	-0.00 (0.00)	-0.00 (0.00)	-2.05 (0.18)
500	1.31 (0.36)	-0.46 (0.05)	-0.04 (0.01)	-0.00 (0.00)	-0.09 (0.02)
1000	2.87 (0.48)	-0.49 (0.06)	0.02 (0.01)	-0.00 (0.00)	0.05 (0.01)
2000	5.44 (0.84)	-0.51 (0.06)	0.06 (0.01)	-0.00 (0.00)	0.05 (0.01)

Table 10: Simulation 3. Difference Between SDD and FGL for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-2948 (157)	0.81 (0.05)	-0.76 (0.03)	0.25 (0.01)	-0.01 (0.01)
200	-2979 (96.9)	0.72 (0.04)	-0.75 (0.02)	0.25 (0.01)	-0.01 (0.02)
500	-7711 (254)	0.63 (0.04)	-0.81 (0.02)	0.66 (0.02)	-0.27 (0.03)
1000	-8771 (137)	0.82 (0.03)	-0.71 (0.02)	0.75 (0.01)	-0.19 (0.02)
2000	-7953 (155)	0.83 (0.01)	-0.34 (0.02)	0.68 (0.01)	-0.08 (0.02)

Table 11: Simulation 3. Difference Between SDD and Naïve Method for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-306 (12.7)	0.80 (0.05)	-0.20 (0.01)	0.03 (0.00)	0.04 (0.00)
200	-558 (20.0)	0.71 (0.04)	-0.21 (0.01)	0.05 (0.00)	-0.02 (0.00)
500	-2324 (34.7)	0.63 (0.04)	-0.72 (0.02)	0.20 (0.00)	0.07 (0.01)
1000	-3523 (49.6)	0.81 (0.03)	-0.65 (0.02)	0.30 (0.00)	0.09 (0.01)
2000	-5008 (61.8)	0.82 (0.01)	-0.33 (0.02)	0.43 (0.01)	0.03 (0.01)

Table 12: Simulation 3. Difference Between SDD and Hard Thresholding Method for Each Metric.

n	# Est edges	Precision	Recall	Accuracy	RRMSE
100	-	-	-	-	-
200	4.22 (0.31)	0.44 (0.06)	0.13 (0.01)	0.00 (0.00)	-0.85 (0.09)
500	5.96 (0.56)	-0.01 (0.06)	0.14 (0.01)	0.00 (0.00)	-0.14 (0.01)
1000	8.05 (0.72)	0.16 (0.06)	0.26 (0.02)	0.00 (0.00)	-0.17 (0.02)
2000	19.4 (1.87)	0.18 (0.05)	0.61 (0.02)	0.00 (0.00)	-0.32 (0.02)

김양진 교수

논문

- Kim, Y-J (2025). Analysis of interval-censored competing risk data using approximate likelihood, Communications for Statistical applications and methods. 32, 249-257

- Che Ramil Amirah and Kim, Y-J (2025). Quantile regression model for interval-censored data with competing risks. Journal of applied Statistics. 52,2438-2447.

- Lee, C, Kim, Y-J, Sohn, i., Kim, J-H, and Lee H.Y. (2025). Tracing metastatic evolutionary patterns in lung adenocarcinoma: Progression dissection based on a multi-state model. Cancer research and treatment. 57(4):1019-1029

Communications for Statistical Applications and Methods
2025, Vol. 32, No. 2, 249–257

https://doi.org/10.29220/CSAM.2025.32.2.249
Print ISSN 2287-7843 / Online ISSN 2383-4757

Analysis of interval-censored competing risk data using approximate likelihood

Yang-Jin Kim[1,a]

[a]Department of Statistics, Sookmyung Women's University, Korea

Abstract

Interval-censored competing risk data occurs when subjects are at risk of several causes and the failure time is known to lie at somewhere between two sequential observation times. For investigating the effect of covariate on the hazard, Goetghebeur and Ryan (2000)s approximate likelihood is extended with a piecewise baseline hazard assumption. The EM algorithm is implemented to update unknown event status and risk status and the IPCW technique is applied to adjust the effect of the subjects experiencing competing events. Simulation results show the estimates based on the suggested method are unbiased and satisfy the coverage probabilities. Two HIV-related datasets (HIV data from intcrr package and the Bangkok Metropolitan Administration (BMA) prospective study HIV-seronegative injecting drug users data) are analyzed as real examples.

Keywords: approximate likelihood, competing risk data, EM algorithm, interval-censoring data

1. Introduction

Competing risk data arise when subjects are at risk of more than one cause and the occurrence of one event precludes the occurrence of the other events. Diverse methods have been proposed based on a cause-specific cumulative incidence function (CIF) defined as the cumulative probability of experiencing a particular cause event until a certain time. Its merit is to consider the subjects cannot experiencing the event because of the occurrence of competing events. In the context of right censored data, Fine and Gray (1999) proposed a sub-distribution proportional hazard model by adopting an inverse probability censoring weight (IPCW) and Geskus (2011) extended to left truncated data using a time-dependent weight. Klein and Andersen (2005) implemented a pseudo-value approach to analyze a competing risk data as well as a multi-state model.

Interval-censored data frequently occurs in the observation study where subjects are monitored to check the occurrence of the event of interest (Sun, 2006). Let L_i and U_i denote the last observation time before the occurrence of event and the first one after it, respectively. As special cases, $L_i = 0$ results in a left censoring and $L_i < U_i = \infty$ indicates a right censoring, respectively. For interval-censored data with competing risk, several approaches to estimate distribution function and regression model have been proposed. Hudgens *et al.* (2001) suggested the NPMLE (nonparametric maximum likelihood estimator) by extending the self-consistency algorithm of Turnbull (1976) and Kim (2017) also considered a weighting method to obtain the NPMLE of CIF. For regression models, Hudgens *et al.* (2014) suggested a parametric regression model and Mao *et al.* (2017) showed the application of several transformation models based on the EM algorithm. Kim and Kim (2016)

[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Yonsan-Gu, Seoul 04310, Korea. E-mail: yjin@sookmyung.ac.kr

Published 31 March 2025 / journal homepage: http://csam.or.kr
© 2025 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

adopted a pseudo-value approach and Do and Kim (2017) applied this approach to solve a missing cause problem. Li (2016) considered a monotone *B*-spline and Bakoyannis *et al.* (2017) considered a general transformation model and showed the procedure based on the sieve maximum likelihood estimation. They also provided the R package **intccr**.

In this article, we propose a regression model based on an approximate likelihood of Goetghebeur and Ryan (2000) who extended the partial likelihood of right censored data to interval-censored regression model. Under no competing risk, a piecewise exponential hazard is assumed on fine-grid of points $0 = w_0 < w_1 < \cdots < w_r$. In more detail, the baseline hazard λ_j is defined on the j^{th} interval $\tilde{a}_j = (w_{j-1}, w_j]$, $j = 1, \ldots, r$, which is derived from the equivalence set (Lindsey and Ryan, 1998). Let dN_{ij} indicate whether a subject has experienced the event at the j^{th} interval and Y_{ij} show whether he or she is at risk in the j^{th} interval, respectively. Given a p-vector of covariate Z and the proportional hazard assumption $\lambda(t;\theta_0) = \lambda(t)\exp(\tau' Z)$, the approximate likelihood of $\theta_0 = (\lambda_1, \ldots, \lambda_r, \tau_1, \ldots, \tau_p)'$ is defined as

$$L(\theta_0) = \prod_{i=1}^{n}\prod_{j=1}^{r}\left[\left\{\lambda_j \exp\left(\tau' Z_i\right)\right\}^{dN_{ij}} \exp\left\{-\sum_{l<j}\lambda_l \exp\left(\tau' Z_i\right) Y_{il}\right\}\right], \tag{1.1}$$

and the score functions of $\lambda_j (j = 1, \ldots, r)$ and τ are

$$\sum_{i=1}^{n}\left\{\frac{1}{\lambda_j} dN_{ij} - \exp\left(\tau' z_i\right) Y_{ij}\right\} = 0, \tag{1.2}$$

$$\sum_{j=1}^{r}\sum_{i=1}^{n} Z_i\left(dN_{ij} - \lambda_j \exp\left(\tau' z_i\right) Y_{ij}\right) = 0. \tag{1.3}$$

However, an interval-censoring brings incomplete information about the occurrence of event and dN_{ij} and Y_{ij} become missing data. We consider the EM algorithm is adopted to recover them. Now, define $e_{ij} = E(dN_{ij}|\tilde{O}_i, \theta_0)$ and $r_{ij} = E(Y_{ij}|\tilde{O}_i, \theta_0)$ as the conditional expectations and they are estimated in the E-step using the updated parameters and data $\tilde{O}_i = (L_i, U_i, Z_i)$ as follows,

$$e_{ij} = \frac{p_{ij}}{\sum_{w_s \in (L_i, U_i]} p_{is}}, \quad w_j \in (L_i, U_i],$$

$$r_{ij} = \sum_{s \geq j} e_{is}\delta_i + (1 - \delta_i) I(w_j < L_i),$$

where $\delta_i = I(U_i \neq \infty)$ and $p_{ij} = \lambda_j \exp(\tau' Z_i)[\exp\{-\sum_{l=1}^{j} r_{il}\lambda_l \exp(\tau' Z_i)\}]$ denotes that the subject i experiences the event of interest at the j^{th} interval $(w_{j-1}, w_j]$. Once updating the missing quantities with e_{ij} and r_{ij}, the solutions of (1.2) and (1.3) are derived in the M-step. In this article, our interest is to extend the EM algorithm to the regression model of interval-censored failure time data with compering risk. In Section 2, we introduce the notations and describe the EM algorithm. Section 3 presents simulation results under several schemes and Section 4 deals the application of suggested method to two real datasets. Several related future works are suggested in Section 5.

2. Approximate likelihood and the EM algorithm

Interval-censored data with competing risks are comprised of $\{L_i, U_i, \eta_i, \delta_i, Z_i, i = 1, \ldots, n\}$, where (L_i, U_i) are observation times satisfying $L_i < \tilde{T}_i \leq U_i$ and $\tilde{T}_i = \min(T_{1i}, \ldots, T_{Ki})$, $\delta_i = I(\tilde{T}_i < C_i)$ where a censoring time C_i, the k^{th} cause event time T_{ki} and the cause indicator $\eta_i = \{1, \ldots, K\}$. In particular, for a censored subject with $\delta_i = 0$, set $L_i = C_i$ and $U_i = \infty$. Given a vector of covariates $Z = z$, the k^{th} cumulative incidence regression function is defined as

$$F_k(t; \theta_k, z) = \Pr(T_k \leq t | \theta_k, z) = \Pr(\tilde{T} \leq t, \eta = k | \theta_k, z), \quad k = 1, \ldots, K.$$

Based on the relation of $F_k(t; \theta_k, z) = 1 - \exp\{-\Lambda_k(t; \theta_k, z)\}$ with $\Lambda_k(t; \theta_k, z) = \int_0^t \lambda_k(s; \theta_k, z) ds$, $\theta_k = (\lambda_1^k, \ldots, \lambda_r^k, \beta_k)$, the corresponding sub-distribution hazard function is defined as

$$\lambda_k(t; \theta_k, z) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \Pr\{t \leq T_k < t + \epsilon \mid T_k > t \cup (T_l \leq t \cap l \neq k), z\}.$$

Under an independent assumption between observation scheme and event occurrence time, the likelihood function is defined as

$$L(\theta) = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \left(F_k(U_i; \theta_k, z_i) - F_k(L_i; \theta_k, z_i) \right)^{I(\eta_i = k)\delta_i} \times \left(1 - \sum_{j=1}^{K} F_j(L_i; \theta_j, z_i) \right)^{1 - \delta_i} \right],$$

where the first term reflects the incidence of the k^{th} cause event and the second one represents at-risk status until the event occurs. Assume that all cause-specific hazards follow the Fine-Gray model (1999),

$$\lambda_k(t; \theta_k, z_i) = \lambda_{0k}(t) \exp\left(\beta_k' z_i\right).$$

In this paper, we apply the approximate likelihood to the subdistribution baseline hazard function which makes it feasible to apply the EM algorithm. The fine-grids of points $u_1 < \cdots < u_r$ are calculated from the equivalence set derived from the pooled data irrespective of causes and the resulting intervals are denoted as $\tilde{u}_j = (u_{j-1}, u_j]$. For each cause, the baseline functions are denoted as λ_j^k, $k = 1, \ldots, K$ and $j = 1, \ldots, r$. Then define dN_{ij}^k as the cause-specific event indicator showing whether the subject i has experienced the k^{th} cause event at the $\tilde{u}_j$th interval and Y_{ij} indicates whether the subject i is at-risk in interval $\tilde{u}_j$. Given the dN_{ij}^k and Y_{ij}, the complete data likelihood is written as

$$L(\theta) = \prod_{k=1}^{K} \prod_{i=1}^{n} \prod_{j=1}^{r} \left[\left(\lambda_j^k \exp\left(\beta_k' z_i\right) \right)^{dN_{ij}^k} \times \exp\left\{ - \sum_{l \leq j} Y_{il} \left(\lambda_l^k e^{\beta_k' z_i} \right) \right\} \right] = \prod_{k=1}^{K} L_k(\theta_k), \tag{2.1}$$

Then the log-likelihood of the k^{th} cause for $\theta_k = (\beta_k, \lambda_1^k, \ldots, \lambda_r^k)$ is written as

$$\log(L_k(\theta_k)) = \sum_{i=1}^{n} \sum_{j=1}^{r} \left\{ dN_{ij}^k \left(\log \lambda_j^k + \beta_k' z_i \right) - \left(\sum_{l=1}^{j} Y_{il} \lambda_l^k e^{\beta_k' z_i} \right) \right\}.$$

Given $\{O_i = (L_i, U_i, \eta_i, \delta_i, z_i), i = 1, \ldots, n\}$ and the estimates of parameters $\tilde{\theta} = (\tilde{\theta}_k, k = 1, \ldots, K)$, two unknown quantities dN_{ij}^k and Y_{ij} are updated with their conditional expectations in the E-step as follows

$$e_{ij}^k = E\left(dN_{ij}^k = 1 \mid O_i, \tilde{\theta}\right) = \delta_i I(\eta_i = k) \frac{p_{ij}^k}{\sum_{L_i < u_s < U_i} p_{is}^k}, \quad u_j \in (L_i, U_i],$$

where $p_{ij}^k = \Pr(\tilde{T}_{ki} \in \tilde{u}_j | z_i)$ is the probability of experiencing the k^{th} cause event at the j^{th} interval and estimated by

$$\hat{p}_{ij}^k = \hat{\lambda}_j^k e^{\hat{\beta}_k' z_i} \exp^{-\sum_{l=1}^K \sum_{s=1}^j Y_{is} \lambda_s^l \exp(\hat{\beta}_l' z_i)}.$$

For the conditional expectation of risk status $r_{ij} = E(Y_{ij} | O_i, \tilde{\theta}_k)$, the subjects with k^{th} event of interest and ones with the competing events have different weights on the risk set which is represented with the IPCW technique,

$$r_{ij}^k = I(\eta_i = k) \sum_{m \geq j} e_{im}^k + I(\eta_i \neq k) \sum_{m \geq j} w_{im} e_{im}^{k'} + I(\delta_i = 0) I(u_j \leq L_i), \tag{2.2}$$

where the second term of the right side represents the contribution of the subjects with competing event and a weight $w_{im} = \hat{S}_c^{-1}(u_m)$ is calculated with the estimated survival function of censoring time. Here $\hat{S}_c(t)$ is obtained with either a Kaplan-Meier estimator or suitable regression models. Given e_{ij}^k and r_{ij}^k for each cause k, the score functions of $\lambda_j^k (j = 1, \ldots, r)$ and $\beta_{kl} (l = 1, \ldots, p)$ are expressed as

$$\frac{\partial l(\theta)}{\partial \lambda_j^k} = \sum_{i=1}^n \left[e_{ij}^k \frac{1}{\lambda_j^k} - r_{ij}^k \exp\left(\beta_k' z_i\right) \right] = 0, \quad j = 1, \ldots, r \tag{2.3}$$

$$\frac{\partial l(\theta)}{\partial \beta_{kl}} = \sum_{i=1}^n \sum_{j=1}^r z_{il} \left\{ e_{ij}^k - r_{ij}^k \lambda_j^k e^{\beta_k' z_i} \right\} = 0, \quad l = 1, \ldots, p, \tag{2.4}$$

where (2.3) and (2.4) are reduced to (1.2) and (1.3) under no competing risk. Then the baseline subdistribution hazard function is estimated by

$$\hat{\lambda}_j^k = \frac{\sum_{i=1}^n e_{ij}^k}{\sum_{i=1}^n \exp(\beta_k' z_i) r_{ij}^k}. \tag{2.5}$$

By inputing (2.5) into (2.4), the score function of $\beta_{kl}, k = 1, \ldots, K; l = 1, \ldots, p$ is derived as follows,

$$\sum_{i=1}^n \sum_{j=1}^r e_{ij}^k \left(z_{il} - \bar{z}_{lj}^k \right), \quad \bar{z}_{lj}^k = \frac{\sum_{i=1}^n z_{il} \exp(\beta_k' z_i) r_{ij}^k}{\sum_{i=1}^n \exp(\beta_k' z_i) r_{ij}^k} \tag{2.6}$$

For the variance estimations, a bootstrap-based method or the fisher information based on the second derivatives can be applied. In the next simulation section, these methods are compared.

Table 1: Simulation results of regression parameters with at $(\beta_{11},\beta_{12}) = (0.5, 0.2), (\beta_{21},\beta_{22}) = (0, 0.3)$

Coef	Est(SD)	SSE	CP	Est(SD)	SSE	CP	Est(SD)	SSE	CP
				$(p_1, p_2, p_0) = (0.35, 0.23, 0.42)$					
$n = 200$	EM			Midpoint			intccr		
β_{11}	0.509 (0.122)	0.123	0.943	0.507 (0.130)	0.123	0.940	0.572 (0.136)	0.171	0.976
β_{12}	0.219 (0.244)	0.241	0.953	0.173 (0.247)	0.241	0.970	0.153 (0.314)	0.315	0.970
β_{21}	−0.008 (0.161)	0.154	0.933	−0.119 (0.160)	0.146	0.843	−0.175 (0.161)	0.191	0.896
β_{22}	0.323 (0.316)	0.309	0.953	0.278 (0.309)	0.299	0.947	0.277 (0.366)	0.371	0.960
$n = 400$									
β_{11}	0.515 (0.084)	0.086	0.950	0.493 (0.092)	0.087	0.940	0.580(0.105)	0.116	0.923
β_{12}	0.195 (0.162)	0.166	0.937	0.147 (0.164)	0.170	0.950	0.172(0.210)	0.215	0.956
β_{21}	0.0005 (0.105)	0.107	0.950	−0.120 (0.188)	0.102	0.753	−0.169(0.106)	0.127	0.830
β_{22}	0.308 (0.209)	0.211	0.950	0.259 (0.201)	0.210	0.956	0.272(0.227)	0.249	0.980
				$(p_1, p_2, p_0) = (0.50, 0.35, 0.15)$					
$n = 200$	EM			Midpoint			intcrr		
β_{11}	0.553 (0.108)	0.102	0.933	0.501 (0.107)	0.106	0.956	0.501(0.120)	0.162	0.996
β_{12}	0.216 (0.202)	0.197	0.943	0.118 (0.205)	0.200	0.933	0.073(0.267)	0.287	0.943
β_{21}	−0.011 (0.128)	0.128	0.933	−0.242(0.128)	0.117	0.450	−0.356(0.122)	0.158	0.347
β_{22}	0.316 (0.245)	0.245	0.960	0.184 (0.242)	0.241	0.913	0.167(0.300)	0.302	0.927
$n = 400$									
β_{11}	0.557 (0.072)	0.070	0.847	0.489 (0.075)	0.075	0.946	0.499 (0.087)	0.110	0.983
β_{12}	0.202 (0.147)	0.137	0.927	0.096 (0.145)	0.141	0.873	0.085 (0.187)	0.197	0.926
β_{21}	−0.011(0.097)	0.089	0.927	−0.235 (0.080)	0.082	0.200	−0.365 (0.092)	0.108	0.067
β_{22}	0.337 (0.183)	0.169	0.917	0.200 (0.170)	0.169	0.923	0.156 (0.187)	0.206	0.938

Est: the mean of $\hat{\beta}$; SD: the standard deviation of $\hat{\beta}$; SSE:mean of standard errors; CP: 95% coverage probability. (p_1, p_2, p_0)=(% of cause 1 event, % of cause 2, % of right censoring)

3. Simulation

Simulation studies are done to assess the performance of the suggested methodology in a finite sample size. We consider two causes of failure and two covariates, one is a continuous variable $Z_1 \sim N(0, 0.5^2)$ and the other is a binary one $Z_2 \sim \text{Ber}(0.5)$. Two event times (T_1, T_2) are generated from the following proportional hazard cumulative incidence function,

$$\log\{-\log(1 - F_k(t \mid \theta_k))\} = \phi_k(t) + \beta_{k1}Z_1 + \beta_{k2}Z_2, \quad k = 1, 2,$$

where $\exp[\phi_k(t)] = \eta_k(1-\exp(-\alpha_k t))/\alpha_k$, where $\exp(\phi_k(t))$ reflects the effects of the baseline incidence function and $(\eta_1, \alpha_1) = (0.5, 0.4)$ and $(\eta_2, \alpha_2) = (0.75, 0.5)$ are adopted by following Park *et al.* (2019). We set $\beta_1 = (\beta_{11},\beta_{12}) = (0.5, 0.2)$ for cause 1 and $\beta_2 = (\beta_{21},\beta_{22}) = (0, 0.3)$ for cause 2. Then define $\tilde{T} = \min(T_1, T_2)$. For making intervals, generate $\{b_1, \ldots, b_m\}$ from a uniform distribution $b_l \sim \text{Unif}(0.05, 0.1)$ and set (L_i, U_i) satisfying with $L_i = \sum_{l=1}^{s-1} b_l < \tilde{T}_i \leq \sum_l^s b_s = U_i$, where m is determined according to censoring rates.

We consider two cases for event proportions and censoring rates; $(p_1, p_2, p_0) = (0.35, 0.23, 0.42)$ and $(0.50, 0.35, 0.15)$, where $p_k = F_k(\infty) = P(\eta = k), k = 1, 2$ and $p_0 = 1 - p_1 - p_2$ represents the right censoring rate. Table 1 summarizes the simulation results of β_1 and β_2 obtained from three methods based on 300 replicates with two sample sizes $n = 200$ and $n = 400$. According to the simulation results, the suggested EM-based method gives unbiased estimates compared to the cmprsk package with midpoint and the method of intccr package when two causes have similar proportions and right censoring is high. Also, the presented standard errors of estimates (SSE) are derived from $B = 30$ bootstrap samples. These are similar with the one based on Fisher information values and

Table 2: Summary of the simulated HIV-ART dataset in *intcrr* package

Cause(n)	CD4 (median)	Age(median)	Sex(male,female)
Loss to care (n = 697)	259.51	43.23	(460,237)
Death (n = 149)	206.61	57.80	(106,43)
In care (n = 2154)	219.38	53.05	(1570,644)

Table 3: Estimates of simulated HIV-ART dataset

	Cause	Loss to care			Death		
		β	SE	p-value	β	SE	p-value
EM-based	cd4	0.0028	0.0003	<0.001	−0.0020	0.0003	<0.001
	age	−0.0206	0.0018	<0.001	0.0146	0.0018	<0.001
	sex	0.1903	0.0726	<0.001	−0.0528	0.0753	0.483
Midpoint	cd4	0.0028	0.0003	<0.001	−0.0022	0.0009	0.017
	age	−0.0209	0.0020	<0.001	0.0182	0.0045	<0.001
	sex	0.1899	0.0803	0.018	−0.1034	0.1808	0.570
Intcrr	cd4	0.0029	0.0005	<0.001	−0.010	0.0018	<0.001
	age	−0.0139	0.0022	<0.001	0.0089	0.0038	<0.001
	sex	0.0498	0.6312	0.1037	−0.030	0.0762	0.694

give similar coverage probabilities which not represented in this draft. These SSEs show very similar with the empirical variances (SD) thus the empirical coverage probabilities are close to the nominal level 0.95. However, when the occurrence of cause 1 is close to 50%, $\hat{\beta}_{11}$ of the cause 1 seem to be overestimated at both sample sizes while the estimates based on midpoint method and intccr one are close to the true value. At all cases, these two methods give serious biases in $\beta_{21} = 0$ and result in lower coverage probabilities. Standard errors (SSE) has larger values at a high right censoring rate ($p_0 = 0.42$) compared with a lower right censoring rate ($p_0 = 0.15$). Compared with intccr package, the suggested EM algorithm has a faster calculation time.

4. Data analysis

The suggested method is applied to the dataset presented in the intccr package (Park *et al.*, 2019) and another HIV dataset from BMA (Hudgens *et al.*, 2001).

4.1. HIV data set from intcrrr Packages

Bakoyannis *et al.* (2017) proposed the semiparametric transformation model and analyzed the HIV-care dataset with two competing events (loss to care and death). Patients in care are supposed to visit the clinic within a 3-month period and no appearance within a 3-month period is regarded as a loss to care. Patients with loss to care would not have a ART (antiretroviral therapy) thus have increased HIV-related mortality and morbidity. While the dataset we analyze is not same with the real data at Bakoyannis *et al.* (2017), the used variables and proportion are similar. The description of variables is presented in Table 2 and a right censoring (In care) has a very high rate (71.8%). For comparing loss to care with death, a younger with a higher CD4 tends to loss to care. Table 3 shows the result of three regression models, (i) ciregic function from intccr package, (ii) crr function from cmprsk package with a midpoint imputation for interval-censored data and (iii) the suggested EM-based method. Also, 50 bootstrap samples are generated for obtaining the standard error B.SE. According to the results in Table 3, young male patients with higher CD4 are more like to lose care. The difference between the ciregic function and EM-based method is the significance of gender effect on the loss to care. That is, the suggested method show significant effect but the former gives insignificant one (p-value =

Table 4: BMA HIV study

	cause	Subtype B			Subtype E		
		β	SE	p-value	β	SE	p-value
EM-based	Gender	0.075	0.501	0.879	0.532	0.388	0.170
	Age	−0.029	0.021	0.169	-0.051	0.013	<0.0001
	No. of needle Sharing	0.397	0.095	<0.001	0.403	0.074	<0.0001
	Jail(Yes(1)/No(0))	0.529	0.346	0.126	0.531	0.2354	0.024
Midpoint	Gender	−0.008	0.746	0.990	0.512	0.331	0.120
	Age	−0.025	0.020	0.220	−0.049	0.013	0.0001
	No. of needle Sharing	0.347	0.091	0.0001	0.395	0.068	<0.0001
	Jail(Yes(1)/No(0))	0.492	0.476	0.300	0.535	0.239	0.025
Intccr	Gender(female = 1)	−0.787	0.418	0.074	0.338	0.168	0.060
	Age	−0.015	0.022	0.181	−0.021	0.009	0.020
	No. of needle Sharing	0.068	0.120	0.549	0.179	0.093	<0.001
	Jail(Yes(1)/No(0))	0.395	0.293	0.186	0.234	0.156	0.013

0.1037). For death event with a small proportion, three methods show similar results. Older patients with lower CD4 counts tend to have a higher risk of death while the gender effect is insignificant factor on death.

4.2. HIV-seronegative injecting drug users dataset

In order to illustrate the suggested method, we analyze another type of HIV vaccine study. This project was established to assess the feasibility of the vaccine to HIV in the injecting drug users (IDU) in Bangkok, Thailand. 1209 HIV seronegative IDU were enrolled and they were supposed to visit about every four months for counseling and assessment of HIV seroconversion. 1124 people had at least one visit and among them, 133 had been diagnosed with HIV seroconversion. In this study, two subtype strains such as subtype B strain and subtype E strain can be occurred. In detail, of the 133 converts, 27 and 99 subjects have been reported to have subtype B and subtype E, respectively, and the remaining seven patients' subtypes were unknown. In our study, patients with missing cause are excluded. For 1117 patients, average age was 31.43 and the numbers of female patients and male ones are 76 and 1041 and among them, 14 female ones and 112 male ones are seroconversed. Hudgens *et al.* (2014) investigated the effect of the needle sharing on the increasing proportion of subtype E. In this analysis, we used the number of sharing needles recorded at the last visiting times as covariate. In addition, three other covariates are included in the model (gender (male(0), female(1)), age and whether he(she) has experienced an imprisonment). According to Table 4, for gender and number of needle sharing, the suggested EM-based method and the midpoint approach have similar result but are different with ones from the **intccr** package. Based on the suggested method, the number of sharing needles is very significant on the incidence of two subtypes. Also young patients with the experience of imprisonment are likely to be more susceptible to subtype E transmission.

5. Discussion

In this paper, the approximate likelihood for interval-censored data (Goetghebeur and Ryan, 2000) is extended to the competing risk data and the EM-based inference is applied. In a simulation study, the performance of the suggested method is compared with two methods; the competing risk with midpoint method and a semiparametric model from intccr package under two type of proportions. According to simulation results, the estimates from the suggested method are close to the true values at almost cases and the bootstrap standard errors are close to the standard deviation of estimates. Un-

der a low right censoring rate, however, $\hat{\beta}_{11}$ of a cause 1 from the suggested method is overestimated but the estimates of $\beta_{21} = 0$ of cause 2 from midpoint and intccr package are underestimated. Also, the computing times of the suggested method are faster compared with the intccr package. As the extension of the suggested method, the application of the smoothing technique would be applied to make predictive models. Li (2016) and Bakoyannis *et al.* (2017) applied a monotone *B*-spline to estimate a baseline function $\phi_k(t)$ to simultaneously estimate β_k but their methods require the specification of suitable knots and provide unstable estimates under some cases according to our simulation result.

As the related future study, predictive accuracy measures such as AUC and C-index are helpful to evaluate the models under interval censored competing risk data. By extending a binary classification to time to event, Heagerty *et al.* (2000) provided time-dependent ROC curves for survival data and suggested diverse definitions of sensitivity and specificity to reflect the disease status changing over time. Among them, the cumulative sensitivity and the dynamic specificity have been commonly applied and defined as follows,

$$Se^C(t,q) = \Pr(M > q \mid T \le t), \quad Sp^D(t,q) = \Pr(M \le q \mid T > t),$$

Then a time-dependent AUC(t) is defined as

$$\mathrm{AUC}(t) = \Pr\left(M_i > M_j \mid T_i \le t, T_j > t\right),$$

For estimating a time-dependent AUC (area under curve) for competing risk data, Jacqmin-Gadda *et al.* (2016) considered the fully modeled-based estimator implementing the maximum likelihood estimator of $F_k(t)$ directly for the calculation of sensitivity and specificity. For following their ideas, two definitions of C-index are considered by Wolbers *et al.* (2014) as follows,

$$C = \frac{E[(M_i > M_j)\int_0^\infty (1 - F_1(s|X_j))dF_1(s|X_i)]}{E[\int_0^\infty (1 - F_1(s|X_j))dF_1(s|X_i)]} \tag{5.1}$$

$$C^* = \frac{E[(M_i > M_j)\int_0^\infty (1 - F_1(s|X_j) - F_2(s|X_j))dF_1(s|X_i)]}{E[\int_0^\infty (1 - F_1(s|X_j) - F_2(s|X_j))dF_1(s|X_i)]} \tag{5.2}$$

Our next step is to apply the time-dependent C-index to evaluate prediction model by incorporating on several smoothing techniques.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by theKorea government (NRF-2020 R1A2C1A01100755).

References

Bakoyannis G, Yu M, and Yiannoutsos CT (2017). Semiparametric regression on cumulative incidence function with interval-censored competing risks data, *Statistics in Medicine*, **36,** 3683–3707.

Do G and Kim YJ (2017). Analysis of interval censored competing risk data with missing causes of failure using pseudo values approach, *Journal of Statistical Computation and Simulation*, **87,** 631–639.

Fine JP and Gray RJ (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, **446**, 496–509.

Geskus RB (2011). Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring, *Biometrics*, **67**, 39–49.

Goetghebeur E and Ryan L (2000). Semiparametric regression analysis of interval-censored data, *Biometrics*, **56**, 1139–1144.

Heagerty PJ, Lumley T, and Pepe MS (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker, *Biometrics,* **56**, 337–344.

Hudgens MG, Li C, and Fine JP (2014). Parametric likelihood inference for interval censored competing risks data, *Biometrics,* **70**, 1–9.

Hudgens MG, Satten GA, and Longini IM (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation, *Biometrics*, **57**, 74–80.

Jacqmin-Gadda H, Blanche P, Chary E, Touraine C, and Dartigues FG (2016). Receiver operating characteristic curve estimation for time-to-event with semicompeting risks and interval censoring, *Statistical Methods in Medical Research*, **25**, 2750–2766.

Kim YJ (2017). Analysis of interval censored failure time data with competing risk, *Journal of Applied Statistics*, **44**, 2778–2790.

Kim S and Kim YJ (2016). Regression analysis of interval censored competing risk data using a pseudo-value approach, *Communications in Statistical Applications and Methods*, **23**, 555–562.

Klein JP and Andersen PK (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function, *Biometrics*, **61**, 223–229.

Li C (2016). The Fine–Gray model under interval censored competing risk data, *Journal of Multivariate Analysis*, **143**, 327–344.

Lindsey J and Ryan L (1998). Methods for interval censored data. Tutorial in biostatistics, *Statistics in Medicine*, **17**, 219–138.

Mao L, Lin DY, and Zeng D (2017). Semiparametric regression analysis of interval-censored competing risk data, *Biometrics*, **73**, 856–865.

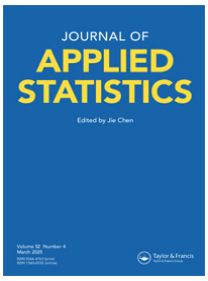
Park J, Bakoyannis G, and Andyiannoutsos CT (2019). Semiparametric competing risks regression under interval censoring using the R package intccr, *Computer Methods and Programs in Biomedicine*, **173**, 167–176.

Sun J (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York.

Turnbull BW (1976). The empirical distribution with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society: Series B*, **38**, 290–295.

Wolbers M, Blanche PMT, Koller MT, Witteman JCM, and Gerds TA (2014). Concordance for prognostic models with competing risks, *Biostatistics,* **15**, 526–539.

Received May 20, 2024; Revised October 14, 2024; Accepted December 26, 2024



Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cjas20

Quantile regression model for interval-censored data with competing risks

Amirah Afiqah binti Che Ramli & Yang-Jin Kim

To cite this article: Amirah Afiqah binti Che Ramli & Yang-Jin Kim (15 Mar 2025): Quantile regression model for interval-censored data with competing risks, Journal of Applied Statistics, DOI: 10.1080/02664763.2025.2474627

To link to this article: https://doi.org/10.1080/02664763.2025.2474627

Published online: 15 Mar 2025.

Submit your article to this journal

Article views: 24

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=cjas20

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2025.2474627

Taylor & Francis
Taylor & Francis Group

Check for updates

Quantile regression model for interval-censored data with competing risks

Amirah Afiqah binti Che Ramli and Yang-Jin Kim

Department of Statistics, Research Institute of Natural Science, Sookmyung Women's University, Seoul, Korea

ABSTRACT

Our interest is to provide the methodology for estimating quantile regression model for interval-censored competing risk data. Lee and Kim [*Analysis of interval censored competing risk data via nonparametric multiple imputation*. Stat. Biopharm. Res. 13 (2020), pp. 367–374.] applied a censoring complete data concept suggested by Ruan and Gray [*Analyses of cumulative incidence function via nonparametric multiple imputation*. Sta. Med. 27 (2008), pp. 5709–5724.] to recover a missing information related with competing events. In this paper, we also applied it to a quantile regression model. The simulated censoring times of the competing events are generated with a multiple imputation technique and the survival function of right censoring times. The performance of suggested methods is evaluated by comparing with the result of a simple imputation method under several distributions and sample sizes. The AIDS dataset is analyzed to estimate the effect of several covariates on the quantiles of cause-specific CIF as a real data analysis.

ARTICLE HISTORY
Received 18 May 2024
Accepted 24 February 2025

KEYWORDS
Interval-censored data; competing risk; quantile regression; multiple imputation

2020 MATHEMATICS SUBJECT CLASSIFICATION
62N02

1. Introduction

Quantile regression has been widely applied in diverse fields such as econometrics, epidemiology and public health. Instead of using the mean, several quantile-based statistics such as the median, the quartiles and percentile difference provide more insight by reflecting the characteristic of a population. For investigating the effect of covariates $\tilde{Z} = (1, Z)$ on quantile of response variable $\tilde{T} \sim F$, Koenker and Basset [8] introduced the following quantile regression model:

$$Q(\tau|Z) = \beta' Z,$$

where $Q(\tau|Z) = \inf\{t; F(t|Z) \geq \tau\}$ is the τ th conditional quantile of $\tilde{T}$ given Z and the corresponding objective function is defined as

$$S_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(\tilde{T}_i - \beta' Z_i), \tag{1}$$

CONTACT Yang-Jin Kim yjin@sookmyung.ac.kr Department of Statistics, Sookmyung Women's University, Yongsan-Gu, Seoul 01432, Korea

© 2025 Informa UK Limited, trading as Taylor & Francis Group

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the quantile loss function. A minimizer of (1.1) is understood as a root of the following estimating equation:

$$D_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} Z_i\{\tau - I(\tilde{T}_i < \beta' Z_i)\}.$$

For a failure time data with right censoring (T_i, δ_i), define $T_i = \min(\tilde{T}_i, C_i)$ and $\delta_i = I(\tilde{T}_i < C_i)$ where G is a survival function of censoring time, $\Pr(C_i \le c|Z) = G(c|Z)$. Several methods for dealing with censoring have been suggested by Portnoy [14], Peng and Huang [13], Wang and Wang [18] and so on. Pornoy [14] and Wang and Wang [16] utilized Efron [1]'s missing information principle (MIP) which redistributes the mass of censored observation to the uncensored ones to the right. Portnoy [14] proposed a linear programming to obtain the weights at censored observations under a global assumption and Wang and Wang [16] suggested a quantile regression model based on a locally weighted Kaplan–Meier estimator thus the weights are more easily calculated in one step. In more detail, define a weight function: $W_i(F) = 1$ for $\delta_i = 1$ or $F(C_i|Z_i) \ge \tau$ and $W_i(F) = \frac{\tau - F(C_i|Z_i)}{1 - F(C_i|Z_i)}$ for $\delta_i = 0$ and $F(C_i|Z_i) < \tau$. Then under a known $F(t|Z)$, the quantile coefficient is estimated by minimizing the following weighted objective function:

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(W_i(F)\rho_\tau(T_i - \beta' Z_i) + (1 - W_i(F))\,\rho_\tau(T^* - \beta' Z_i)\right),$$

where T^* is set as a very large value. As the gradient of $Q_n(\beta)$ with respect to β, the estimating equation is expressed as follows:

$$S_n(\beta, \tau) = \sum_{i=1}^{n} Z_i\left(\tau - \frac{I(T_i \le \beta' Z_i)\delta_i}{\hat{G}(T_i|Z_i)}\right).$$

For event history studies, a failure sometimes occurs with several causes. The available data is composed of $\{T, \epsilon, Z\}$ where ϵ denotes a cause indicator and a corresponding cause-specific cumulative incidence function (CIF) is denoted as $F_k(t|Z) = \Pr(T \le t, \epsilon = k|Z)$. The most widely applied approach for analyzing competing risk data is a cause-specific subdistribution hazard model (Fine and Gray [2]) which is also extended to the effect of covariates on the cumulative incidence function. For a quantile regression model $Q_k(\tau|Z) = \inf\{t; F_k(t|Z) \ge \tau\}$ of competing risk data, Peng and Fine [12] proposed a transformed regression model based on the AFT model and Li *et al.* [10] extended Wang and Wang [16]'s local weighting procedure to apply the re-distributed method, along with a variable selection technique.

In this paper, we extend a multiple imputation technique proposed in Lee and Kim [9] to a quantile regression model for interval-censored data with competing risk. Interval-censored failure data frequently occur in observational studies and the available data is composed of a pair of monitoring times (L_i, R_i) satisfying $L_i < \tilde{T}_i < R_i$. In detail, L_i is the last observation time before the occurrence of the event and R_i is the first one after the occurrence of the event, respectively. There are several quantile regression models related with interval-censored data. Among them, Kim *et al.* [7] applied the missing information principle of Efron [1] for a median regression model of interval-censored data and

Shen [16] extended it to the left truncated data. Zhou *et al.* [19] proposed a quantile regression model and derived the asymptotic properties. Recently, Frumento [3] proposed a quantile regression model with a following estimating equation with a weight ω_i:

$$S_n(\beta, \tau) = \sum_{i=1}^{n} Z_i (\tau - \omega_i), \tag{2}$$

where (i) $\omega_i = 1$ if $L_i < R_i \leq Z_i'\beta$, (ii) $\omega_i = 0$ if $Z_i'\beta \leq L_i < R_i$ and (iii) $\frac{F(Z_i'\beta|Z_i) - F(L_i|Z_i)}{F(R_i|Z_i) - F(L_i|Z_i)}$ if $L_i < Z_i'\beta \leq R_i$. The estimation procedure is composed of two stages: (i) estimate $\hat{F}(|Z)$ by adopting a sieve method for a maximizer of the log-likelihood $\sum_{i=1}^{n} \log\{F(R_i|Z_i) - F(L_i|Z_i)$ and (ii) estimate β from Equation (2). A corresponding variance estimator is derived from a sandwich estimator and the related procedure is implemented in R package `ctqr` [3].

In Section 2, we applied a multiple imputation technique to generate the censoring times of competing events and then the resampled dataset is applied to derive estimates and variances. Section 3 presents the simulation result to evaluate the finite sample performance and compares the suggested method with the result based on a simple imputation method. The application to HIV-related dataset is presented in Section 5. Lastly, Section 6 discusses the related future studies.

2. Quantile regression model for interval-censored competing risks

Suppose that our interest is the event related with the cause 1, $\epsilon_i = 1$. In a context of right censored data, Fine and Gray [2] propose a censoring complete data where all subjects including ones with competing events are supposed to have potential censoring times. Define $R(t)$ as a risk set at t,

$$R(t) = \{j : (C_j \wedge T_j \geq t) \cup (T_j \leq t, \epsilon_j \neq 1 \cap C_j \geq t)\}.$$

However, it is impossible to observe censoring times for the subjects experiencing competing events under a random censored. For solving this problem, Fine and Gray [2] adopted an inverse probability censoring weight (IPCW) technique to reflect a probability to stay in the risk set where the probability is determined by both the competing event time and the survival function of a censoring time $G(c) = P(C > c)$. As an alternative to the IPCW method, Ruan and Gray [15] considered a potential censoring time defined as $V^* = I(\delta = 0 \text{ or } \epsilon = 1)T + I(\epsilon > 1)C$ for the subjects with competing risk events. Taylor *et al.* [17] proposed using a non-parametric procedure to impute missing event times for censored observations in the ordinary survival data setting.

However, since C is still unknown for the subjects with $\epsilon > 1$ experiencing competing events, it is regarded as missing data. Therefore, they applied a multiple imputation technique to generate the censoring times based on the NPMLE estimate of G such as a Kaplan–Meier estimator. It is assumed that censoring distribution is independent of all failure types. The merit of imputing the censoring times is a simple method in that the distribution for the censoring times is only needed but the other failure times need not be modeled when there is interest in only one failure type.

Lee and Kim [9] applied this approach to interval-censored competing risk data. Let $\{L_i, R_i, \epsilon_i, \delta_i, Z_i,\ i = 1, \ldots, n\}$ denote the available data where for a subject with $\delta_i = 0$ right

censored at L_i, set $R_i = \infty$ and $\epsilon_i = 0$. For analyzing interval censored competing risks data, Hudgens *et al.* [4] suggested two types of estimates of CIF under two different time supports. Kim [6] also considered the estimates of F_k based on the approximate likelihood under a piecewise constant assumption. Let $\{0 = s_0 < s_1 < \cdots < s_{m+1} = \infty\}$ denote the ordered time points derived from the all causes. Define λ_{jk} as the hazard function of the kth cause at s_j. Let dN_{ij}^k be an event indicator showing whether the i subject experiences the kth cause event at s_j. At risk indicator Y_{ij} represents that ith subject is still in the risk group at s_j. By using these notations, the approximate likelihood is written as follows:

$$L_a = \prod_{j=1}^{K} \prod_{k=1}^{n} \prod_{i=1} \left[\lambda_{jk}^{dN_{ij}^k} \exp\left(- \sum_{1<l<j} Y_{il} \lambda_l \right) \right]$$

where dN_{ij}^k and Y_{ij} are unknown thus the EM algorithm is applied to obtain $\hat{\lambda} = \sum_{i=1}^{n} \hat{E}(dN_{ij}^k | O_i, \hat{\lambda}) / \sum_{i=1}^{n} \sum_{l \geq j} \hat{E}(Y_{ij} | O_i, \hat{\lambda})$. Then resulting estimate is derived as follows:

$$\hat{F}_k(t) = \sum_{s_j < t} \hat{\lambda_{jk}} \left(- \sum_{k'=1}^{K} \sum_{l=1}^{j-1} \hat{\lambda}_{lk'} \right) \tag{3}$$

For extending Ruan and Gray [15]'s method to interval-censored data, a potential censoring time is defined as $C_i = I(\epsilon_i = 0) * L_i + I(\epsilon_i = 1) * R_i + I(\epsilon_i > 1) * C_i^*$, where C_i^* denotes the censoring time for the subject experiencing competing event but is unknown instead $C_i^* > R_i$ is available information [9]. Therefore, the augmented data replacing C^* is generated through a following multiple imputation scheme:

Step 1: Estimate a Kaplan–Meier estimator $\hat{G}(c)$, the survival function of the right censoring time with $\{V_i, \delta_i^c,\ i = 1, \ldots, n\}$, where $\delta_i^c = 1 - \delta_i$ and set $V_i = L_i$ for $\delta_i = 0$ and $V_i = R_i$ for $\delta_i = 1$.

Step 2: At the mth imputation step ($m = 1, \ldots, M$),

(i) For a subject experiencing a competing event, generate the censoring time $C_i^{*(m)}$ from $\hat{G}$ satisfying $\Pr(C_i^{*(m)} > t | C_i^{*(m)} > R_i) = \frac{\hat{G}(t)}{\hat{G}(R_i)}$ for $t > R_i$.

(ii) With $C_i^{*(m)}$ sampled in (i), the mth imputed dataset is set as follows: $(L_i^{*(m)}, R_i^{*(m)}, \delta_i^{*(m)})$, where for subject with $\delta_i = 0$, set $L_i^{*(m)} = L_i$, $R_i^{*(m)} = \infty$ and $\delta_i^{*(m)} = 0$, for one with $\epsilon_i = 1$, $L_i^{*(m)} = L_i$, $R_i^{*(m)} = R_i$ and $\delta_i^{*(m)} = 1$. Finally, for subject with competing events $\epsilon_i > 1$, $L_i^{*(m)} = C_i^{*(m)}$, $R_i^{*(m)} = \infty$ and $\delta_i^{*(m)} = 0$.

(iii) For the mth imputed dataset, the ctqr function in the R package `ctqr` is applied to obtain the estimates of the quantile regression model.

Step 3. For the mth imputed dataset, the regression coefficients $\hat{\beta}^{(m)}$ and the corresponding variances $\hat{I}^{-1(m)}$ are obtained.

Step 4. The final estimates and variances are computed as follows:

$$\bar{\hat{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)},$$

$$\hat{Var}(\hat{\beta}) = \frac{1}{M}\sum_{m=1}^{M}\hat{I}^{-1(m)} + \left(1+\frac{1}{M}\right)\sum_{j=1}^{M}\frac{(\hat{\beta}^{(m)}-\bar{\hat{\beta}})(\hat{\beta}_{\tau}^{(m)}-\bar{\hat{\beta}})'}{M-1}.$$

3. Simulation

Simulation studies were conducted to evaluate the finite sample performance of the suggested method. 1000 simulated datasets were generated with sample size of 300 and 500. Two covariates were considered where one is continuous with $Z_1 \sim Unif(0,1)$ and the other is binary $Z_2 \sim Ber(0.5)$. Therefore, the covariate vector is defined as $Z = (1, Z_1, Z_2)$ and each covariate is assumed to have varying effects on cumulative incidence function. Two causes of failures are assumed and the event times (T_1, T_2) are generated from $\Pr(\epsilon = 1|Z) = p_0 I(Z_2 = 0) + p_1 I(Z_2 = 1)$, $\Pr(T \le t|\epsilon = 1, Z) = \Phi(\log t - \gamma' Z)$ and $\Pr(T \le t|\epsilon = 2, Z) = \Phi(\log t - \alpha' Z)$ with $p_0 = 0.8, p_1 = 0.6, \gamma = (0.0, 0.5, 0.5)'$ and $\alpha = (-0.1, 0.5, -0.5)'$. Then resulting quantile regression model of $\epsilon = 1$ is expressed as

$$\log Q_1(\tau|Z) = \gamma_1 + \Phi^{-1}\left(\frac{\tau}{p_0}\right) + \gamma_2 Z_1 + \left(\gamma_3 + \Phi^{-1}\left(\frac{\tau}{p_1}\right) - \Phi^{-1}\left(\frac{\tau}{p_0}\right)\right) Z_2, \quad (4)$$

where Φ denotes the cumulative distribution function of four distributions: (i) normal, (ii) logistic, (iii) Cauchy and (iv) t-distribution, respectively. From (4), while Z_1 have a constant effect on $Q_1(\tau|Z)$ over τ while the effects of intercept and Z_2 vary according to values of τ and Φ. Then $\beta_0 = \gamma_1 + \Phi^{-1}(\frac{\tau}{p_0})$, $\beta_1 = \gamma_2$ and $\beta_2 = (\gamma_3 + \Phi^{-1}(\frac{\tau}{p_1}) - \Phi^{-1}(\frac{\tau}{p_0}))$. For evaluating the proposed method, the complete censoring method and the midpoint approach using a midpoint of each interval are compared. At the complete censored method, the censoring times of competing events are assumed to be known and are applied for the analysis instead of interval-censored data. Therefore the quantile regression model for interval censored data is directly applied irrespective of interval-censored data of competing events.

Tables 1–4 show the simulation results for $(\beta_0, \beta_1, \beta_2)$ at three quantile levels $\tau = (0.1, 0.2, 0.4)$ and two sample sizes under four distributions. Each table contains the mean of estimated parameters (Est), average of the standard errors (ASE), empirical standard error (ESE) and empirical coverage probabilities (CP).

According to results, the suggested multiple imputation method and the complete censoring method have similar result. At $\tau = 0.2$ and $\tau = 0.4$, the two methods give almost unbiased estimates and reasonable standard errors along with coverage probabilities close to the nominal level. For $\tau = 0.1$, the estimates show serious biases at logistic and Cauchy distribution while the results are suitable at normal and t-distributions. The results of a midpoint approach show that estimates are underestimated at lower quantile and are overestimated at upper quantile at all distributions. The proposed method demonstrates a suitable result at small sample size and several distributions but results in biases at lower quantile levels.

Table 1. Simulation result for normal distribution

	Multiple imputation					Complete censoring				Midpoint			
	True	Est	ASE	ESE	Cp	Est	ASE	ESE	Cp	Est	ASE	ESE	Cp
$\tau = 0.1, n = 300$													
β_0	−1.05	−0.96	0.22	0.24	0.82	−0.95	0.23	0.25	0.80	0.31	0.15	0.13	0.00
β_1	0.50	0.46	0.34	0.36	0.92	0.45	0.37	0.38	0.90	0.31	0.27	0.24	0.88
β_2	0.68	0.63	0.20	0.21	0.90	0.63	0.21	0.22	0.89	0.45	0.18	0.17	0.69
$\tau = 0.1, n = 500$													
β_0	−1.05	−0.94	0.17	0.18	0.79	−0.96	0.16	0.18	0.81	−0.31	0.11	0.10	0.00
β_1	0.50	0.43	0.27	0.28	0.92	0.46	0.26	0.27	0.93	0.30	0.21	0.19	0.80
β_2	0.68	0.62	0.16	0.17	0.88	0.64	0.16	0.17	0.90	0.45	0.13	0.13	0.54
$\tau = 0.2, n = 300$													
β_0	−0.57	−0.55	0.19	0.20	0.91	−0.53	0.18	0.20	0.88	0.51	0.18	0.20	0.00
β_1	0.50	0.50	0.31	0.31	0.92	0.48	0.32	0.33	0.93	0.51	0.37	0.41	0.91
β_2	0.74	0.75	0.19	0.19	0.94	0.75	0.19	0.21	0.92	0.80	0.25	0.27	0.90
$\tau = 0.2, n = 500$													
β_0	−0.57	−0.57	0.16	0.15	0.92	−0.57	0.14	0.15	0.92	0.51	0.13	0.14	0.00
β_1	0.50	0.49	0.26	0.24	0.93	0.52	0.24	0.25	0.93	0.47	0.27	0.30	0.93
β_2	0.74	0.75	0.16	0.15	0.94	0.76	0.15	0.16	0.93	0.80	0.19	0.21	0.91
$\tau = 0.4, n = 300$													
β_0	0.10	0.12	0.22	0.23	0.92	0.14	0.23	0.24	0.89	1.01	0.39	0.43	0.43
β_1	0.50	0.53	0.38	0.36	0.93	0.50	0.40	0.40	0.91	0.97	0.84	0.95	0.92
β_2	0.93	0.98	0.25	0.22	0.94	0.97	0.25	0.25	0.93	2.26	0.74	0.86	0.66
$\tau = 0.4, n = 500$													
β_0	0.10	0.10	0.18	0.17	0.92	0.12	0.17	0.16	0.93	1.00	0.27	0.30	0.11
β_1	0.50	0.50	0.28	0.27	0.93	0.54	0.28	0.29	0.93	0.92	0.60	0.65	0.88
β_2	0.93	0.94	0.18	0.17	0.92	0.98	0.19	0.20	0.94	2.22	0.60	0.67	0.49

Table 2. Simulation result for Logistic distribution.

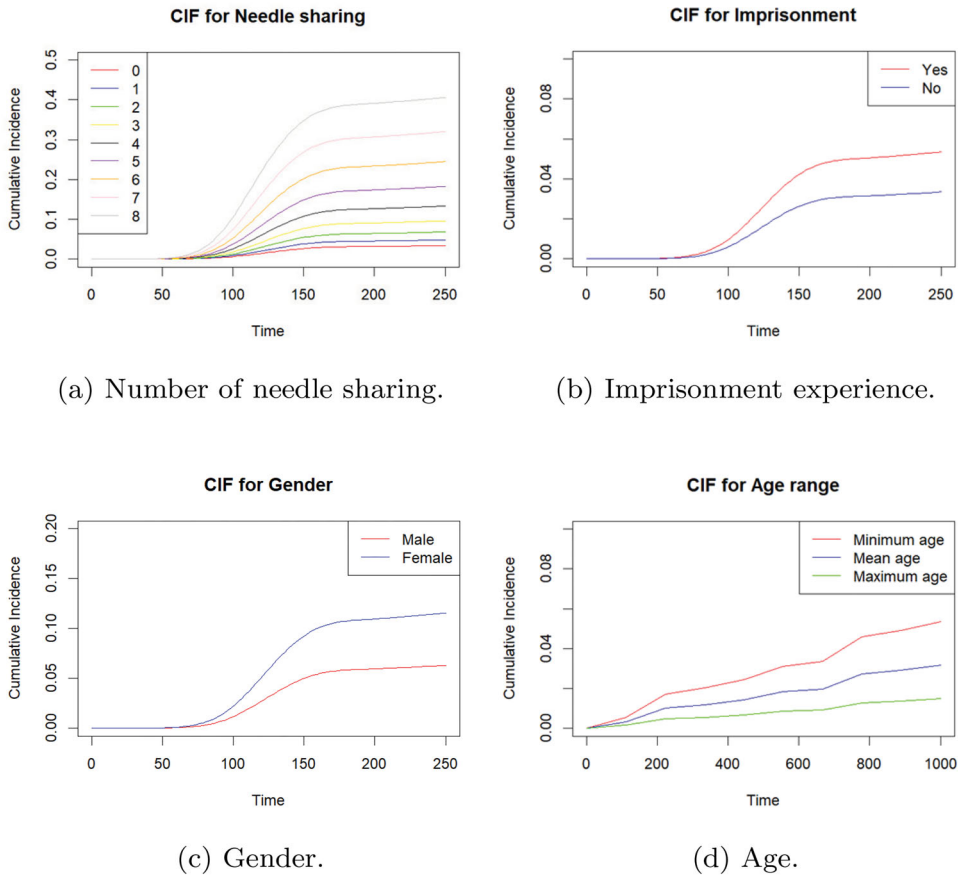
	Multiple imputation					Complete censoring				Midpoint			
	True	Est	ASE	ESE	Cp	Est	ASE	ESE	Cp	Est	ASE	ESE	Cp
$\tau = 0.1, n = 300$													
β_0	−1.85	−1.52	0.25	0.24	0.73	−1.52	0.25	0.25	0.74	0.15	0.11	0.11	0.00
β_1	0.50	0.32	0.46	0.45	0.92	0.33	0.45	0.44	0.92	0.12	0.23	0.19	0.56
β_2	0.84	0.57	0.30	0.30	0.83	0.61	0.30	0.31	0.83	0.27	0.16	0.16	0.14
$\tau = 0.1, n = 500$													
β_0	−1.85	−1.56	0.19	0.17	0.70	−1.54	0.19	0.19	0.67	0.15	0.07	0.06	0.0
β_1	0.50	0.35	0.37	0.35	0.93	0.31	0.36	0.26	0.89	0.12	0.17	0.14	0.40
β_2	0.84	0.64	0.24	0.25	0.83	0.64	0.25	0.26	0.85	0.28	0.12	0.12	0.05
$\tau = 0.2, n = 300$													
β_0	**−1.00**	−0.94	0.32	0.35	0.88	−0.95	0.32	0.33	0.88	0.34	0.19	0.21	0.00
β_1	0.50	0.47	0.54	0.57	0.92	0.50	0.52	0.51	0.93	0.33	0.38	0.43	0.92
β_2	0.91	0.83	0.33	0.32	0.93	0.86	0.32	0.31	0.93	0.69	0.30	0.34	0.84
$\tau = 0.2, n = 500$													
β_0	−1.00	−0.98	0.28	0.26	0.88	−0.95	0.28	0.29	0.88	0.33	0.15	0.17	0.00
β_1	0.50	0.51	0.43	0.43	0.93	0.50	0.44	0.45	0.92	0.32	0.28	0.33	0.92
β_2	0.91	0.90	0.26	0.25	0.94	0.88	0.26	0.25	0.94	0.70	0.24	0.27	0.82
$\tau = 0.4, n = 300$													
β_0	0.10	0.17	0.39	0.40	0.91	0.13	0.40	0.38	0.95	1.03	0.59	0.59	0.67
β_1	0.50	0.51	0.69	0.69	0.94	0.57	0.67	0.63	0.95	0.95	1.21	1.26	0.86
β_2	1.19	1.16	0.46	0.35	0.96	1.18	0.44	0.36	0.96	3.22	1.35	1.39	0.64
$\tau = 0.4, n = 500$													
β_0	0.10	0.13	0.26	0.28	0.93	0.14	0.32	0.33	0.92	1.00	0.45	0.48	0.57
β_1	0.50	0.50	0.48	0.50	0.95	0.56	0.56	0.56	0.95	0.97	0.94	1.02	0.88
β_2	1.19	1.20	0.27	0.31	0.95	1.24	0.37	0.29	0.97	3.34	1.13	1.25	0.60

Table 3. Simulation result for Cauchy distribution.

	Multiple imputation					Complete censoring				Midpoint			
	True	Est	ASE	ESE	Cp	Est	ASE	ESE	Cp	Est	ASE	ESE	Cp
$\tau = 0.1, n = 300$													
β_0	−2.31	−1.49	0.32	0.35	0.30	−1.53	0.35	0.36	0.38	0.15	0.13	0.10	0.0
β_1	0.50	0.30	0.61	0.65	0.88	0.29	0.62	0.64	0.89	0.09	0.27	0.22	0.55
β_2	1.18	0.62	0.38	0.40	0.65	1.18	0.68	0.39	0.42	0.25	0.21	0.20	0.05
$\tau = 0.1, n = 500$													
β_0	−2.31	−1.54	0.25	0.21	0.12	−1.55	0.25	0.26	0.14	0.14	0.08	0.07	0.00
β_1	0.50	0.26	0.46	0.50	0.89	0.28	0.47	0.48	0.92	0.09	0.19	0.16	0.39
β_2	1.18	0.70	0.31	0.36	0.67	0.66	0.32	0.36	0.63	0.27	0.17	0.16	0.02
$\tau = 0.2, n = 300$													
β_0	−0.90	−0.81	0.37	0.34	0.85	−0.81	0.36	0.37	0.86	0.35	0.23	0.24	0.00
β_1	0.50	0.48	0.56	0.54	0.91	0.50	0.54	0.54	0.91	0.37	0.46	0.48	0.89
β_2	0.92	0.87	0.31	0.32	0.91	0.88	0.33	0.33	0.93	0.75	0.34	0.35	0.86
$\tau = 0.2, n = 500$													
β_0	−0.90	−0.83	0.29	0.28	0.89	−0.78	0.28	0.28	0.86	0.35	0.17	0.19	0.00
β_1	0.50	0.52	0.46	0.43	0.91	0.50	0.47	0.46	0.94	0.39	0.34	0.37	0.92
β_2	0.92	0.89	0.26	0.25	0.94	0.86	0.25	0.24	0.93	0.78	0.26	0.27	0.88
$\tau = 0.4, n = 300$													
β_0	0.10	0.15	0.31	0.31	0.92	0.17	0.37	0.36	0.94	1.03	0.54	0.53	0.57
β_1	0.50	0.49	0.53	0.53	0.94	0.53	0.62	0.60	0.96	0.90	1.06	1.12	0.87
β_2	1.08	1.07	0.30	0.35	0.95	1.09	0.39	0.30	0.97	2.75	1.33	1.30	0.67
$\tau = 0.4, n = 500$													
β_0	0.10	0.15	0.24	0.23	0.92	0.17	0.27	0.27	0.93	1.00	0.37	0.39	0.32
β_1	0.50	0.49	0.43	0.41	0.93	0.53	0.47	0.46	0.95	0.97	0.78	0.84	0.89
β_2	1.08	1.10	0.26	0.27	0.94	1.15	0.33	0.27	0.97	2.87	1.02	1.14	0.67

Table 4. Simulation result for t-distribution.

	Multiple imputation					Complete censoring				Midpoint			
	True	Est	ASE	ESE	Cp	Est	ASE	ESE	Cp	Est	ASE	ESE	Cp
$\tau = 0.1, n = 300$													
β_0	−1.32	−1.17	0.29	0.30	0.86	−1.19	0.28	0.29	0.84	0.24	0.17	0.15	0.01
β_1	0.50	0.41	0.40	0.40	0.92	0.44	0.46	0.46	0.94	0.24	0.32	0.29	0.83
β_2	0.77	0.79	0.24	0.23	0.94	0.28	0.30	0.90	0.77	0.43	0.21	0.19	0.55
$\tau = 0.1, n = 500$													
β_0	−1.32	−1.22	0.20	0.23	0.87	−1.21	0.19	0.21	0.86	0.24	0.13	0.12	0.00
β_1	0.50	0.41	0.36	0.37	0.91	0.43	0.37	0.36	0.92	0.24	0.25	0.22	0.82
β_2	0.77	0.73	0.21	0.22	0.94	0.73	0.21	0.24	0.91	0.38	0.16	0.15	0.34
$\tau = 0.2, n = 300$													
β_0	−0.66	−0.63	0.23	0.23	0.88	−0.63	0.24	0.23	0.92	0.47	0.19	0.21	0.02
β_1	0.50	0.51	0.38	0.38	0.92	0.50	0.38	0.37	0.94	0.37	0.37	0.42	0.94
β_2	0.79	0.80	0.22	0.23	0.93	0.79	0.24	0.23	0.94	0.78	0.28	0.32	0.92
$\tau = 0.2, n = 500$													
β_0	−0.66	−0.64	0.20	0.19	0.89	−0.64	0.19	0.20	0.92	0.45	0.16	0.17	0.00
β_1	0.50	0.49	0.32	0.31	0.94	0.52	0.32	0.32	0.93	0.47	0.31	0.34	0.94
β_2	0.79	0.78	0.18	0.18	0.94	0.79	0.18	0.18	0.93	0.79	0.22	0.24	0.93
$\tau = 0.4, n = 300$													
β_0	0.10	0.12	0.26	0.25	0.94	0.13	0.25	0.23	0.92	1.02	0.39	0.44	0.46
β_1	0.50	0.51	0.44	0.45	0.94	0.54	0.44	0.39	0.94	0.94	0.80	0.94	0.90
β_2	0.98	1.00	0.30	0.27	0.95	1.04	0.29	0.27	0.94	2.44	0.93	1.09	0.72
$\tau = 0.4, n = 500$													
β_0	0.10	0.12	0.19	0.18	0.92	0.133	0.19	0.19	0.92	1.00	0.33	0.37	0.29
β_1	0.50	0.49	0.33	0.32	0.93	0.54	0.35	0.35	0.93	0.94	0.69	0.76	0.89
β_2	0.98	0.99	0.20	0.20	0.94	1.04	0.23	0.22	0.95	2.42	0.70	0.85	0.60



(a) Number of needle sharing. (b) Imprisonment experience.

(c) Gender. (d) Age.

Figure 1. NPMLE of CIFS by group in Table 5: (a) number of needle sharing, (b) imprisonment experience, (c) gender and (d) age.

When comparing the results across all four distributions, logistic distribution and Cauchy distribution show slightly lower precision and coverage probabilities compared to other distributions. Increasing the sample size generally leads to more accurate parameter estimates across different distributions. Across different quantile levels, the accuracy of estimates tends to improve as the quantile increases while larger standard errors for estimates are observed at higher quantile, especially for normal and logistic distributions with smaller sample sizes.

4. Data analysis

To illustrate the suggested method, we analyze HIV vaccine study [4,5]. This project was established to assess the feasibility of the vaccine to HIV (Human Immunodeficiency Virus) in the injecting drug users (IDU) in Bangkok, Thailand. 1209 HIV seronegative IDU were enrolled and they were supposed to visit about every 4 months for counseling and assessment of HIV seroconversion. 1124 peoples had at least one visit and among them, 133 ones had been diagnosed with HIV seroconversion. In this study, two subtype strains

Table 5. Estimation of quantile regression model of Subtype E.

		Est	Standard error	*p*-values
$\tau = 0.2$	(Intercept)	7.792	0.221	<0.001
	Gender	−0.016	0.210	0.938
	Age	−0.0004	0.003	0.900
	Needle sharing	−0.139	0.023	<0.001
	Imprisonment	−0.699	0.088	<0.001
$\tau = 0.3$	(Intercept)	8.235	0.395	<0.001
	Gender	0.207	0.330	0.531
	Age	−0.002	0.007	0.780
	Needle sharing	−0.083	0.075	0.271
	Imprisonment	−1.244	0.132	<0.001
$\tau = 0.4$	(Intercept)	8.738	0.640	<0.001
	Gender	0.561	0.514	0.279
	Age	−0.003	0.011	0.780
	Needle sharing	−0.068	0.131	0.607
	Imprisonment	−2.052	0.186	<0.001

such as subtype B strain and subtype E strain can be occurred. In detail, of the 133 converts, 27 and 99 subjects have been reported to have subtype B and subtype E, respectively, and the remaining seven patients' subtypes were unknown. In our study, patients with missing cause are excluded. For 1117 subjects, average age was 31.43 and the numbers of female patients and male ones are 76 and 1041. Among them, 14 female ones and 112 male ones are seroconverted. Hudgens *et al.* [6] investigated the effect of the needle sharing on the increasing proportion of subtype E. In this analysis, we investigate the effect of four covariates (gender (male (0); female (1)), age, the number of needle sharing and whether he(she) has experienced an imprisonment).

In Table 5, the estimated regression coefficients represent the associations between each covariate and three quantiles ($\tau = 0.2, 0.3, 0.4$) of HIV seroconversion of subtype E strain. Gender and age are insignificant on all quantiles and the number of needle sharing is only significant for patients at lower quantiles but become insignificant for higher quantiles while an imprisonment experience has a negative relation for all quantiles. That is, the coefficients of imprisonment become decreasing as τ increases and a negative value indicates that a subject with an experience of imprisonment is more likely to get subtype E as time increases. Figure 1 shows the estimated CIFs by groups based on the covariates analyzed in Table 5 using the NPMLE(2.1) [6]. Female group and imprisonment experience have higher probabilities as time increases.

5. Concluding remarks

This study introduces the usage of quantile regression for analyzing an interval-censored data with competing risks by incorporating the censoring complete data by Ruan and Gray [15] as an alternative to the traditional subdistribution hazard methods. According to the result, the suggested method and the complete censoring method have similar result and satisfy unbiasedness and variation under several distributions when the quantiles are not too lower. For a lower quantile, however, the Cauchy and logistic distribution results in biased estimates. As the reviewer pointed out, when all competing events are interest, the imputation procedure is individually applied to generate the censoring distribution. Related works can be considered to improve this procedure. However, the proposed method depends on the independent censoring assumption and a two-stage procedure of

`ctqr` packages requiring the estimation of F. It may entail much serious biases under a dependent censoring thus a suitable modeling for a dependent censoring $G(|Z)$ would be applied. The AIDS dataset contains a high rate of right censoring and may need the implementation of a risk-free group into a quantile regression model [11].

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (NRF-2020 R1A2C1A01100755).

References

[1] B. Efron, *The two-sample problem with censored data*, in: *Proceedings of the fifth Berkley symposium in mathematical statistics*, L. Le Cam, J. Neyman, eds., vol. IV, Prentice-Hall, New York, 1967, pp. 831–853.
[2] J.P. Fine and R.J. Gray, *A proportional hazards model for the subdistribution of a competing risk*, J. Am. Stat. Assoc. 94 (1999), pp. 496–509.
[3] *ctqr: Censored and truncated quantile regression*. 2021. R package version 2.0. Available from: http://CRAN.R-project.org/package=ctqr.
[4] M.G. Hudgens, G.A. Satten, and I.M. Longini, *Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation*, Biometrics 57 (2001), pp. 74–80.
[5] M.G. Hudgens, C. Li, and J.P. Fine, *Parametric likelihood inference for interval censored competing risks data*, Biometrics 70 (2014), pp. 1–9.
[6] Y-J. Kim, *Analysis of interval censored failure time data with competing risk*, J. Appl. Stat. 44 (2017), pp. 2278–2290.
[7] Y-J. Kim, H.J. Cho J. Kim, and M. Jhun, *Median regression model with interval censored data*, Biom. J. 52 (2010), pp. 201–208.
[8] R. Koenker and J.G. Bassett, *Regression quantiles*, Econometrica 46 (1978), pp. 33–50.
[9] H. Lee and Y-J. Kim, *Analysis of interval censored competing risk data via nonparametric multiple imputation*, Stat. Biopharm. Res. 13 (2020), pp. 367–374.
[10] E. Li, M. Tian, and M-L. Tang, *Variable selection in competing risks models based on quantile regression*, Stat. Med. 38 (2019), pp. 4670–4685.
[11] N. Narisetty and R. Koenker, *Censored quantile regression survival models with a cure proportion*, J. Econometr. 226 (2022), pp. 192–203.
[12] L. Peng and J.P. Fine, *Competing risks quantile regression*, J. Am. Stat. Assoc. 104 (2009), pp. 1440–1453.
[13] L. Peng and Y. Huang, *Survival analysis with quantile regression models*, J. Am. Stat. Assoc. 103 (2008), pp. 637–649.
[14] S. Portnoy, *Censored regression quantiles*, J. Am. Stat. Assoc. 98 (2003), pp. 1001–1012.
[15] P.K. Ruan and R.J. Gray, *Analyses of cumulative incidence function via non-parametric multiple imputation*, Stat. Med. 27 (2008), pp. 5709–5724.
[16] P.S. Shen, *Median regression model with left truncated and interval-censored data*, Kor. J. Stat. 42 (2013), pp. 469–479.
[17] J.M.G. Taylor, S. Murray, and C.-H. Hsu, *Survival estimation and testing via multiple imputation*, Stat. Probabil. Lett. 58 (2002), pp. 221–232.
[18] H.J. Wang and L. Wang, *Locally weighted censored quantile regression*, J. Am. Stat. Assoc. 104 (2009), pp. 1117–1128.
[19] X. Zhou, Y. Feng, and X. Du, *Quantile regression for interval censored data*, Commun. Stat. Theor. Methods. 46 (2017), pp. 3848–3863.

pISSN 1598-2998, eISSN 2005-9256
https://doi.org/10.4143/crt.2024.700 Cancer Res Treat. 2025;57(4):1019-1029

Original Article

Tracing Metastatic Evolutionary Patterns in Lung Adenocarcinoma: Prognostic Dissection Based on a Multi-state Model

Geewon Lee[1,2], **Yang-Jin Kim**[3], **Insuk Sohn**[4], **Jong Hoon Kim**[5], **Ho Yun Lee**[1,6]

[1]*Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul,* [2]*Department of Radiology and Medical Research Institute, Pusan National University Hospital, Pusan National University School of Medicine, Busan,* [3]*Department of Statistics, Sookmyung Women's University, Seoul,* [4]*Arontier Inc., Seoul,* [5]*Industrial Biomaterial Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon,* [6]*Department of Health Sciences and Technology, SAIHST, Sungkyunkwan University, Seoul, Korea*

Purpose After surgery for lung adenocarcinoma, a patient may experience various states of recurrence, with multiple factors potentially influencing the transitions between these states. Our purpose was to investigate the effects of clinical and pathological factors on tumor recurrence, death, and prognosis across various metastasizing pathways.

Materials and Methods Our study group included 335 patients with all demographic and pathologic data available who underwent surgical resection for lung adenocarcinoma for more than 10 years. The following states of disease were defined: initial state, operation (OP); three intermediate states of local recurrence (LR), metastasis (Meta), and concurrent LR with metastasis (LR+Meta); and a terminal state, death. We identified eight transitions representing various pathways of tumor progression. We employed a multi-state model (MSM) to separate the impacts of multiple prognostic factors on the transitions following surgery.

Results After surgery, approximately half of patients experienced recurrence. Specifically, 142 (42.4%), 54 (16.1%), and seven (2.1%) patients developed Meta, LR+Meta, and LR, respectively. Clinical and pathological factors associated with the transitions were different. Impact of pathological lymph node remained a risk factor for both OP to Meta (λ_{02}, p=0.001) and OP to LR+Meta (λ_{03}, p=0.001).

Conclusion Lung adenocarcinoma displays a broad spectrum of clinical scenarios even after curative surgery. Incidence, risk factors, and prognosis varied across different pathways of recurrence in lung adenocarcinoma patients. The greatest implication of this MSM is its ability to predict the timing and type of clinical intervention that will have the greatest impact on survival.

Key words Non-small cell lung carcinoma, Multi-state model, Transition intensity

Introduction

Lung cancer is the leading cause of cancer-related death worldwide in both men and women [1]. However, a large proportion of lung cancer patients develop recurrence even after curative surgery, with reported rates as high as 68%, eventually leading to increased risk of death [2-4]. Recurrence patterns after surgery are heterogeneous as tumor recurrence encompasses multiple states such as local recurrence (LR), distal metastasis (Meta), and concurrent LR and distal Meta. In fact, a patient may develop several different states and move among the states throughout the course of the disease [5]. In particular, lung adenocarcinoma displays a broad spectrum of biological variations, ranging from indolent to aggressive forms, affecting prognostic results [6]. In other words, numerous clinical scenarios are evident in lung adenocarcinoma, leading to a wide range of treatment options.

Most lung cancer trials use the traditional end points of overall survival and disease-free survival. These endpoints offer the advantages of simplicity and straightforwardness for describing patient mortality or cancer advancement. However, besides death, surviving patients may experience various states and multiple transitions between states during the course of the disease. For example, a patient may develop an ipsilateral metastatic lymph node (LN), classified as LR and receive limited treatment. Following limited treatment, the patient may be disease-free for many years, and then may develop distal Meta in other body regions, necessitating systemic therapy. Yet, this tumor trajectory is not revealed using traditional endpoints in lung adenocarcinoma. Thus, an improved understanding of the various metastasizing pathways and their risks can enhance patient care after surgery, allowing for a more personalized and effective treatment approach.

Multi-state model (MSM) analysis is used to analyze the movement of individuals between defined states over a period of time [7]. For example, MSM analysis has been used

Correspondence: Ho Yun Lee
Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Korea
Tel: 82-2-3410-2502 E-mail: hoyunlee96@gmail.com
Received July 25, 2024 **Accepted** January 22, 2025 **Published Online** January 24, 2025
*Geewon Lee and Yang-Jin Kim contributed equally to this work.

Copyright © 2025 by the Korean Cancer Association

This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

in prostate and oropharyngeal cancer studies when patients experience intermediate (transition) states corresponding to particular stages of disease [8-10]. As mentioned above, even after curative resection, the prognosis is extremely heterogenous with diverse pathways and multiple treatment options in lung adenocarcinoma. Thus, for a more detailed understanding of metastatic evolutionary patterns, we implemented MSM analysis in lung adenocarcinoma patients. The main goal of this study was to investigate the effects of clinical and pathological factors on tumor recurrence, death, and prognosis in lung adenocarcinoma patients across various metastasizing pathways.

Materials and Methods

1. Patients

Data were retrospectively collected from patients who underwent operation (OP) for lung adenocarcinoma from October 2003 to August 2011. To conduct a comprehensive analysis of long-term survival after lung cancer surgery, we continued follow-up until the date of death or study end date (June 20, 2023), whichever was earliest, resulting in data collected over a period of more than a decade.

From a thoracic surgical database that exclusively focused on lung adenocarcinoma, 340 patients who had undergone complete surgical resection for invasive lung adenocarcinoma were identified. Five patients who had incomplete pathologic analysis and histologic subtyping were excluded. Thus, our final study group included 335 patients (186 men, 149 women). Demographic variables including age, sex, smoking history, adjuvant chemotherapy, and OP type were collected by reviewing electronic medical charts. Pathological variables such as overall tumor-node-metastasis (TNM) stage, tumor size, T status (pT), and LN status (pN) were evaluated. The pathologic staging was recorded based on the 7th edition TNM staging in non–small cell lung cancer (NSCLC) patients, at the beginning of this study.

Furthermore, a comprehensive evaluation of the pathological factors including tumor differentiation, predominant histologic pattern, pleural invasion, vascular invasion, lymphatic invasion, and perineural invasion was performed.

2. Definition of disease status

Electronic medical records were reviewed for date of last follow-up, documented LR or distal Meta, or death. Medical images including computed tomography (CT), magnetic resonance imaging (MRI), bone scan, ^{18}F-fluorodeoxyglucose positron emission tomography/CT (^{18}F-FDG PET/CT), and pathologic results were used to confirm recurrences. In this study, LR was defined as tumor recurrence in the entire ipsilateral hemithorax (including the pleura and chest wall) or/and the ipsilateral mediastinal LNs [11,12]. When it came to identifying subsolid and solid nodules in the ipsilateral lung, we initially searched for pathologic results of biopsies or surgeries. If there were no pathologic results present, a comprehensive evaluation of multiple follow-up CT scans was performed. If the nodule continued to increase in size, and benign conditions such as infection or inflammation were excluded, the nodule was presumed to be recurrence from the date of initial detection on the CT scan by a thoracic radiologist. In addition, distinguishing multifocal primary lung cancers from intrapulmonary metastases is an important issue, yet can be very challenging. Especially in complex situations such as matching histology, the nodule was confirmed after extensive radiologic review and follow-up, comprehensive histologic assessment, and multidisciplinary tumor board discussion [13-16].

Any tumor recurrence occurring elsewhere than the ipsilateral hemithorax and ipsilateral LNs was considered distal Meta. S1 Table presents the information regarding confirmation of distal metastases. The coexistence of LR and distal metastasis (LR+Meta) was classified independently from LR and distal Meta. The Korean statistical information service was utilized to investigate the dates and reasons for death.

3. Statistical analysis

MSM is useful to investigate the effect of treatment on several paths from a study entry to intermediate to death. While classical methods such as log-rank test and Cox proportional hazard model focus on effect of treatment on a single event, MSM assesses the comprehensive patterns of the effects of treatment on the intermediate events resulting in the final outcome. The MSM consists of states and transitions between states. In this study, to represent several event types occurred after OP and the transition between events, five states are considered with one initial state (OP), three intermediate states (LR, Meta, and LR+Meta), and a terminal state (death). For notational simplicity, "metastasis" is denoted as Meta in the statistical analysis.

Under the assumption of the Markov property, a transition probability $\pi_{lk}(s,t)=\Pr(X(t)=k \mid X(s)=l\}$ defining the probability of staying state l at $T=s$ and moving to state k at $T=t$. To reflect the effect of covariates on transition, a transition intensity is defined as

$$\lambda_{lk}(t)=\lim_{\Delta t\to 0}\frac{\pi lk(t,t+\Delta t)}{\Delta t},\ \lambda_{kk}=-\sum_{k\neq l}\lambda_{lk}$$

which is a similar concept of Cox hazard function.

Fig. 1 shows the MSM with five states $K=\{0,1,2,3,4\}=\{$OP, LR, Meta, LR+Meta, Death$\}$ and eight transitions. The initial state is OP and there are three intermediate states (LR, Meta,

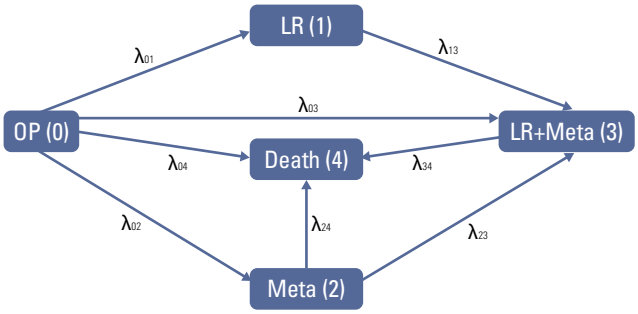


Fig. 1. Visualization of the multiple transitions consisting the multistate model with five states and eight transitions. Operation (OP) is the initial state and there are three intermediate states (local recurrence [LR], metastasis [Meta], concurrent local recurrence and metastasis [LR+Meta]) and a terminal state (Death). All patients start in state OP and may transit to one of the three intermediate states (transitions λ_{01}, λ_{02}, λ_{03}) or transit directly to a terminal state (transition λ_{04}) or stay on OP state without events.

LR+Meta) and a terminal event (Death) as an absorbing state. All patients start at OP (state 0) and may transfer to one of three intermediate states (state 1, 2, 3) or directly to a terminal state (state 4) or remain in OP state without events. Once patients move into one intermediate state, they are still at risk for other states or death. Therefore, they may also move forward to another intermediate state or death or stay. In detail, there are four possible transitions since OP. 0→1 denotes a transition OP→LR and the corresponding transition intensity is denoted as λ_{01} (t). Similarly, 0→2 denotes OP→ Meta with λ_{02} (t); 0→3 represents OP→ LR+Meta with λ_{03} (t) and 0→4 means the transition OP→Death with λ_{04} (t). Next, for the transition to as the intermediate states, 1→3 means the transition LR→LR+Meta with λ_{13} (t), 2→3 represents the transition Meta→LR+Meta with λ_{23} (t). Finally, some patients has experienced the terminal event, death. 2→4 represents the transition Meta→Death with λ_{24} (t) and 3→4 means the transition LR+Meta→Death with λ_{34} (t).

To estimate the effects of risk factors on each transition, the following proportional transition intensity model is assumed with a covariate vector $z=(z_1, \ldots z_p)$,

$$\lambda_{lk}(t \mid z)=\lambda_{lk0}(t)\exp(\beta_{lk}'z),$$

where $\lambda_{lk0}(t)$ is the baseline intensity function of the $l \rightarrow k$ transition and β_{lk} is the vector of the regression coefficient to measure the association between the $l \rightarrow k$ th transition and covariates z. All analyses were performed with R ver. 4.3.1 (R Foundation for Statistical Computing) and the function "probtrans" from R package 'mstate' is implemented to estimate the transition probability and the function "coxph" from R package 'survival' package is used to estimate the effect of covariates on each transition. Also, for comparing the time distributions of recurrent events at three OP types, F-test and Kruskal test are applied depending on the sample size. At all hypothesis tests, $p < 0.05$ was regarded as statistically significant.

Table 1. Demographics of the study population

Variable	No. of patients (%) (n=335)
Age (yr), mean (25th to 75th percentile)	61 (54-68)
Sex	
Male	186 (55.5)
Female	149 (44.5)
Smoking	
Never-smoker	180 (53.7)
Current or former-smoker	155 (46.3)
Adjuvant chemotherapy	
No	291 (86.9)
Yes	44 (13.1)
Local recurrence[a]	
No	281 (83.9)
Yes	54 (16.1)
Distant metastasis	
No	171 (51.0)
Yes	164 (49.0)
Death	
No	186 (55.5)
Yes	149 (44.5)
Stage	
I	220 (65.7)
II	65 (19.4)
III	50 (14.9)
pT	
T1	156 (46.6)
T2	159 (47.4)
T3	20 (6.0)
pN	
N0	242 (72.2)
N1	48 (14.3)
N2	45 (13.5)
Tumor size (mm), mean (25th to 75th percentile)	28 (20-34)
Operation type	
Sublobar resection	18 (5.4)
Lobectomy/Pneumonectomy	317 (94.6)

[a]Including those with concurrent local recurrence and metastasis.

Table 2. Frequency table of transitions

From	To				
	OP (state 0)	LR (state 1)	Meta (state 2)	LR+Meta (state 3)	Death (state 4)
OP (state 0) (n=335)	142 (42.4)	7 (2.1) (λ_{01})	142 (42.4) (λ_{02})	15 (4.5) (λ_{03})	29 (8.7) (λ_{04})
LR (state 1) (n=7)	0	0	0	7 (100) (λ_{13})	0
Meta (state 2) (n=142)	0	0	43 (30.3)	32 (22.5) (λ_{23})	67 (47.2) (λ_{24})
LR+Meta (state 3) (n=54)	0	0	0	1 (1.9)	53 (98.1) (λ_{34})

Values are presented as number of patients (%). Percentages reflect the proportion of patients in the state before transitioning. For example, among 335 patients from OP (state 0), 142 patients (42.4%) stayed at OP state without any further events. Seven patients (2.1%) developed LR (state 1). Another 142 patients developed Meta (42.4%), 15 patients developed LR+Meta (4.5%), and 29 patients (8.7%) died. Among 54 patients from LR+Meta (state 3), only one patient (1.9%) stayed at LR+Meta without any further event, and 53 patients (98.1%) died (state 4). Percent totals may not add up to 100% due to rounding error. LR, local recurrence; LR+Meta, concurrent local recurrence and metastasis; Meta, metastasis; OP, operation.

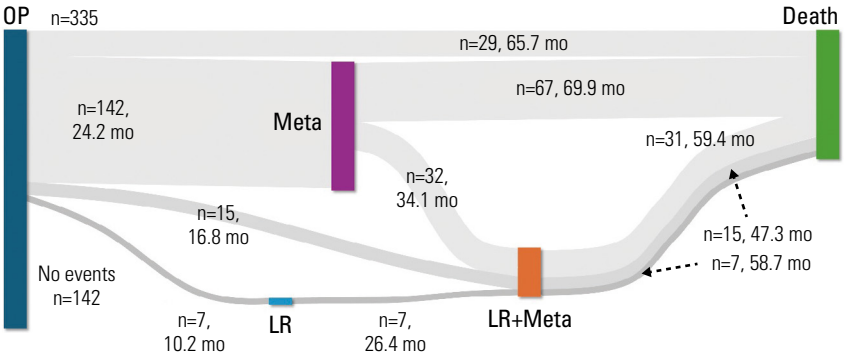


Fig. 2. Overflow of the various metastasizing pathways after operation and their corresponding patient numbers and median length of time to each status. The thickness of each transition line is proportional to the number of patients. LR, local recurrence; LR+Meta, concurrent local recurrence with metastasis; Meta, metastasis; mo, months; OP, operation.

Results

1. Patient characteristics and survival

Patient characteristics are summarized in Table 1. The final study group included 335 patients (186 men, 149 women) with a mean age of 61 years (25th to 75th percentile, 54 to 68 years). Regarding smoking history, 180 (53.7%) were never smokers and 155 (46.3%) were current or former smokers. Among the study group, 291 (86.9%) had never experienced adjuvant chemotherapy. Pathological TNM stages were classified as follows: stage I, 220 (65.7%); stage II, 65 (19.4%); and stage III, 50 (14.9%). The pT has three groups: T1, 156 (46.6%); T2, 159 (47.4%); and T3, 20 (6.0). The pN has also three groups: N0, 242 (72.2%); N1: 48 (14.3%); and N2, 45 (13.5%).

Mean tumor size was 28 mm (25th to 75th percentile, 20 to 34). Most patients underwent lobectomy or pneumonectomy (317 patients, 94.6%) and only 18 patients underwent sublobar resection.

Table 2 shows the numbers of patients corresponding to each transition path (Fig. 2). At the end of the study, among 335 patients starting from OP, 142 patients (42.4%) had no further events and survived until the end of the study, while 54 (16.1%) patients experienced LR+Meta and 142 patients (42.4%) experienced Meta, where the detection of LR and Meta may occur simultaneously or sequentially. Notably, all seven patients (2.1%) who experienced LR eventually developed concurrent Meta (transition λ_{13}).

Overall, 149 patients (44.5%) died by the end of the study. Among these, 67 patients (45.0%) died from Meta, 53 patients (35.6%) died from LR+Meta, and 29 patients (19.5%) died without intermediate events. Notably, with the exception of one patient, all other LR+Meta patients died.

For 142 patients who survived until the end of the study without any events, the mean follow-up period for these patients was 9.8±4.3 years after OP (S2 Table). The majority of these patients was diagnosed at early pathologic stages (stage I, 114 patients) and the mean tumor size was 2.7±1.3 cm. Considering predominant subtype, acinar type (79 patients, 55.6%) was the most frequent subtype in this group.

Table 3. Results of multivariate regression analysis of risk factors

Risk factor	Transitions							
	λ_{01} OP→LR	λ_{02} OP→Meta	λ_{03} OP→LR+Meta	λ_{04} OP→Death	λ_{13} LR→LR+Meta	λ_{23} Meta→LR+Meta	λ_{24} Meta→Death	λ_{34} LR+Meta→Death
Sex								
Female	0.645 (0.607)	–0.262 (0.280)	–0.768 (0.309)	0.303 (0.670)	0.279 (0.691)	–2.02 (< 0.001)[a]	0.245 (0.537)	–0.036 (0.938)
Age	0.004 (0.412)	0.010 (0.244)	0.034 (0.232)	0.122 (< 0.001)[a]	0.004 (0.412)	–0.015 (0.425)	0.062 (< 0.001)[a]	–0.007 (0.696)
Smoking								
Never-smoker	–1.34 (0.254)	0.217 (0.391)	0.381 (0.576)	–0.964 (0.170)	–1.34 (0.254)	0.699 (0.111)	–0.552 (0.151)	–0.637 (0.117)
Adjuvant chemotherapy								
No	–37.6 (0.998)	0.608 (0.176)	1.957 (0.107)	1.075 (0.389)	–37.6 (0.998)	1.000 (0.186)	0.979 (0.121)	1.808 (0.009)[a]
pT	0.208 (0.757)	0.360 (0.024)[a]	0.362 (0.455)	0.463 (0.187)	–0.272 (0.99)	0.674 (0.070)	–0.066 (0.811)	0.595 (0.073)
pN	–1.88 (0.998)	0.729 (0.001)[a]	1.957 (0.001)[a]	0.762 (0.216)	–26.9 (0.998)	1.189 (0.005)[a]	0.598 (0.059)	0.853 (0.015)[a]
Size	0.276 (0.416)	0.093 (0.180)	0.115 (0.538)	0.278 (0.061)	27.91 (0.991)	–0.158 (0.218)	1.113 (0.281)	–0.014 (0.912)
Operation type								
Sublobar resection	–2.75 (0.015)[a]	1.264 (0.079)	–1.094 (0.192)	–0.832 (0.163)	–16.4 (0.995)	0.338 (0.749)	–1.489 (0.171)	–0.445 (0.403)

Data are β coefficient values of risk factors and data in parentheses are p-values. Age, pT, pN, and size were regarded as continuous variables. LR, local recurrence; LR+Meta, concurrent local recurrence and metastasis; Meta, metastasis; OP, operation. [a]$p < 0.05$.

2. Risk factors for LR, Meta, and/or death

Table 3 demonstrates the results of Cox regression models to evaluate risk factors for each transition. Female patients were more likely to transition from Meta to LR+Meta (transition λ_{23}, β=–2.02, $p < 0.001$). Patients receiving adjuvant therapy were more likely to transit from LR+Meta to death (transition λ_{34}, β=1.808, p=0.009).

Age was only a significant risk factor for the transition from OP to death (transition λ_{04}, β=0.122, $p < 0.001$) and the transition from Meta to death (transition λ_{24}, β=0.062, $p < 0.001$). The pT status was significantly correlated with the transition from OP to Meta (transition λ_{02}, β=0.360, p=0.024). After OP, pN status was significantly correlated with the two transitions to Meta (transition λ_{02}, β=0.729, p=0.001) and to LR+Meta (transition λ_{03}, β=1.957, p=0.001).

In terms of OP type, compared to lobectomy / pneumonectomy, sublobar resection was strongly associated with the transition to LR after OP (transition λ_{01}, β=–2.72, p=0.015).

3. Transition probabilities

Fig. 3A shows the stacked transition probabilities to four states over time after OP. At early times, the transition probability to Meta was the largest but death became more likely to occur as time passed. In addition, there were two transitions from Meta as shown in Fig. 3B. According to the transition probabilities over time of two competing states (death, LR+Meta), the transitions of two states have very different patterns. At 2 years since OP, the transition probabilities from Meta to LR+Meta (transition λ_{23}) and death (transition λ_{24}) were 0.23 and 0.36, respectively. These probabilities at 4 years and 6 years since OP were (0.17, 0.55) and (0.06, 0.73), respectively. In other words, the transition to LR+Meta has a higher probability at 2-4 years compared with other time periods.

Median and 25th and 75th percentiles of four transition times (transitions λ_{01}, λ_{02}, λ_{03}, and λ_{04}) occurred since OP are displayed (S3 Table). While a median Meta occurrence time is 24.2 months (about 2 years) since OP, 29 death cases occurred over a median of 65.7 months (about 5.4 years) since OP. As shown in Table 2 and Fig. 3A, B, the transition to LR has low incidence and a very small band between OP and Meta.

4. Event incidence according to OP type

Table 4 demonstrates the event frequency and time-to-event according to OP type. The three OP types displayed different recurrence patterns ($p < 0.001$). For checking the distribution of the occurrence time of four events (LR, Meta, LR+Meta, and death) according to OP type, there was no significant different among events in the sublobar resection and pneumonectomy group. However, there was a significant difference in the lobectomy group. After OP, the median occurrence time of Meta was 2.05 years, the median time of

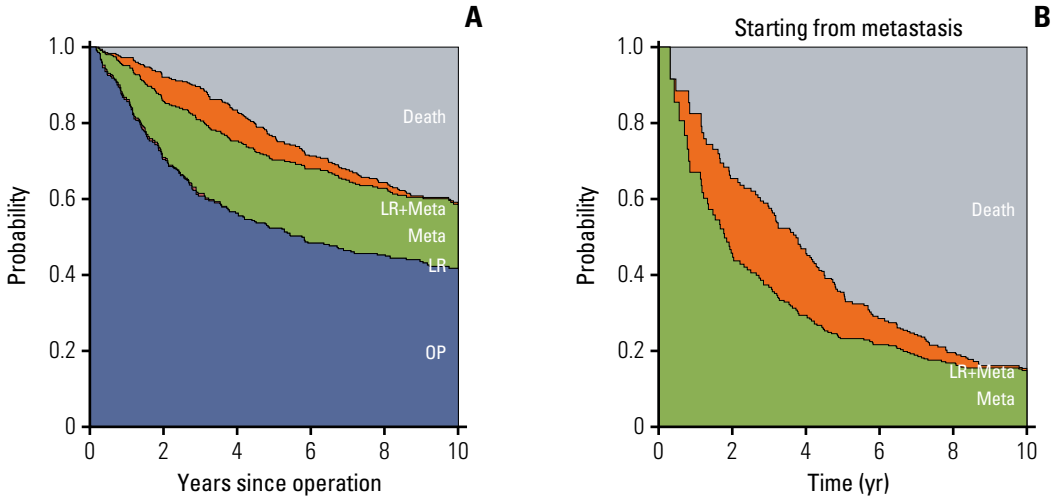


Fig. 3. (A) Transition probabilities starting from operation to death, concurrent local recurrence with metastasis (LR+Meta), metastasis (Meta), local recurrence (LR), or operation (OP). (B) Transition probabilities starting from metastasis to death, LR+Meta, or metastasis.

Table 4. Event frequency and time-to-event according to operation type

Operation type	Event type	Frequency		Time-to-event (yr)	
		Sample size (n)	Test-statistics[a]	Median (25%-75% percentile)	Test-statistics
Sublobar resection (n=18)	No event	7	Fisher's exact test $p < 0.001$	11.65 (10.59-13.44)	χ^2=2.209 (df=3) p=0.530
	LR(0→1)	2		1.53[b]	
	Meta(0→2)	2		1.48[b]	
	LR+Meta(0→3)	2		0.89[b]	
	Death(0→4)	5		4.79 (1.93-8.59)	
Lobectomy (n=317)	No event	133		9.89 (5.76-12.58)	F=16.81 (df1=3, df2=172) $p < 0.001$
	LR(0→1)	5		0.84 (0.67-1.37)	
	Meta(0→2)	136		2.05 (1.04-3.69)	
	LR+Meta(0→3)	13		1.38 (1.14-1.92)	
	Death(0→4)	22		6.28 (2.24-10.63)	
Pneumonectomy (n=317)	No event	2		10.33[b]	χ^2=0.214 (df=1) p=0.643
	LR(0→1)	0		-	
	Meta(0→2)	4		0.424[b]	
	LR+Meta(0→3)	0		-	
	Death(0→4)	2		0.563[b]	

LR, local recurrence; LR+Meta, concurrent local recurrence and metastasis; Meta, metastasis. [a]H_0: Three operation types have homogeneous frequencies of events; Fisher's exact test statistics, [b](25%-75%) quantile points are not available at n=2 and n=4.

LR+Meta was 1.38 years and the median time of death was 6.28 years, respectively. Fig. 4 shows the boxplot of the occurrence time of four recurrent events after lobectomy.

5. Analysis of comprehensive pathological data

An independent analysis of patients with comprehensive pathological data was performed. S4 Table shows the number of patients corresponding to each transition pathway. All five patients who transitioned from OP to LR subsequently moved on to LR+Meta. Thus, this transition (λ_{13}) was disregarded in the pathological analysis due to the extremely high regression coefficient value.

Table 5 shows the results of multivariate regression analysis of pathological risk factors for each transition. The presence of any vascular, lymphatic, or perineural invasion was a risk factor for transitions from OP to Meta (transition λ_{02}, β=0.635, $p < 0.001$) and OP to LR+Meta (transition λ_{03}, β=1.882, p=0.014).

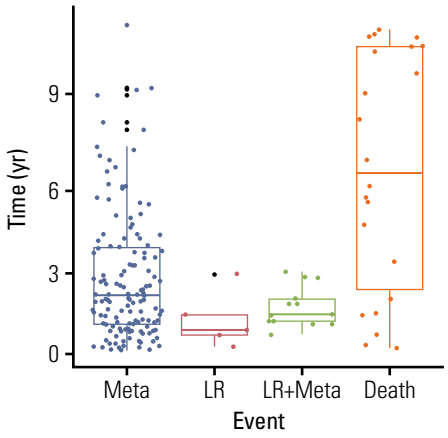


Fig. 4. Boxplot of time by event after lobectomy operation. After operation, the median occurrence time of metastasis (Meta) is 2.05 years, the median time of concurrent local recurrence (LR) and metastasis (LR+Meta) is 1.38 years and the median time of death is 6.28 years, respectively.

Discussion

Our findings can be summarized as follows: (1) After OP, approximately half of the patients in our study group developed recurrence with Meta as the most common route. To be specific, 142 (42.4%), 54 (16.1%), and seven (2.1%) patients developed Meta, LR+Meta, and LR, respectively.; (2) The prognosis and transition patterns from OP to intermediate states (LR, Meta, and LR+Meta) were different.; and (3) Clinical and pathological factors had different effect on each transition. (4) The presence of any vascular, lymphatic, or perineural invasion was a risk factor for transitions from OP to Meta and OP to LR+Meta. The key strength of our study is that we revealed multiple tumor recurrence pathways, each with its own prognosis. Moreover, we were able to separate the impact of multiple prognostic factors on the transitions following lung adenocarcinoma surgery. Initially, the transition probability to Meta is the largest, but the probability of death increases with time.

Our results revealed that nearly 50% of patients developed recurrence, with the most frequent pathway being Meta, followed by LR+Meta and LR. This confirms the results of earlier studies, which indicated that recurrence rates for stages I to III of NSCLC ranges from 30% to 75%, depending on pathologic stage and the length of follow-up [2,17-19]. In a large study by Boyd et al. [20], the incidences of recurrence after surgery were highest in distant Meta followed by local and distant Meta, and finally that confined to local regions. Distinguishing between LR, LR+Meta, and Meta is crucial because an increasing number of clinicians are adopting radical local treatment options that support systemic therapy, underscoring its importance [5].

Consistent with previous research, LR was found to occur quickly after surgery, with reported ranges of 50%-90% within 2 years of the initial OP [21]. In addition, a separate study demonstrated that 24% of patients had LR after lung cancer surgery [11]. The incidence of LR was lower (2.1%) in our study. This is probably because our study was conducted at a nationwide top tier, tertiary hospital that specializes in thoracic surgery and treatment. It should be noted that all patients with LR concurrently showed LR+Meta. Despite the small number of patients in our samples, our findings suggest that LR patients are at risk for developing Meta, so careful monitoring should be performed in this subgroup. Furthermore, all but one LR+Meta patients died, indicating poor prognosis.

In patients with Meta, the transition to LR+Meta has a higher probability at 2-4 years compared with other time periods. Hence, based on our findings, it is important to closely monitor metastatic patients for LR at 2-4 years. In a meta-analysis, Fedor et al recommended that patients with distant recurrence should be carefully examined for local lesions, as the presence of distant recurrence may imply that LR has had sufficient time to develop [11].

Age was a risk factor for OP to death (transition λ_{04}) and Meta to death (transition λ_{24}) in our study. Aging may present as a challenge after OP as the respiratory system undergoes physiologic changes and multiple chronic conditions tend to coexist in older patients. Furthermore, the immune system functionally declines in elderly patients, increasing the risk of respiratory tract infections [22]. Hence, elderly patients are a heterogeneous group that can be at risk for both over and under-treatment in lung cancer patients. However, age itself was not a risk factor for OP to LR (transition λ_{01}) or OP to Meta (transition λ_{02}) in our study. This may reflect that lung tumors in elderly patients tend to be biologically less aggressive. In a large study, Sacher et al. [23] found that younger NSCLC patients were more likely to have a targetable genotype, but the survival rate for young patients was unexpectedly poor compared to other age groups, indicating more aggressive disease biology.

Our results further emphasize the significant role of pT and pN, two important staging descriptors. Instead of the initial TNM staging, we assessed the effects of pT and pN separately in this study. The pT was a risk factor for OP to Meta (transition λ_{02}). The impact of pN was still evident even after surgery, as it remained a risk factor for both OP to Meta (transition λ_{02}) and OP to LR+Meta (transition λ_{03}). In addition, pN was also a risk factor for Meta to LR+Meta (transition λ_{23}). The presence of cancer cells in the LNs is a significant prognostic factor for most cancers, but the exact mechanism remains uncertain. Originally, the commonly

Table 5. Results of multivariate regression analysis of pathological risk factors (n=307)

Risk factor	Reference	Transitions							
		λ_{01} OP→LR	λ_{02} OP→Meta	λ_{03} OP→LR+Meta	λ_{04} OP→Death	λ_{23} Meta→LR+Meta	λ_{24} Meta→Death	λ_{34} LR+Meta→Death
Tumor differentiation	Well and moderate	–14.95 (0.989)	0.163 (0.469)	–0.668 (0.357)	0.857 (0.087)	–0.170 (0.728)	0.636 (0.069)	–0.313 (0.425)
Predominant histologic pattern	Lepidic, acinar, and papillary	1.238 (0.195)	0.199 (0.369)	0.492 (0.426)	–0.619 (0.259)	0.536 (0.256)	–0.369 (0.294)	0.647 (0.063)
Pleural invasion	No	–0.556 (0.559)	0.363 (0.056)	0.627 (0.288)	–0.262 (0.511)	0.571 (0.219)	–0.417 (0.151)	0.463 (0.215)
Vascular, lymphatic, or perineural invasion	No	–0.412 (0.654)	0.635 (< 0.001)[a]	1.882 (0.014)[a]	0.396 (0.326)	–0.212 (0.600)	0.076 (0.796)	0.008 (0.812)

Data are β coefficient values of risk factors and data in parentheses are p-values. LR, local recurrence; LR+Meta, concurrent local recurrence and metastasis; Meta, distal metastasis; OP, operation. [a]$p < 0.05$.

accepted concept regarding metastatic progression was that it follows a sequential pattern, with the presence of tumor cells in regional LNs leading to an increased chance of tumor cells throughout the lymphatic system, ultimately leading to distant metastases [24]. In contrast, Naxerova et al. [25] found that a considerable portion of colorectal cancer patients exhibited distinct genetic origins for their LNs and distant lesions. These findings propose that there are two distinct lineage relationships between lymphatic and distant metastases [25].

Prior studies have established lobectomy as the gold standard for lung cancer surgery, as it results in a lower risk of LR compared to wedge resection or segmentectomy [18]. In the decades since those initial studies, advancements in imaging techniques have made it possible to detect tumors at earlier stages and smaller sizes, leading to controversy and debates regarding sublobar resection in patients with early lung cancer. A recent large, randomized trial concluded that sublobar resection is recommended solely for NSCLC patients with a tumor size of 2 cm or smaller, as well as confirmed absence of LN involvement [26]. The evidence from our study reinforces this fact, as the risk of LR after sublobar resection was found to be higher, despite the small number of patients undergoing sublobar resection.

It is worth noting that 142 patients (42.4%) had no tumor recurrence and survived until the end of the study period. People who live for more than 5 years following a lung cancer diagnosis have been referred to as long-term lung cancer patients. Our results exceed the average 5-year survival rates reported worldwide, ranging from 10%-20% in most regions, and the highest rate of 32.9% observed in Japan [27]. This might be attributed to early detection through lung cancer screening and advances in treatment options in Korea. In our study, 80% of patients who survived without tumor recurrence were found to have pathologic stage I disease. Studies have confirmed that patients diagnosed with lung cancer through annual screening programs and those diagnosed in the early stages have a higher lung cancer-specific survival rate [28,29].

Finally, we performed an independent analysis of patients with comprehensive pathological data. Based on our results, the detection of any vascular, lymphatic, or perineural invasion is likely to be more closely related to the development of metastases. Earlier research by Higgins et al. [30] has also indicated that the presence of lymphovascular invasion is a significant adverse prognostic factor for both the development of distant metastases and long-term survival outcomes. Despite being recognized as significant risk factors for unfavorable outcomes in lung adenocarcinoma patients, factors such as tumor differentiation, predominant histologic pattern, and the presence of pleural invasion were not

identified as risk factors among the transitions in this study. Pleural invasion is a commonly acknowledged risk factor, and according to the latest staging system, tumors that are 3cm or smaller with visceral pleural invasion are upstaged to T2 status, influencing the type of surgery conducted [31]. Tumors identified as high risk based on their differentiation or predominant histologic pattern serves as an indicator for more intense adjuvant treatment strategies. Consequently, we believe that these factors of pleural invasion, predominant histologic type, and tumor differentiation were already incorporated into the surgical procedure and subsequent treatment, compensating their significance.

Two traditional statistical models, the log-rank test and Cox proportional hazard model, are often used for the analysis of survival or time-to-event outcomes. However, these models only allow for evaluation of the impact of covariates on a single outcome, such as death. In this regard, MSM enables clinicians to partition prognostic factors according to various outcomes as well as to elucidate the evolutionary pattern over time. To date, only a few studies have employed this approach in oncology studies. In a study of colorectal cancer patients, MSM separated the effects of prognostic factors on recurrence and mortality [32]. Ferrer et al. [33] applied MSM to prostate cancer patients and were able to formulate a complete model of prostate cancer progression that considers both classical prognostic factors and prostate-specific antigen dynamics between clinical health states. Jeong et al. [34] also applied a multi-state model to describe the transition probability and to estimate the covariate effect on each transition. The significant contribution of our study was to consider the concurrence of Meta and local recurrence (LR+Meta). That is, they considered only two events (LR and Meta) and five transition pathways after OP [34]. Our study included three events and eight transitions between the states. Hence, our study provides a wider range of clinical scenarios as we tried to reflect the complexity of the real-world data. However, compared with the previous study of 949 patients, our dataset has smaller sample size, which may result in fewer occurrences in OP to LR and LR to LR+Meta.

There were several limitations in this study. First, although patient data from a specialized cancer hospital were analyzed, this was a retrospective study based on the data from a single institution, and the findings may not be universal and needs to be verified in a larger multicenter study. Second, imaging studies such as CT, MRI, ^{18}F-FDG PET/CT, and bone scan were performed at varying intervals during follow-up or upon clinical suspicion. In line with typical observational studies when compared to clinical trials, our results may have been influenced by lead time bias caused by the time gap between the date of diagnosis and true progression of the disease. However, our institute is highly specialized in cancer treatment and maintains a very high level of patient compliance. Hence, the majority of patients was treated according to standard protocols and underwent regular follow-up examinations. Third, new techniques such as detecting circulating tumor DNA in the blood have shown potential and accuracy in monitoring and detecting recurrence in cancer patients. However, due to the retrospective design of this study, we were unable to correlate our findings with cancer profile blood levels. Fourth, a considerable proportion of patients simply underwent surgery without additional chemotherapy. The explanation for this was that some patients were too elderly and in poor general condition and some patients decided against undergoing chemotherapy. Some patients refrained from undergoing chemotherapy due to financial reasons. It is important to highlight that the results we obtained is from the real-world data. In addition, treatments such as 3rd generation tyrosine kinase inhibitors and immunotherapy agents has been widely adopted for lung cancer treatment because of their high selectivity and minimal adverse effects in recent years. However, our study cohort had an extensive follow-up period of more than 10 years, meaning it was established before the widespread availability of current treatments. Thus, further investigations including a more recent patient cohort which includes the recent treatment options would be interesting. Finally, we were unable to examine the presence of spread through air spaces, as our patient cohort included patients who underwent OPs more than ten years prior, when spread through air spaces was not an essential part of assessment process.

In conclusion, lung adenocarcinoma displays a broad spectrum of clinical scenarios even after curative surgery. The incidence, risk factors, and prognosis varied across different pathways of recurrence in patients with lung adenocarcinoma. The greatest implication of this MSM is its ability to predict the timing and type of clinical intervention that will have the greatest impact on survival in lung adenocarcinoma patients.

Electronic Supplementary Material

Supplementary materials are available at Cancer Research and Treatment website (https://www.e-crt.org).

Ethical Statement

This study was conducted in accordance with the 1964 Declaration of Helsinki and was approved by the Institutional Review Board of Samsung Medical Center, Seoul, Korea (IRB approval number: 2017-09-045). Written informed consent was waived due to the retrospective nature of this study.

Author Contributions

Conceived and designed the analysis: Lee G, Kim YJ, Lee HY.
Collected the data: Lee G, Lee HY.
Contributed data or analysis tools: Lee G, Kim YJ, Sohn I, Kim JH, Lee HY.
Performed the analysis: Kim YJ.
Wrote the paper: Lee G, Kim YJ, Sohn I, Kim JH, Lee HY.

ORCID iDs

Geewon Lee : https://orcid.org/0000-0002-8278-0500
Yang-Jin Kim : https://orcid.org/0000-0002-0882-7622
Ho Yun Lee : https://orcid.org/0000-0001-9960-5648

Conflicts of Interest

Conflict of interest relevant to this article was not reported.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2C1003999) and was supported by Future Medicine 20*30 Project of the Samsung Medical Center (#SMO1240791) and partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub).

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023;73:17-48.
2. Martin J, Ginsberg RJ, Venkatraman ES, Bains MS, Downey RJ, Korst RJ, et al. Long-term results of combined-modality therapy in resectable non-small-cell lung cancer. J Clin Oncol. 2002;20:1989-95.
3. Takenaka T, Yano T, Yamazaki K, Okamoto T, Hamatake M, Shimokawa M, et al. Survival after recurrence following surgical resected non-small cell lung cancer: a multicenter, prospective cohort study. JTCVS Open. 2022;10:370-81.
4. Uramoto H, Tanaka F. Prediction of recurrence after complete resection in patients with NSCLC. Anticancer Res. 2012;32:3953-60.
5. Guckenberger M, Lievens Y, Bouma AB, Collette L, Dekker A, deSouza NM, et al. Characterisation and classification of oligometastatic disease: a European Society for Radiotherapy and Oncology and European Organisation for Research and Treatment of Cancer consensus recommendation. Lancet Oncol. 2020;21:e18-28.
6. Hung JJ, Yeh YC, Jeng WJ, Wu KJ, Huang BS, Wu YC, et al. Predictive value of the international association for the study of lung cancer/American Thoracic Society/European Respiratory Society classification of lung adenocarcinoma in tumor recurrence and patient survival. J Clin Oncol. 2014;32:2357-64.
7. Meira-Machado L, de Una-Alvarez J, Cadarso-Suarez C, Andersen PK. Multi-state models for the analysis of time-to-event data. Stat Methods Med Res. 2009;18:195-222.
8. Beesley LJ, Morgan TM, Spratt DE, Singhal U, Feng FY, Furgal AC, et al. Individual and population comparisons of surgery and radiotherapy outcomes in prostate cancer using Bayesian multistate models. JAMA Netw Open. 2019;2:e187765.
9. Beesley LJ, Shuman AG, Mierzwa ML, Bellile EL, Rosen BS, Casper KA, et al. Development and assessment of a model for predicting individualized outcomes in patients with oropharyngeal cancer. JAMA Netw Open. 2021;4:e2120055.
10. Le-Rademacher JG, Peterson RA, Therneau TM, Sanford BL, Stone RM, Mandrekar SJ. Application of multi-state models in cancer clinical trials. Clin Trials. 2018;15:489-98.
11. Fedor D, Johnson WR, Singhal S. Local recurrence following lung cancer surgery: incidence, risk factors, and outcomes. Surg Oncol. 2013;22:156-61.
12. Taylor MD, Nagji AS, Bhamidipati CM, Theodosakis N, Kozower BD, Lau CL, et al. Tumor recurrence after complete resection for non-small cell lung cancer. Ann Thorac Surg. 2012;93:1813-20.
13. Detterbeck FC, Franklin WA, Nicholson AG, Girard N, Arenberg DA, Travis WD, et al. The IASLC Lung cancer staging project: background data and proposed criteria to distinguish separate primary lung cancers from metastatic foci in patients with two lung tumors in the forthcoming eighth edition of the TNM classification for lung cancer. J Thorac Oncol. 2016;11:651-65.
14. Detterbeck FC, Marom EM, Arenberg DA, Franklin WA, Nicholson AG, Travis WD, et al. The IASLC lung cancer staging project: background data and proposals for the application of TNM staging rules to lung cancer presenting as multiple nodules with ground glass or lepidic features or a pneumonic type of involvement in the forthcoming eighth edition of the TNM classification. J Thorac Oncol. 2016;11:666-80.
15. Voulaz E, Novellis P, Rossetti F, Solinas M, Rossi S, Alloisio M, et al. Distinguishing multiple lung primaries from intrapulmonary metastases and treatment implications. Expert Rev Anticancer Ther. 2020;20:985-95.
16. Werner R, Steinmann N, Decaluwe H, Date H, De Ruysscher D, Opitz I. Complex situations in lung cancer: multifocal disease, oligoprogression and oligorecurrence. Eur Respir Rev. 2024;33:230200.
17. al-Kattan K, Sepsas E, Fountain SW, Townsend ER. Disease recurrence after resection for stage I lung cancer. Eur J Cardiothorac Surg. 1997;12:380-4.
18. Martini N, Bains MS, Burt ME, Zakowski MF, McCormack P, Rusch VW, et al. Incidence of local recurrence and second primary tumors in resected stage I lung cancer. J Thorac Cardiovasc Surg. 1995;109:120-9.
19. Yoshino I, Yohena T, Kitajima M, Ushijima C, Nishioka K, Ichinose Y, et al. Survival of non-small cell lung cancer patients

with postoperative recurrence at distant organs. Ann Thorac Cardiovasc Surg. 2001;7:204-9.
20. Boyd JA, Hubbs JL, Kim DW, Hollis D, Marks LB, Kelsey CR. Timing of local and distant failure in resected lung cancer: implications for reported rates of local failure. J Thorac Oncol. 2010;5:211-4.
21. Alberts WM. Follow up and surveillance of the patient with lung cancer: what do you do after surgery? Respirology. 2007; 12:16-21.
22. Meyer KC. Lung infections and aging. Ageing Res Rev. 2004; 3:55-67.
23. Sacher AG, Dahlberg SE, Heng J, Mach S, Janne PA, Oxnard GR. Association between younger age and targetable genomic alterations and prognosis in non-small-cell lung cancer. JAMA Oncol. 2016;2:313-20.
24. Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y. Molecular principles of metastasis: a hallmark of cancer revisited. Signal Transduct Target Ther. 2020;5:28.
25. Naxerova K, Reiter JG, Brachtel E, Lennerz JK, van de Wetering M, Rowan A, et al. Origins of lymphatic and distant metastases in human colorectal cancer. Science. 2017;357:55-60.
26. Altorki N, Wang X, Kozono D, Watt C, Landrenau R, Wigle D, et al. Lobar or sublobar resection for peripheral stage IA non-small-cell lung cancer. N Engl J Med. 2023;388:489-98.
27. Bi JH, Tuo JY, Xiao YX, Tang DD, Zhou XH, Jiang YF, et al. Observed and relative survival trends of lung cancer: a systematic review of population-based cancer registration data. Thorac Cancer. 2024;15:142-51.
28. He S, Li H, Cao M, Sun D, Yang F, Yan X, et al. Survival of 7,311 lung cancer patients by pathological stage and histological classification: a multicenter hospital-based study in China. Transl Lung Cancer Res. 2022;11:1591-605.
29. Henschke CI, Yip R, Shaham D, Markowitz S, Cervera Deval J, Zulueta JJ, et al. A 20-year follow-up of the International Early Lung Cancer Action Program (I-ELCAP). Radiology. 2023; 309:e231988.
30. Higgins KA, Chino JP, Ready N, D'Amico TA, Berry MF, Sporn T, et al. Lymphovascular invasion in non-small-cell lung cancer: implications for staging and adjuvant therapy. J Thorac Oncol. 2012;7:1141-7.
31. Rami-Porta R, Nishimura KK, Giroux DJ, Detterbeck F, Cardillo G, Edwards JG, et al. The International Association for the Study of Lung Cancer lung cancer staging project: proposals for revision of the TNM stage groups in the forthcoming (ninth) edition of the TNM classification for lung cancer. J Thorac Oncol. 2024;19:1007-27.
32. Gilard-Pioc S, Abrahamowicz M, Mahboubi A, Bouvier AM, Dejardin O, Huszti E, et al. Multi-state relative survival modelling of colorectal cancer progression and mortality. Cancer Epidemiol. 2015;39:447-55.
33. Ferrer L, Rondeau V, Dignam J, Pickles T, Jacqmin-Gadda H, Proust-Lima C. Joint modelling of longitudinal and multistate processes: application to clinical progressions in prostate cancer. Stat Med. 2016;35:3933-48.
34. Jeong WG, Choi H, Chae KJ, Kim J. Prognosis and recurrence patterns in patients with early stage lung cancer: a multi-state model approach. Transl Lung Cancer Res. 2022;11:1279-91.

박호영 교수

논문

- Kim, J., Park, J., and Park, H. (2025). High dimensional discriminant rules with shrinkage estimators of the covariance matrix and mean vector. Journal of Statistical Planning and Inference, 234, 106199.

Journal of Statistical Planning and Inference 234 (2025) 106199

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

High dimensional discriminant rules with shrinkage estimators of the covariance matrix and mean vector

Jaehoan Kim [a], Junyong Park [b], Hoyoung Park [c,*]

[a] Department of Statistical Science, Duke University, NC, USA
[b] Department of Statistics, Seoul National University, Seoul, Republic of Korea
[c] Department of Statistics, Sookmyung Women's University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
High dimensional discriminant analysis
Nonparametric maximum likelihood estimation
Nonparametric empirical bayes
Estimation of precision matrix

ABSTRACT

Linear discriminant analysis (LDA) is a typical method for classification problems with large dimensions and small samples. There are various types of LDA methods that are based on the different types of estimators for the covariance matrices and mean vectors. In this paper, we consider shrinkage methods based on a non-parametric approach. For the precision matrix, methods based on the sparsity structure or data splitting are examined. Regarding the estimation of mean vectors, Non-parametric Empirical Bayes (NPEB) methods and Non-parametric Maximum Likelihood Estimation (NPMLE) methods, also known as f-modeling and g-modeling, respectively, are adopted. The performance of linear discriminant rules based on combined estimation strategies of the covariance matrix and mean vectors are analyzed in this study. Particularly, the study presents a theoretical result on the performance of the NPEB method and compares it with previous studies. Simulation studies with various covariance matrices and mean vector structures are conducted to evaluate the methods discussed in this paper. Furthermore, real data examples such as gene expressions and EEG data are also presented.

1. Introduction

Discriminant analysis is one of the most widely emerging problems in various fields, including marketing, biomedical studies, sociology, and psychology. Due to the development of data collection technology, the data which contains the larger number of features than the number of samples became prevalent. Therefore, discriminant analysis in high-dimensional situations became important. Owing to its simplicity and optimality under the knowledge of parameters (see Anderson (2003)), the Fisher's linear discriminant analysis (LDA) received a lot of attention and has been studied intensively in both theory and practical applications. Since this linear discriminant rule requires mean vectors and the precision matrix, they should be estimated in practical situations. However, especially for high dimensional situations, precision matrix should be estimated carefully. Since the sample covariance matrix is not invertible anymore, the inverse of the sample covariance matrix is no more valid as an estimator. Therefore, there have been numerous studies to resolve this problem and construct the linear discriminant rules.

First, there were several approaches to estimating the precision matrix as a diagonal one, by simply assuming the independence among features. Since this assumption makes the true covariance matrix to be diagonal, it only requires each feature's precision to be estimated. The discriminant rule based on this assumption is called the independence rule (IR). In addition to that, feature selection had been added on this assumption with intuition to "reduce the noise from data". This led to the feature annealed independence

* Corresponding author.
E-mail addresses: jaehoan.kim@duke.edu (J. Kim), junyongpark@snu.ac.kr (J. Park), hyparks@sookmyung.ac.kr (H. Park).

https://doi.org/10.1016/j.jspi.2024.106199
Received 1 February 2023; Received in revised form 10 May 2024; Accepted 2 June 2024
Available online 8 June 2024
0378-3758/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

rule (FAIR), studied in Fan and Fan (2008), showing that the FAIR method can improve the error rate of the IR method in real data situations.

However, the IR and the FAIR are restrictive since they ignore all the possible dependence among features, which might be crucial if there exist some features which are crucially related, for example, voxels in fMRI data. Therefore, there have been several approaches to preserve the 'structure', which also can be considered as connectivity, of original data.

One possible approach is to slightly modify the sample covariance matrix into an invertible form using the idea of a shrinkage estimator. For the sample covariance matrix S and $0 < \lambda < 1$, one can take $S' = (1-\lambda)S + \lambda I$ or $S' = (1-\lambda)\nu S + \lambda I$ as the covariance estimator and use its inverse as the estimator of the precision matrix. Ledoit and Wolf (2004) broadened Stein's shrinkage estimation technique from the mean vector to the covariance matrix, and suggested how to choose the coefficients to minimize the L_2 risk. Additionally, Ledoit and Wolf (2012, 2020, 2022) have proposed nonlinear shrinkage estimation of the covariance matrix based on the spectral decomposition of the sample covariance matrix. Additionally, other approaches of precision estimation are proposed to solve the optimization problem with penalty terms. See Cai et al. (2011) and Friedman et al. (2008) for more detailed explanations. Lam (2016) suggested nonparametric sample splitting technique for the estimation.

In addition to the estimation of covariance or precision matrix, mean vector estimation has several approaches as well. Starting from the groundbreaking work by James (1961), shrinkage estimation methods have been actively researched, with notable contributions from researchers such as Chételat and Wells (2012), Bodnar et al. (2019), who focused on high-dimensional shrinkage mean estimation methods. Additionally, Wang et al. (2014) extended and proposed nonparametric shrinkage methods under unknown covariance matrices. For the general mean estimation approach, Efron (2014) categorized estimation strategies in empirical Bayes into the f-modeling and g-modeling which had been actually used in many estimation problems before. There are several attempts to utilize these strategies into discriminant analysis problem. See Greenshtein and Park (2009) and Efron (2009) for f-modeling and Dicker and Zhao (2016) and Park et al. (2022) for g-modeling in high dimensional discriminant analysis. The theoretical properties of the discriminant rules based on these estimators are gradually unveiled especially for g-modeling rather than f-modeling; see Dicker and Zhao (2016) and Saha and Guntuboyina (2020) for theoretical studies on the discriminant rule based on g-modeling. Although Greenshtein and Park (2009) and Efron (2009) used the f-modeling, to the best of our knowledge, there has been no study on theoretical properties of high dimensional discriminant rule based on f-modeling for estimating mean vectors. It is worth noting that both f-modeling and g-modeling in mean vector estimations assume independence among features, which requires appropriate data decorrelation before applying these strategies.

There is some research on utilizing these estimators of covariance or precision matrix in high dimensional discriminant analysis. For example, Hyodo et al. (2012) provided numerical experiments when $S' = \alpha S + \beta I$ is used in LDA. Sifaou et al. (2020) investigated a different estimator under spiked covariance model. For scale invariance, Li et al. (2025) utilized $\mathrm{diag}(S)$ instead of I. Xu et al. (2009) compared the performance of various discriminant rules for the microarray data. Park et al. (2022) presented intensive comparisons among discriminant rules based on various estimators of covariance or precision matrix. It is shown that the non-parametric mean vector estimation method and the precision estimator proposed in Lam (2016) provided the best overall performance. Although Park et al. (2022) presented various results, some important methodologies are missing. For example, mean vector estimation based on the NPEB in Greenshtein and Park (2009) has not been considered in their study and the sparse precision matrix has not been combined with such mean vector estimations, either.

In this context, our paper has two main contributions.

- First, we provide theoretical results on the performance of the discriminant rule based on f-modeling mean vector estimator. By providing such a theoretical performance on f-modeling, it is possible to compare the performances of f-modeling and g-modeling in high dimensional discriminant analysis where the theoretical result on g-modeling is provided in Park et al. (2022).
- Second, we elaborate on the impact of the aforementioned estimation methods for precision matrices and mean vectors on the constructed discriminant rules. Since mean vector estimation methods require the precision matrix to remove the dependency in the data before application, the effect of the two estimators is intertwined. Hence, we cross check the estimation method, we analyze the impact of the estimators on real data examples.

This paper is organized as follows. In Section 2, we review the Fisher's rule and estimation of the covariance matrix and mean vector. In Section 3, we provide the asymptotic results for nonparametric empirical Bayes method and comparison with other methods. Section 4 includes simulations studies to evaluate all the discriminant rules considered in this paper for various combinations of mean vectors and covariance matrices. In Section 5, real data examples are also presented to compare all the discriminant rules. We summarize all the results in Section 6 as concluding remarks.

2. Construction of discriminant rules

In this section, we review the methods of estimating precision matrix and mean vector in high dimensions. Throughout this paper, we assume $\mathbf{X}_{ij} \sim N_p(\boldsymbol{\mu}_i, \Sigma)$ for $i = 1, 2$ and $j = 1, \ldots, n_i$. To avoid confusion, we use bold symbols to indicate vectors in this section.

A well-known linear discriminant rule, Fisher's linear discriminant analysis (LDA), can be written as

$$f(\boldsymbol{x}^{new}) = \frac{1}{2}\left[3 - \mathrm{sign}\left\{\delta(\boldsymbol{x}^{new})\right\}\right]$$

for the new data $\boldsymbol{x}^{new} \in \mathbb{R}^p$, where

$$\delta(\boldsymbol{x}^{new}) = \left(\boldsymbol{x}^{new} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right). \tag{1}$$

$\delta(\boldsymbol{x}^{new})$ has the value 1 or 2 to indicate the group which the new observation is categorized. As one can see, Fisher's LDA requests two estimates, for $\boldsymbol{\mu}_i$ and $\Omega = \Sigma^{-1}$.

First, when estimating Ω, difficulty arises in high-dimensional situations since one cannot directly plug in the inverse of the sample covariance matrix due to its rank deficiency. Therefore, several alternatives exist to estimate Ω directly in high-dimensional situations. In this paper, we utilized the precision estimation methods using (1) L_1 penalty of the precision matrix (which is called the graphical Lasso method; we call this method 'glasso' in this paper; see Friedman et al. (2008)), and, (2) random sample splitting (we call this method 'LAM'; see Lam (2016)). Of course, we can assume the precision matrix to be diagonal and naively estimate each element. We write this method as 'IR,' which is the shorthand for 'Independence Rule'.

Regarding the estimation of $\boldsymbol{\mu}_i$, we use two methods based on two types of nonparametric empirical Bayes estimations, namely g-modeling and f-modeling introduced in Efron (2014). Such mean vector estimations have been used in Greenshtein and Park (2009) and Dicker and Zhao (2016) under the independent assumption of variables. Obviously, we also consider the sample mean as an estimate of $\boldsymbol{\mu}_i$.

The mainstream of this paper is to examine the performance of different types of discriminant rules in high-dimensional situations and analyze their performance both practically and theoretically.

As a building block, we define the decorrelated observation $\mathbf{Z}_{ij}$ as

$$\mathbf{Z}_{ij} = \Omega^{1/2}\mathbf{X}_{ij} \sim N(\Omega^{1/2}\boldsymbol{\mu}_i, I_p) \equiv N_p(\boldsymbol{\mu}_i^*, I_p) \tag{2}$$

where I_p is $p \times p$ identity matrix. Based on $\mathbf{Z}_{ij}$, the Fisher's rule is expressed as follows:

$$\begin{aligned}
\delta(\boldsymbol{x}^{new}) &= \left(\boldsymbol{x}^{new} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) \\
&= \left\{\Sigma^{-1/2}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right\}^\top \Sigma^{-1/2}\left(\boldsymbol{x}^{new} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^\top \\
&= \left(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\right)^\top \left(\boldsymbol{z}^{new} - \frac{\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*}{2}\right) \\
&\equiv \sum_{i=1}^{p} a_i z_i^{new} + a_0
\end{aligned} \tag{3}$$

where $\boldsymbol{z}^{new} = \Omega^{1/2}\boldsymbol{x}^{new} = \Sigma^{-1/2}\boldsymbol{x}^{new}$, $a_i = \mu_{1i}^* - \mu_{2i}^*$ and $a_0 = -\left(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\right)^\top \left(\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*\right)/2$. z_i^{new} denotes the ith component of $\boldsymbol{z}^{new}$.

In practice, of course, $\Omega = \Sigma^{-1}$ is unknown, so we need to estimate $\hat{\Omega}$ and decorrelate $\mathbf{X}_{ij}$ with $\hat{\Omega}^{1/2}$. If such estimates are accurate, one can expect that all variables in $\mathbf{Z}_{ij}$ are nearly independent, hence we can apply the mean estimation methods described later, which assumes the independence of variables.

To summarize, we need two estimates to construct Fisher's discriminant rule: one for Ω (or Σ) and the other one for $\boldsymbol{\mu}_i$ (or $\boldsymbol{\mu}_i^*$). We present these two procedures in the following sections.

2.1. Estimation of precision matrix

In the case of the precision matrix estimation, we present three approaches: one is based on some special structural assumption, and the others are for the general cases without this assumption, using random sample splitting and shrinkage to diagonal matrix, respectively. For the former case, it is widely assumed that the precision matrix has sparsity (relatively small number of nonzero components within the matrix) or graphical structure, as Cai et al. (2011) and Friedman et al. (2008). However, this structural assumption has limitations to be applied for general multivariate data. On the other hand, the latter cases can be generally utilized since it is free of any structural assumption. See, for example, Lam (2016). However, for the random splitting method, the data dimension gives a computational restriction compared others. In this paper, we consider the method based on the graphical model, called glasso in Friedman et al. (2008) as an example of the first approach. Subsequently, we introduce the method based on random sample data splitting, called the LAM method in Lam (2016). We then present additional two shrinkage methods of covariance matrix introduced in Ledoit and Wolf (2004) and Li et al. (2025), where we obtain the shrinkage estimators for the covariance matrix and invert them. We name these methods as 'Ridge1' and 'Ridge2', respectively.

First, as previously mentioned, the glasso method is initially devised to estimate the precision matrix under the sparsity assumption. To utilize this assumption, Friedman et al. (2008) converted this estimation problem into the optimization problem, maximizing a penalized log-likelihood function with respect to the matrix Θ. The penalty term is given in the L_1-norm of the precision matrix, which can favor the sparsity assumption. The objective function in glasso is written as

$$f(\Theta) = \log\det(\Theta) - tr(S_n\Theta) - \rho\|\Theta\|_1 \tag{4}$$

where S_n is sample covariance matrix, $\det(\Theta)$ is the determinant of Θ and ρ is the regularization parameter. The optimal value $\Theta^* = \text{argmax}_\Theta\, f(\Theta)$ is used as the estimator of Ω.

Subsequently, the LAM algorithm uses a random sample splitting idea. The fundamental issue for high dimensional data is that S_n is not invertible, which hinders S_n^{-1} to be used as an estimator for the precision matrix. To detour this problem, Lam (2016) applied the following strategy to construct an invertible estimator for the covariance matrix.

First, the true covariance matrix Σ can be represented as

$$\Sigma = PDP^T, D \succeq 0, \tag{5}$$

where P is the orthogonal matrix and D diagonal matrix. This can be also written as

$$\Sigma = PDP^T = Pdiag(P^T \Sigma P)P^T, \tag{6}$$

where $diag(A)$ denotes the diagonal matrix with diagonal elements of A. Based on this, to construct an invertible estimator $\hat{\Sigma}$ for Σ, Lam (2016) first split n observations into two groups of n_1 and $n_2 = n - n_1$ observations. Then, they used one group to estimate P and the other to estimate Σ for $diag(P^T \Sigma P)$. Since we already have two groups of observations, we can use the first group of observations to estimate P, and the second group of n_2 observations to estimate Σ on the right-hand side of (6). Therefore, when we denote the sample covariance matrix of two groups as $S_{n,1}$, $S_{n,2}$, and $\hat{P}_1$ the eigenvectors of $S_{n,1}$, the estimator is represented as

$$\hat{\Sigma}_{LAM} = \hat{P}_1 \mathrm{diag}(\hat{P}_1^T S_{n,2} \hat{P}_1) \hat{P}_1^T. \tag{7}$$

Lastly, we present two shrinkage methods that apply shrinkage on sample covariance matrix S to the diagonal matrices, to make the resulting estimator invertible. First one, Ridge1, has the form of $S' = \alpha S + \beta I$. Ledoit and Wolf (2004) studied how to obtain α and β to minimize L_2 loss. However, this estimator is not scale-invariant. That being said, changing the unit of the data results in the different result in the classifier. To remedy this, Li et al. (2025) proposed the estimator $S' = \alpha S + \beta \mathrm{diag}(S)$. We call this as Ridge2 and use the coefficients suggested in Li et al. (2025). We then take the inverse of them to obtain the estimators of precision matrix.

2.2. Estimation of mean vector

Estimation of the mean vector is another important component of Fisher linear discriminant analysis (LDA). In many cases, the estimation of the covariance matrix has been focused on high-dimensional discriminant analysis, and the sample means vector has been commonly used. Shrinkage estimator of the mean vector in high dimensional classification has been seriously considered under the assumption that the covariance matrix is a diagonal matrix, which is a modified form of the naive Bayes rule. Greenshtein and Park (2009) and Dicker and Zhao (2016) are representative work applying f-modeling and g-modeling to discriminant analysis under the assumption that all variables are independent.

When we have independent random variables $Y_i \sim N(\mu_i, 1)$ for $1 \le i \le p$, a lot of research has been done on the simultaneous estimation of the mean vector, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$. A typical example is the James–Stein estimator. Starting from James (1961), this type of shrinkage estimator has been developed in many aspects, including high dimensional settings. See Chételat and Wells (2012) and Bodnar et al. (2019) for detail. However, from an empirical Bayesian point of view, it is an estimator under the assumption that each mean parameter is generated from a normal distribution (Efron and Morris, 1973), so it has the disadvantage of not reflecting the various structures of the mean values such as bi-modality of mean values which may occur for the case of the sparsity of mean values. For more detailed information, refer to the paper by Jiang and Zhang (2009).

Additionally, even though we coerce them into the discriminant rules, it becomes problematic. Suppose that $Y_{ij} \sim N(\mu_{ij}, 1)$ for $i = 1, 2$ and $1 \le j \le p$. The key ideas used in the papers which dealt with the James–Stein estimator are to use $c(\boldsymbol{Y}_2 - \boldsymbol{Y}_1)$ or $c(\boldsymbol{Y}_2 - \boldsymbol{Y}_1) + d\boldsymbol{\mu}_0$ for estimating $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. When there are multiple number of observations, the sample mean is used with adequate scaling. One can check that the former case becomes identical to the case where the difference in sample mean is used when plugged into the linear discriminant rule. The latter one lacks the justification of choosing $\boldsymbol{\mu}_0$. Hence, we exclude the James–Stein type estimators for the rest of our discussion.

When p is large, there are two typical methods in nonparametric empirical Bayes methods discussed in Efron (2014) : f-modeling based on the estimation of marginal density and g-modeling based on the estimation of mixing distribution or prior distribution as a non-parametric method. The detail of these two methods are discussed shortly. In particular, f-modeling has a simpler calculation process than g-modeling, however with the recent development of various algorithms that can be used for g-modeling, empirical Bayesian methods based on g-modeling are also widely used.

In fact, when estimating the mean vector via these methods, the observed values are assumed to be independent. However, in classification, this assumption is not satisfied in general. Although there exist correlations among all variables, existing studies such as Greenshtein and Park (2009) applied f-modeling method to correlated variables in high dimensional classification problems. Instead, since we can obtain an estimate of the precision matrix following the previous section, we can remove the correlation among the variables using the square root matrix of this obtained estimate.

More specifically, consider $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$ with $i = 1, \dots, n$ and $p \times p$ positive definite matrix $\Omega^{1/2}$ where $\Omega = \Sigma^{-1}$. We consider the random variables $\boldsymbol{Z}_i$ as follows:

$$\boldsymbol{Z}_i \equiv \Omega^{1/2} \boldsymbol{X}_i \sim N_p(\Omega^{1/2}\boldsymbol{\mu}, I_p) \equiv N_p(\boldsymbol{\mu}^*, I_p), \tag{8}$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^T$, $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_p^*)^T$, $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^T$ and $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{ip})^T$. Therefore, the jth component of $\boldsymbol{Z}_i$, Z_{ij}, satisfies $Z_{ij} \sim N(\mu_j^*, 1)$ and is independently distributed. We estimate the mean vector $\boldsymbol{\mu}_i^*$ based on decorrelated observations $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{ip})^T$ and then transform back to $\hat{\boldsymbol{\mu}}_i = R^{-1}\hat{\boldsymbol{\mu}}_i^*$, aligned with the assumption of f-modeling and g-modeling.

Now we provide f- and g-modeling based estimation methods in detail. We consider one dimensional random variable Z and its observed value $z_1, \dots, z_n$, under the hierarchical structure

$$Z_i \sim N(\mu_i, 1), \quad \mu_i \sim G. \tag{9}$$

Here, G is a cumulative distribution function of μ. Under the Bayesian scheme, the probability distribution function of Z, g^*, is written as

$$g^*(z) = \int \phi(z - v)\, dG(v), \tag{10}$$

where $\phi(\cdot)$ is the probability distribution function of standard normal distribution, say $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Under this structure, the Bayes estimator for μ_i is

$$E(\mu_i|z_i) = z_i + \frac{(g^*)'(z_i)}{g^*(z_i)}, \tag{11}$$

where $(g^*)'(z) = dg^*(z)/dz$.

There are two categories of estimator for $E(\mu_i|z_i)$: First one is $\widehat{g^*(z)} = \int \phi(z - v) d\hat{G}(v)$ with an estimator for G, called g-modeling. See Jiang and Zhang (2009) and Efron (2014). The other one is estimating g^* directly by using density estimation based on the observed values $z_1, \dots, z_p$, called f-modeling. See Greenshtein and Park (2009). We denote the first method as 'NPMLE' (NonParametric Maximum Likelihood Estimation) since we would estimate G with maximum likelihood estimator (MLE), and the latter one as 'NPEB' (NonParametric Empirical Bayes).

To begin with, the NPMLE method requires G to be estimated and MLE can be used as the solution of the optimization problem

$$\hat{G} = \underset{F \in \mathcal{F}}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left\{ \int \phi(z_i - \mu) dF(\mu) \right\}, \tag{12}$$

where $\mathcal{F}$ is a set of distribution functions. It cannot be solved in its original form since $\mathcal{F}$ is infinite-dimensional. Therefore, we must constrain this as a finite-dimensional one. Koenker and Mizera (2014) restricted $\mathcal{F}$ into the set of piecewise constant distribution functions with $K+1$ regular grid points and converted this problem into K dimensional convex optimization problem. See Dicker and Zhao (2016) for details. The well-performing behavior of $\hat{G}$ via this algorithm, even with a relatively small number of K ($K \approx \sqrt{n}$), is justified in Dicker and Zhao (2016). This algorithm can be implemented in R via `REBayes` package.

In our data setting (8), we use the notation of z_{ij} to denote the jth element of the $\boldsymbol{z}_i$ vector. Let $\bar{z}_j = \sum_{k=1}^{n} z_{kj}/n_g$, the jth component of the sample mean vector $\bar{\boldsymbol{z}}$. Note that $\sqrt{n}\bar{z}_j \sim N(\sqrt{n}\mu_j^*, 1)$. Therefore, we substitute $\sqrt{n}\bar{z}_i$ for z_i in (12) to eventually obtain the estimate for $\sqrt{n}\boldsymbol{\mu}_j^*$ and divide it by $\sqrt{n}$.

Subsequently, f-modeling method estimates $(g^*)'(z_i)$ and $g^*(z_i)$ of (11) using the observed data z_i, $i = 1, \dots, n$. This paper uses kernel estimators with normal kernels. See section 2 of Brown and Greenshtein (2009) for detail. We present the eventual form of the estimator $\hat{\mu}_{EB}$ as follows.

$$(\hat{\boldsymbol{\mu}}_{EB})_i = \widehat{E(\boldsymbol{\mu}_i|\bar{\boldsymbol{z}})} = \bar{z}_i + \frac{\sum_{j=1}^{p} (\bar{z}_j - \bar{z}_i)\, \phi\left\{\sqrt{n}(\bar{z}_i - \bar{z}_i)/h\right\}}{h^2 \sum_{j=1}^{p} \phi\left\{\sqrt{n}(\bar{z}_i - \bar{z}_i)/h\right\}}. \tag{13}$$

Here, $(\hat{\mu}_{EB})_i$ denotes the ith component of $\hat{\boldsymbol{\mu}}_{EB}$, and h the bandwidth of the kernel estimator. Throughout our paper, we use $h = 1/\sqrt{\log p}$, with which good theoretical properties are known to hold (see Brown and Greenshtein (2009)). It is worth noting that both estimation methods provide the estimator of $\boldsymbol{\mu}^*$ solely from the sample mean $\bar{\boldsymbol{z}}$.

2.3. Construction of discriminant rules

First, recall that when the prior probability of group 1 and group 2 is provided, the Bayes discriminant rule δ has the form of

$$\delta(\boldsymbol{x}^{new}) = \left(\boldsymbol{x}^{new} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^{\top} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \log \frac{p_1}{p_2}, \tag{14}$$

where p_1, p_2 denotes the prior probability of each group. The only difference from (1) is the last term, which denotes the prior odds. In real data situations, when n_1, n_2 samples are obtained from group 1 and group 2, respectively, prior odds can be estimated as n_1/n_2. As discussed in (3), we can write

$$\delta(\boldsymbol{x}^{new}) = (\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)^{\top} \left(\boldsymbol{z}^{new} - \frac{\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*}{2}\right) - \log \frac{p_1}{p_2}. \tag{15}$$

As we previously mentioned, we first estimate the precision matrix, decorrelate the data (including $\boldsymbol{x}_{new}$), and then estimate the decorrelated mean vector, μ^*. For the classification, we would have two groups of datasets. In each group, we obtain the estimator for the precision matrix ($\hat{\Omega}_1$, $\hat{\Omega}_2$). Then, we use the pooled estimator as a final precision estimator. Namely,

$$\hat{\Sigma}^{-1} = \hat{\Omega} = \frac{(n_1 - 1)\hat{\Omega}_1 + (n_2 - 1)\hat{\Omega}_2}{n_1 + n_2 - 2}.$$

For $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \Omega^{-1})$, one can write

$$\hat{\Omega}^{1/2}\boldsymbol{X}_i \sim N_p(\hat{\Omega}^{1/2}\boldsymbol{\mu}, \hat{\Omega}^{1/2}\Omega^{-1}\hat{\Omega}^{1/2}),\ i = 1, \dots, n.$$

Assuming that $\hat{\Omega}$ estimates Ω properly, $\hat{\Omega}^{1/2}\Omega^{-1}\hat{\Omega}^{1/2}$ can be assumed to be similar to I_p. Therefore, we can apply NPEB and NPMLE methods under desirable settings, to finally get the estimator of $\hat{\Omega}^{1/2}\boldsymbol{\mu}$. Since we have two groups of data, we denote the decorrelated mean of group i and its estimate as $\boldsymbol{\mu}_i^*$ and $\widehat{\boldsymbol{\mu}_i^*}$, respectively ($i = 1, 2$).

Following the procedure, one can obtain the practical version of the discriminant rule as

$$\delta(\boldsymbol{x}^{new}) = \left(\widehat{\boldsymbol{\mu}_1^*} - \widehat{\boldsymbol{\mu}_2^*}\right)^\top \left(\hat{\Omega}^{1/2}\boldsymbol{x}^{new} - \frac{\widehat{\boldsymbol{\mu}_1^*} + \widehat{\boldsymbol{\mu}_2^*}}{2}\right) - \log\frac{p_1}{p_2}. \tag{16}$$

Instead of estimating $\boldsymbol{\mu}_1^*$ and $\boldsymbol{\mu}_2^*$ separately, we can estimate $\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ at once. When $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ denotes the decorrelated sample mean (with true precision matrix) of the group 1 and 2,

$$\boldsymbol{z}_1 \sim N_p(\boldsymbol{\mu}_1^*, {n_1}^{-1} I_p),\ \boldsymbol{z}_2 \sim N_p(\boldsymbol{\mu}_2^*, {n_2}^{-1} I_p).$$

Since $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ are independent,

$$\boldsymbol{z}_1 - \boldsymbol{z}_2 \sim N_p\left(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*, \left(n_1^{-1} + n_2^{-1}\right) I_p\right).$$

Therefore, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ can be directly estimated from $\boldsymbol{z}_1 - \boldsymbol{z}_2$ using NPEB and NPMLE method by multiplying and dividing the constant $a_n = (1/n_1 + 1/n_2)^{-1/2}$. Since the mean difference is considered to be the main key to separate two groups, we can simply plug in $\boldsymbol{z}_1, \boldsymbol{z}_2$ for $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ in $\left(\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*\right)/2$ of (15). Therefore, the classifier built on this method is written as

$$\delta(\boldsymbol{x}^{new}) = \left(\widehat{\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*}\right)^\top \left(\hat{\Omega}^{1/2}\boldsymbol{x}^{new} - \frac{\boldsymbol{z}_1^* + \boldsymbol{z}_2^*}{2}\right) - \log\frac{p_1}{p_2}. \tag{17}$$

In practice, we noticed that two discriminant rules show similar performance when the same estimation strategies are used. Hence, aligned with the theoretical assumptions in Theorem 1 presented in Section 3, we use (17) as a discriminant rule. Note that (16) and (17) become identical when we choose sample mean vectors as estimators of mean vectors.

3. Asymptotic results of the NPEB method

In this section, we present the asymptotic result for the discriminant rule with the NPEB (Non-Parametric Empirical Bayes) method. Wang and Jiang (2018) studied the performance of the discriminant rule (17) when using sample mean. Park et al. (2022) studied the performance of (17) using the mean vector estimator with NPMLE (Non-Parametric Maximum Likelihood Estimation) in (11) and showed a region for the parameters related to the strength of mean values and the level of sparsity such that the error rate is asymptotically 0 as the dimension increases. We provide an analogous result for the NPEB method under the setting in Park et al. (2022) and compare the result with those for NPMLE, hard threshold, and naive Bayes rule (which uses the sample mean for (17)).

As in Park et al. (2022) and Wang and Jiang (2018), we assume that the precision matrix is consistently estimated so that the data are decorrelated in an appropriate way, for example, glasso method provides the uniformly consistent estimators of all components in the precision matrix. We emphasize the effects of mean vector estimations after data are decorrelated.

Under (17), assuming $\Sigma = I_p$ or the precision matrix is consistently estimated, the misclassification error rate of a discriminant rule δ is written as

$$\begin{aligned} P\left(\delta(x^{new}) > 0 \,\middle|\, x^{new} \in \text{group } 2\right) &= P\left((z^{new} - \frac{\bar{z}_1 + \bar{z}_2}{2})^T \hat{\mu}_D^* > 0 \,\middle|\, z^{new} \sim N_p(\mu_2^*, I)\right) \\ &= \Phi\left\{-\frac{1}{2}\left(\frac{\mu_D^{*T}\hat{\mu}_D^*}{\|\hat{\mu}_D^*\|_2} + \frac{(\bar{z}_1 + \bar{z}_2 - (\mu_1^* + \mu_2^*))^T \hat{\mu}_D^*}{\|\hat{\mu}_D^*\|_2}\right)\right\}, \end{aligned}$$

where $\mu_D^* = \mu_1^* - \mu_2^*$, $\hat{\mu}_D^* = \widehat{\mu_1^* - \mu_2^*}$. Since

$$\bar{z}_1 + \bar{z}_2 - (\mu_1^* + \mu_2^*) \sim N\left(0, a_n^{-2} I_p\right),$$

when n_1 and n_2 are fixed, $\bar{z}_1 + \bar{z}_2 - (\mu_1^* + \mu_2^*) = O_p(1)$ holds. Therefore, from the Cauchy–Schwarz inequality, we have

$$\frac{\left\{\bar{z}_1 + \bar{z}_2 - (\mu_1^* + \mu_2^*)\right\}^T \hat{\mu}_D^*}{\|\hat{\mu}_D^*\|_2} = O_p(1).$$

Thus, as $p \to \infty$, we have the result that $P\left(\delta(x^{new}) > 0 \,\middle|\, x^{new} \in \text{group } 2\right) \to 0$ is equivalent to

$$V := \frac{\mu_D^{*T}\hat{\mu}_D^*}{\|\hat{\mu}_D^*\|_2} \xrightarrow{\text{p}} \infty. \tag{18}$$

This implies that if $V \xrightarrow{p} \infty$ as $p \to \infty$, a given decision rule δ has the error rate 0 asymptotically. However, as $p/n \to \infty$, it is widely known that δ has the error rate $1/2$ asymptotically, which matches the intuition (see Bickel and Levina (2004)). Therefore, we consider the situation in which both the number of meaningful components (say, the number of 'signals') and the degree of signal increase in a positive power of p.

To implement different levels of sparsity of signals in high dimension as $p \to \infty$, we parametrize the number of nonzero mean differences and their sizes which are commonly used in the study of sparse signals in high dimensional data analysis. See Park et al. (2022) and Dicker and Zhao (2016) for the examples. More specifically, we set $\mu_1^* = (0, \dots, 0)$ and $\mu_2^* = (\boldsymbol{\Delta}_l^T, \mathbf{0}_{p-l}^T)$ where $\boldsymbol{\Delta}_l$ is the l dimensional column vector with all components Δ. Then we assign

$$\Delta = p^b, l = [p^a] \tag{19}$$

for $b > 0$ and $0 < a < 1$. Δ and l represent the strength of the signal for each component and the level of sparsity, respectively. As p increases, the sparsity level $l = [p^a]$ and the strength of signal Δ are represented as functions of p. This setting was used in the theoretical study by Park et al. (2022) and the simulation experiments by Dicker and Zhao (2016). This setting will include various forms, including cases where Δ converges to infinity or to zero, in other words, the strength of the signal Δ is very large or very small. This will be accompanied by the number of variables l resulting in comparisons of various methodologies for various cases.

We first set the range of interest for the constants a and b since if either a or b is too large, the two groups can be easily separated regardless of the method used for classification. Therefore, as mentioned in Park et al. (2022), we want to compare various methods only in the (a, b) range as follows:

$$(a, b) \in \mathcal{R} \equiv (T_1 \cup T_2)^c$$

where $T_1 = \left\{(a, b) \in \mathbb{R}^2, a + 2b \geq 1/2, 0 < a \leq 0.5\right\}$ and $T_2 = \left\{(a, b) \in \mathbb{R}^2 : 0.5 < a \leq 1,\ b > 0\right\}$. In T_1 or T_2, the combination of Δ and l provides large strengths of signals leading to clear separation of two groups by various methods. See more details in Park et al. (2022). We consider the outside of T_1 and T_2 which is $(T_1 \cup T_2)^c$ in comparison of various methods.

Note that if a discriminant rule has the property of (18) in a wider range of (a, b) than another one, then the first discriminant rule has superiority to the other one asymptotically. Park et al. (2022) derived some conditions on (a, b) under which the classifier (17) using the NPMLE, the sample mean, and hard threshold method achieves (18). Park et al. (2022) showed that the discriminant rule with NPMLE has an advantage over naive Bayes rule and hard threshold in that (18) of NPMLE method is obtained in a wider range than the other two methods.

Here, we suggest the range of parameters to achieve (18) in the NPEB method and compare this with the regions of NPMLE, the sample mean(SM), and hard threshold(Hard) methodologies in the following theorem. We first define $\mathcal{R}_{NPEB}, \mathcal{R}_{NPMLE}, \mathcal{R}_{SM}$ and $\mathcal{R}_{Hard}$ as the areas of (a, b) satisfying (18) under the corresponding mean estimation methods.

Theorem 1. *Suppose* $\mu_1 = (0, \dots, 0)$ *and* $\mu_2 = (\boldsymbol{\Delta}_l^T, \mathbf{0}_{p-l}^T)$ *where* $\Delta = p^b, l = [p^a]$. *Then,* $\mathcal{R}_{NPEB}, \mathcal{R}_{NPMLE}$ *cover both* $\mathcal{R}_{SM}$ *and* $\mathcal{R}_{Hard}$. *Specifically,*

$$\mathcal{R}_{Hard} = C, \mathcal{R}_{SM} = D,$$
$$\mathcal{R}_{NPEB} \supset A \cup C \cup D, \mathcal{R}_{NPMLE} \supset B \cup C \cup D,$$

where

$$A = \left\{(a, b) \in \mathbb{R}^2,\ 0 \leq a + 2b \leq 1/2 : -0.25 < b < 0 \text{ and } 0 < a < 1\right\},$$
$$B = \left\{(a, b) \in \mathbb{R}^2,\ a + b \geq 1/2 : -0.5 < b < 0 \text{ and } 0.5 < a < 1\right\},$$
$$C = \left\{(a, b) \in \mathbb{R}^2,\ a + 2b \leq 1/2 : 0 < b < 0.25 \text{ and } 0 < a < 0.5\right\},$$
$$D = \left\{(a, b) \in \mathbb{R}^2,\ a + 2b \geq 1/2 : -0.25 < b < 0 \text{ and } 0.5 < a < 1\right\}.$$

Proof. See Appendix A. □

Remark 1. We would like to highlight the contribution of Theorem 1. There has not been any theoretical result on the linear discriminant rule based on f-modeling after it was first suggested in Efron (2009) and Greenshtein and Park (2009) while Park et al. (2022) provided the asymptotic results for g-modeling in high dimensional linear discriminant rule. Through Theorem 1, we shed a light on the comparison of two major different methods, f-modeling and g-modeling in high dimensional discriminant analysis.

Note that $\mathcal{R}_{NPEB}$ and $\mathcal{R}_{NPMLE}$ can be enlarged further, while $\mathcal{R}_{SM}$ and $\mathcal{R}_{Hard}$ exactly matches with D and C. The visual illustration of each region is in Fig. 1. We can see that the NPEB method covers all the areas of the naive Bayes rule and hard threshold method. On the other hand, neither $\mathcal{R}_{NPEB}$ nor $\mathcal{R}_{NPMLE}$ is included in each other. Both methods are expected to work well in both sparse and dense cases while the hard threshold method and naive Bayes rule are designed for only sparse and dense cases, respectively.

We provide numerical studies in Section 4 and real data analysis in Section 5 to compare these methods, and see that there does not seem to be any tendency that either of NPMLE and NPEB dominates the other method.

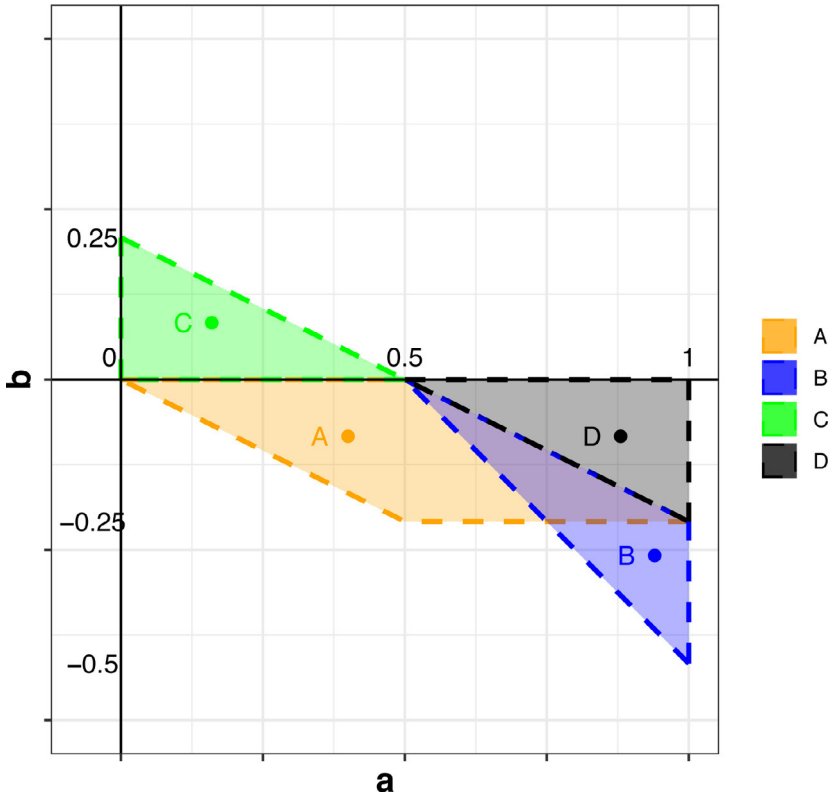


Fig. 1. The area of (a, b) under the setting (19).

4. Simulations

In this section, we provide simulation results for various combinations of mean vectors and covariance matrices. We divided simulation settings according to the sparseness and denseness of the precision matrix and those of the difference of mean vectors. One of our main contributions is that we provide systematic simulation studies and provide some guidelines on which combinations of estimations of mean vector and covariance matrix are desirable for numerous different situations.

Under the assumption of multivariate normal distribution, $\Omega_{ij} = 0$ means that X_i and X_j are conditionally independent. To verify the strength of glasso and LAM methods separately, we divided simulation settings according to the sparseness of the precision matrix or covariance matrix. To compare the mean vector estimation method individually, we assumed the simulation setting as follows. We set $\mu_1^T = (0, \ldots, 0)$, $\mu_2^T = (\boldsymbol{\Delta}_l^T, 0_{p-l}^T)$ where $\boldsymbol{\Delta}_l$ is the l dimensional column vector with all component Δ and $\mathbf{0}_{p-l}$ is the $(p-l)$ dimensional vector with all 0 components. In some situation, we set $\mu_{2s} = \Sigma^{1/2}\mu_2 = (\boldsymbol{\Delta}_l^T, \mathbf{0}_{p-l}^T)$. Throughout the simulation, we set $p = 500$ with the number of training data and test data as $n_{1,train} = n_{2,train} = 50$ and $n_{1,test} = n_{2,test} = 250$.

We consider the following simulation settings for various configurations of l and Δ as well as different structure of Σ.

1. Setting 1: AR(1) structured precision matrix. We set Σ^{-1} as $(\Sigma^{-1})_{ij} = \rho^{|i-j|}$.
   -simulation 1-1: $\Delta = 3, l = 20$ in μ_{2s} with $\rho = 0.8$.
   -simulation 1-2: $\Delta = 0.4, l = 400$ in μ_{2s} with $\rho = 0.8$.
2. Setting 2: Blocked AR(1) structured precision matrix. We set Σ^{-1} as $\begin{bmatrix} \Sigma_q^{-1} & 0 \\ 0 & I_{p-q} \end{bmatrix}$, where $(\Sigma_q^{-1})_{ij} = \rho^{|i-j|}$ for $1 \le i, j \le q$.
   -simulation 2-1: $\Delta = 3, l = 20$ in μ_{2s} with $\rho = 0.9, q = 50$.
   -simulation 2-2: $\Delta = 0.15, l = 400$ in μ_{2s} with $\rho = 0.9, q = 50$.
3. Setting 3: Exchangeable precision matrix. We set Σ^{-1} as $(\Sigma^{-1})_{ij}$ to have 1 as a diagonal components and ρ as all off-diagonal elements.
   -simulation 3-1: $\Delta = 0.7, l = 20$ in μ_2 with $\rho = 0.3$.
   -simulation 3-2: $\Delta = 0.5$, $l = 20$ in μ_{2s} with $\rho = 0.3$.
4. Setting 4: Toeplitz covariance matrix. We set Σ as $\Sigma_{ij} = 1 / (|i - j| + 1)$.
   -simulation 4-1: $\Delta = 1, l = 20$ in μ_2.
   -simulation 4-2: $\Delta = 0.4, l = 20$ in μ_{2s}.

Table 1
Simulation 1-1 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.0000	0.1145	0.0978	0.1197	0.1048	0.1311
	(0.0000)	(0.0142)	(0.0129)	(0.0135)	(0.0131)	(0.0145)
NPMLE	0.0000	0.1137	0.0988	0.1274	0.1106	0.1313
	(0.0000)	(0.0141)	(0.0128)	(0.0140)	(0.0133)	(0.0145)
SM	0.0000	0.1156	0.1568	0.1547	0.1643	0.1423
	(0.0000)	(0.0142)	(0.0159)	(0.0159)	(0.0160)	(0.0154)

Table 2
Simulation 1-2 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.0000	0.3960	0.2239	0.2255	0.1563	0.4140
	(0.0002)	(0.0221)	(0.0176)	(0.0182)	(0.0161)	(0.0223)
NPMLE	0.0000	0.3960	0.0002	0.0007	0.0002	0.4138
	(0.0003)	(0.0219)	(0.0006)	(0.0011)	(0.0006)	(0.0225)
SM	0.0001	0.3970	0.4246	0.4299	0.4251	0.4218
	(0.0005)	(0.0219)	(0.0229)	(0.0225)	(0.0233)	(0.0224)

Table 3
Simulation 2-1 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.0000	0.0905	0.1705	0.1944	0.1157	0.2270
	(0.0000)	(0.0120)	(0.0156)	(0.0164)	(0.0141)	(0.0173)
NPMLE	0.0000	0.0904	0.1721	0.1971	0.1122	0.2272
	(0.0000)	(0.0120)	(0.0159)	(0.0163)	(0.0139)	(0.0173)
SM	0.0000	0.0918	0.1669	0.2350	0.1139	0.2351
	(0.0000)	(0.0123)	(0.0159)	(0.0182)	(0.0141)	(0.0179)

Table 4
Simulation 2-2 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.0991	0.3197	0.2179	0.1194	0.3418	0.2082
	(0.0138)	(0.0211)	(0.0165)	(0.0141)	(0.0207)	(0.0172)
NPMLE	0.0874	0.3192	0.2019	0.1026	0.3365	0.2063
	(0.0125)	(0.0207)	(0.0163)	(0.0132)	(0.0213)	(0.0170)
SM	0.1812	0.3211	0.2825	0.2165	0.3617	0.2170
	(0.0164)	(0.0207)	(0.0200)	(0.0178)	(0.0212)	(0.0176)

5. Setting 5: Banded covariance matrix. We set Σ as $\Sigma_{ij} = \max(1 - |i-j| / 10, 0)$.
   -simulation 5-1: $\Delta = 0.7, l = 20$ in μ_2.
   -simulation 5-2: $\Delta = 2$, $l = 20$ in μ_{2s}.

We present misclassification error rates in Table 1 to Table 10, in which the columns and rows denote the precision matrix and mean vector estimation methods, respectively. To assess the performance of classification, we consider various combinations of mean vector and precision matrix estimations. Each row in the tables corresponds to mean vector estimation, sequentially represented by Non-parametric Empirical Bayes (NPEB) methods, Non-parametric Maximum Likelihood Estimation (NPMLE), and Sample mean (SM) method. Meanwhile, each column represents precision matrix estimation methods, including the actual precision matrix (True.prec), estimation based on graphical models introduced by Friedman et al. (2008) (glasso), estimation based on random sample data splitting proposed by Lam (2016), estimation assuming independent rule using the inverse of the sample variance of each feature (IR), and methods based on Ridge shrinkage proposed by Ledoit and Wolf (2004) and Li et al. (2025), named (Ridge1) and (Ridge2), respectively. In particular, True.prec applies the true precision matrix, so it is not a method for estimating the precision matrix. However, we consider it to verify the effect of the precision matrix estimation. Specially, the combinations of SM method with Ridge1 and Ridge2 method were studied in Hyodo et al. (2012) and Li et al. (2025), respectively.

Table 5
Simulation 3-1 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.0000	0.2508	0.1493	0.1590	0.1481	0.2398
	(0.0003)	(0.0199)	(0.0159)	(0.0168)	(0.0164)	(0.0191)
NPMLE	0.0000	0.2509	0.1315	0.1419	0.1342	0.2382
	(0.0003)	(0.0199)	(0.0147)	(0.0154)	(0.0150)	(0.0189)
SM	0.0001	0.2544	0.2356	0.2408	0.2384	0.2491
	(0.0005)	(0.0201)	(0.0184)	(0.0189)	(0.0186)	(0.0194)

Table 6
Simulation 3-2 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.2423	0.3125	0.2468	0.2581	0.2546	0.3055
	(0.0181)	(0.0207)	(0.0185)	(0.0196)	(0.0192)	(0.0202)
NPMLE	0.2212	0.3127	0.2187	0.2326	0.2315	0.3060
	(0.0189)	(0.0206)	(0.0184)	(0.0191)	(0.0189)	(0.0202)
SM	0.2911	0.3145	0.2969	0.3020	0.3004	0.3099
	(0.0194)	(0.0208)	(0.0196)	(0.0199)	(0.0199)	(0.0200)

Table 7
Simulation 4-1 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.2248	0.2041	0.1596	0.1591	0.1587	0.1942
	(0.0180)	(0.0180)	(0.0171)	(0.0168)	(0.0170)	(0.0186)
NPMLE	0.2081	0.2034	0.1564	0.1579	0.1556	0.1945
	(0.0185)	(0.0182)	(0.0165)	(0.0168)	(0.0167)	(0.0182)
SM	0.2724	0.2073	0.1965	0.1930	0.1948	0.2014
	(0.0186)	(0.0183)	(0.0181)	(0.0175)	(0.0180)	(0.0181)

Table 8
Simulation 4-2 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.3257	0.2652	0.2192	0.2135	0.2152	0.2571
	(0.0200)	(0.0197)	(0.0184)	(0.0183)	(0.0184)	(0.0189)
NPMLE	0.3027	0.2659	0.2129	0.2122	0.2107	0.2568
	(0.0204)	(0.0196)	(0.0182)	(0.0186)	(0.0191)	(0.0192)
SM	0.3539	0.2684	0.2574	0.2502	0.2546	0.2630
	(0.0205)	(0.0197)	(0.0196)	(0.0196)	(0.0194)	(0.0194)

Table 9
Simulation 5-1 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.0000	0.4904	0.4930	0.4957	0.4896	0.4684
	(0.0000)	(0.0226)	(0.0232)	(0.0225)	(0.0221)	(0.0231)
NPMLE	0.0000	0.4908	0.1851	0.4897	0.1756	0.4692
	(0.0000)	(0.0228)	(0.0170)	(0.0223)	(0.0165)	(0.0230)
SM	0.0000	0.4906	0.4932	0.4961	0.4946	0.4745
	(0.0002)	(0.0227)	(0.0230)	(0.0239)	(0.0228)	(0.0228)

- *Impact of decorrelation of the LAM on the mean vector estimations:*
  From all the results from the LAM and the IR, we noticed that mean vector estimation methods have a consistent impact on the error rates. Especially the SM method is improved by the NPMLE, and the NPEB since the error rates of the NPEB and the NPMLE methods are smaller than those of the SM.

Table 10
Simulation 5-2 result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$					
	True.prec	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	0.0000	0.4502	0.4050	0.4245	0.3783	0.3775
	(0.0001)	(0.0236)	(0.0213)	(0.0211)	(0.0206)	(0.0223)
NPMLE	0.0000	0.4504	0.3397	0.3692	0.3343	0.3770
	(0.0000)	(0.0233)	(0.0201)	(0.0218)	(0.0208)	(0.0224)
SM	0.0000	0.4518	0.4624	0.4698	0.4655	0.3999
	(0.0002)	(0.0234)	(0.0234)	(0.0237)	(0.0232)	(0.0232)

- *Impact of decorrelation using glasso on the mean vector estimations:*
  When the precision matrix is estimated by glasso, the mean vector estimation itself did not show a clear impact on the error rate. Table 2 illustrates this tendency. For the LAM method, the error rate differs from 0.00 to 0.42 according to mean estimation methods, whereas it remains almost the same on glasso based discriminant rule.
- *Impact of glasso, LAM, Ridge1 and Ridge2:*
  The glasso is designed for sparse structure while the other precision estimators (LAM, Ridge1, Ridge2) are designed for more general structure. Therefore, the performance of the discriminant rules varies depending on whether the structural assumptions are satisfied or not. For example, in simulation 2-1, the glasso method tends to produce smaller error rates than those from the other methods. Simulation 2-1 shows the sparse structure of the precision matrix since the AR(1) structure is the banded matrix. On the other hand, in situation 3-1, using the precision matrix with all the same off-diagonal terms, the LAM and Ridge1 methods perform better than glasso method as the precision matrix in simulation 3-1 is a dense matrix. These two simulations coincide with the structural assumptions on which the glasso and the other estimators are suggested, such as sparsity.
- *No uniform dominance among precision estimation strategies:*
  Comparing the results in simulations 2-1 and 2-2, one can notice that even though two situations are generated with the same precision matrix, a tendency in error rates is dissimilar. In situation 2-1, the linear discriminant rule using the graphical lasso method shows the lowest error rate with every mean estimation strategy, which is completely opposite to simulation 2-2. Likewise, the discriminant rules based on Ridge1 perform better than those based on Ridge2 in simulation 2-1, while the opposite result happens in simulation 2-2. Therefore, the dominant precision estimation strategy for the linear discriminant rule is not determined depending on the precision matrix only.
- *Comparison with theoretical result:*
  The theoretical result can elucidate the significant dominance of the NPMLE and NPEB method over the SM method in most of the simulation results under the IR and LAM method for the estimation of the precision matrix. According to Theorem 1, when there is a dense signal with low signal intensity ($5/6 < a < 1$, $b < 0$), the NPMLE and NPEB methods ensure a wider range of b in which the asymptotic error rate is zero compared to SM method. One can check that simulation 2-2 nearly falls into this area. The error rates in Table 4 supports the theorem.
- *Impact of mean vector estimation by shrinkage:*
  From the results in Tables 1 to 10, we can observe conspicuous improvements in the SM method in error rates by using the NPMLE and NPEB methods for any given estimation of the precision matrix. When the sparse signals have an intensity that is proportional to the positive power of p, namely, when $0 < a < 1/2$, $0 < b < 1/4$, and $a + 2b < 1/2$ holds, the linear discriminant rules built from the SM method are known to asymptotically act as random guessing, while the discriminant rules using NPEB and NPMLE have the error rate which asymptotically converges to zero.

5. Real data analysis

In this section, we consider five real data sets to compare the performance of various binary classification methods discussed in this paper. Through these real data examples, we investigate the role of mean vector and precision matrix estimation in high-dimensional classification.

Four real datasets are used for comparing the performance of discriminant rules.

- EEG data: 122 observations (77: group 1, 45: group 2) with 512 features
  We used EEG (Electroencephalography) data to validate linear discriminant rules under multiple estimation methods (which is available at https://archive.ics.uci.edu/ml/datasets/eeg+database). The data was initially generated for the large study about the genetic predisposition of alcoholism. Measurements from 64 electrodes placed on the subject's scalps were sampled at 256 Hz for 1 s. For data pre-processing, we parsed this time series data into 8 intervals and extracted each interval's median value. Therefore, we finally made the data of $64 \times 8 = 512$ dimensional vectors from each subject. With 122 observations in sum ($n_1 = 77$ subjects from the alcoholic group and $n_2 = 45$ from the control group), we examined the performance of each discriminant rule using the LOOCV method.

Table 11
EEG data result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$				
	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	22/122	25/122	27/122	25/122	27/122
NPMLE	23/122	26/122	27/122	26/122	27/122
SM	24/122	24/122	28/122	26/122	27/122

- Gravier data: 168 observations (111: group 1, 57: group 2) with 1000 features
  Gene array data obtained for the study pertaining to the prediction of metastasis of small node-negative breast carcinoma is also utilized, which we named Gravier data. Small invasive ductal carcinomas without axillary lymph node involvement (T1T2N0) were analyzed from 168 patients. Among these, 111 subjects who had no events for 5 years afterward were categorized as group 1, and the other 57 subjects with early metastasis were assigned as group 2. We processed the original data by extracting 1000 features that have the most significant t values among the 2905 features beforehand.
- Ham data: 214 observations (111: group 1, 103: group 2) with 431 features.
  Ham data was obtained by food spectrograph about 19 Spanish and 18 French dry-cured hams. A food spectrograph is utilized in chemometrics to classify food types, which can be directly used to assure food safety and quality. This data is publicly available at http://www.timeseriesclassification.com/description.php?Dataset=Ham. From 37 hams in total, up to 6 observations were obtained. Therefore, 214 observations were categorized into two groups (111 observations for group 1 and 103 for group 2), with 431 feature vectors for each observation. The data preprocessing procedure is described in 'Sodium dodecyl sulphate-polyacrylamide gel electrophoresis of proteins in dry-cured hams: Data registration and multivariate analysis across multiple gels' (to be in reference).
- IMVigor data: 298 observations (230: group 1, 68: group 2) with 4792 features (divided into 7 groups)
  Clinical outcomes of metastatic urothelial cancer patients were collected and are available at http://research-pub.gene.com/IMvigor210CoreBiologies/ (see Mariathasan et al. (2018)). Observations about 230 non-responders and 68 responders are composed of 4792 feature vectors. These features are known to be correlated within seven groups, which contain 1583, 975, 569, 548, 546, 341, and 230 features each. However, we do not utilize this information when constructing the discriminant rule.

Within these data, classification performance has been measured through Leave-One-Out Cross Validation (LOOCV) method. As introduced in Section 2.3, linear discriminant rules can be constructed by estimating μ_1^* and μ_2^* respectively, or $\mu_2^* - \mu_1^*$ at once. Therefore, we compared the error rate of each discriminant rule according to the mean estimation method, precision estimation method, and discriminant rule construction method.

As one can see from Tables 11–14, classification errors largely depend on precision matrix estimation methods. A more suitable precision estimation method relies on the data. For example, in HAM data, the LAM method shows a lower error rate, while glasso method performs well with EEG data. Ridge1 and Ridge2 methods showed similar performances to the LAM method across all datasets. Among mean vector estimation methods, error rates do not vary significantly in most cases. However, when classifying Gravier data with glasso method, the error rate varied significantly according to the mean estimation method. Therefore, we compared each component of $\widehat{\mu_2^* - \mu_1^*}$ among mean estimation methods.

Fig. 2 shows the plots of NPEB vs. SM and NPMLE vs. SM after two methods of decorrelation. The NPEB and NPMLE tend to have the shrinkage effect of the SM, especially for large SM values. It is seen that the NPMLE in (11) with the estimated $\hat{G}$ has the monotonicity property of SM. On the other hand, the NPEB with estimated marginal density $\hat{f}$ and $\hat{f}'$ is not guaranteed to have monotonicity, so there exist some wiggly patterns in local regions, however, overall patterns of NPEB are similar to those of NPMLE. In general, the posterior mean in (11) based on f-modeling is not guaranteed to have the monotonicity in z while the g-modeling such as plugging in estimated $\hat{G}$ into (11) guarantees the monotonicity in z. See the discussion on this issue in Ignatiadis and Wager (2022) for the Poisson model, which is also true for the case of normal mean estimation.

To summarize, the correlation, such as glasso and LAM methods, is effective in most of the cases in that the IR without decorrelation is improved although those two methods make the IR worse in some data sets. In particular, the glasso is sensitive to different cases since the glasso is designed for a sparse structured precision matrix. On the other hand, the LAM method is robust to all data sets except Gravier data in which the performance of all classification methods is worse than the glasso and the IR. For the glasso method, different classification methods have large variations in error rate while the LAM method produces quite similar error rates for different rules.

6. Concluding remarks

In this study, our research contributions can be broadly summarized into two main points. Firstly, numerous prior studies (Xu et al., 2009; Hyodo et al., 2012; Bodnar et al., 2019; Sifaou et al., 2020; Li et al., 2025) in Linear Discriminant Analysis (LDA) commonly employ covariance matrix estimation methods, substituting the mean parameter with a sample version. This prompts us to identify potential areas for enhancement in the mean parameter estimation method of the sample version within these earlier

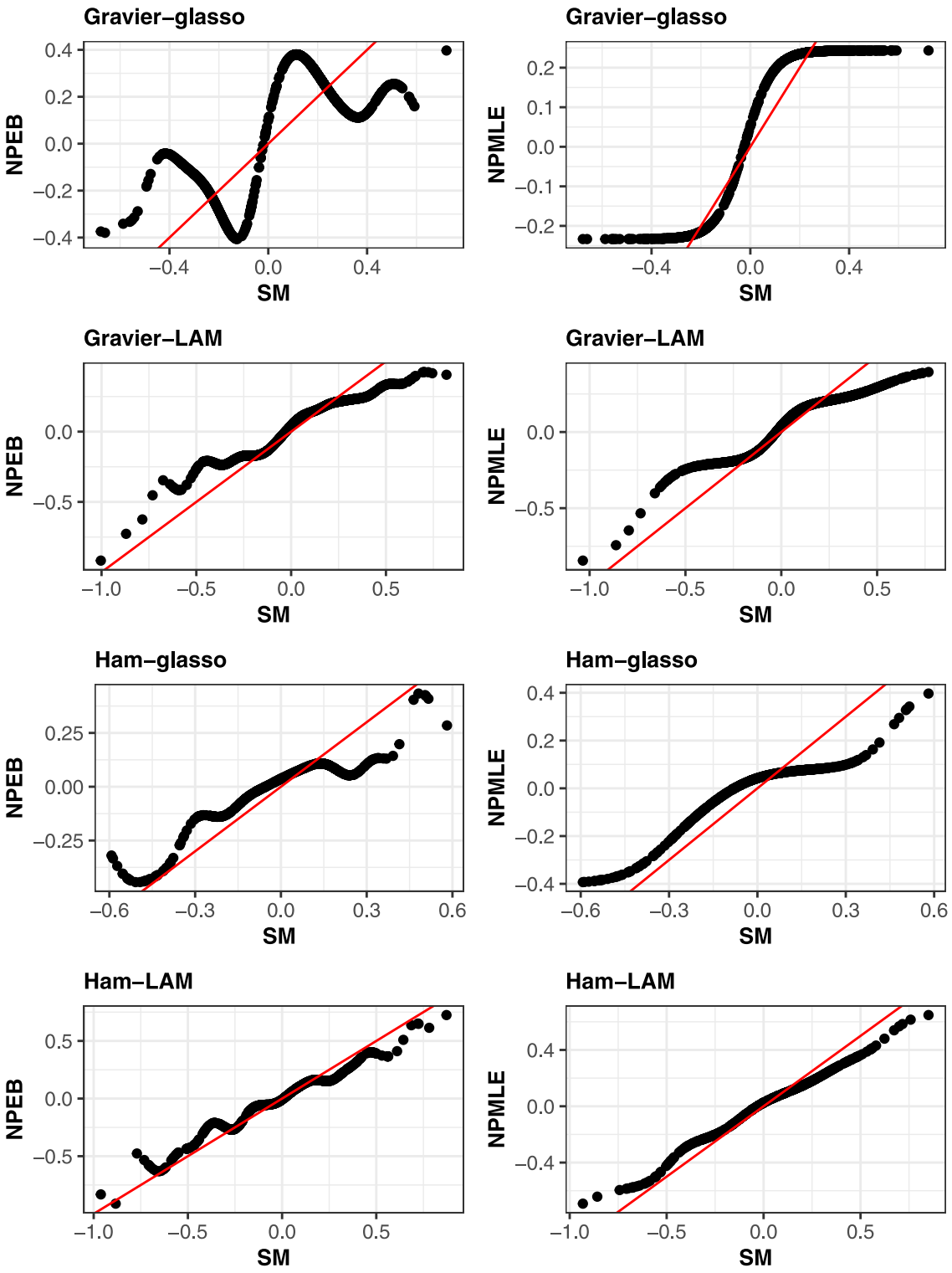


Fig. 2. $(\widehat{\mu_2^* - \mu_1^*})$ plot - NPEB, NPMLE method to SM method.

Table 12
Gravier data result.

$\hat{\mu}$	$\hat{\Sigma}^{-1}$				
	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	119/168	33/168	24/168	34/168	35/168
NPMLE	50/168	31/168	29/168	34/168	33/168
SM	32/168	33/168	31/168	33/168	35/168

Table 13
Ham data result.

	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	52/214	32/214	49/214	31/214	31/214
NPMLE	54/214	32/214	49/214	32/214	31/214
SM	52/214	30/214	50/214	31/214	30/214

Table 14
IMVigor data result.

	Glasso	LAM	IR	Ridge1	Ridge2
NPEB	156/298	71/298	111/298	73/298	69/298
NPMLE	130/298	70/298	112/298	74/298	68/298
SM	107/298	70/298	111/298	72/298	73/298

studies. In contrast, our methodology elevates LDA by standardizing the data with a robust non-parametric covariance estimator and non-parametrically estimating the mean parameters. This approach offers the advantage of enhancing LDA performance across various data structures through the non-parametric estimation of both the covariance matrix and mean parameters. By combining different strategies for estimating precision matrices and mean vectors, we explore the performance of various linear discriminant rules. We assess these rules under various controlled settings. Overall, the application of nonparametric estimation methods for the mean vector, in addition to the estimation of the precision matrix, significantly enhances the classification performance of LDA. Firstly, we theoretically demonstrated that, under the consistency of the covariance matrix estimator, our Nonparametric Empirical Bayes (NPEB) approach is more effective across a broader range of data structures compared to the sample mean estimator or threshold-based estimator. We believe that this is an interesting result in the sense that we can compare f-modeling and g-modeling theoretically, which correspond to the NPEB and NPMLE where the performance of the NPMLE is studied in Park et al. (2022). One interesting phenomenon is that the structure of the mean vector may be changed after decorrelation. Since our results in this paper show that the different methods of estimation of mean vectors have different performances depending on the structure of mean vectors, so it will be interesting to investigate the changes in mean vector structures after decorrelation and their effect on the choice of estimation methods. We leave this as a future work.

Acknowledgments

The authors are deeply grateful to the reviewers for their helpful and stimulating comments that led to the improvement of this paper. Hoyoung Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00212502) and additional funding was provided by Sookmyung Women's University Research, South Korea Grants (1-2303-2020). Research of J. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C1A01100526).

Appendix A. Lemmas for Theorem 1

We present the following four lemmas which are used in the proof of Theorem 1.

Lemma 1. *For $p > 1$,*

$$\left(1 - \frac{1}{p^2}\right)^p \geq 1 - \frac{1}{p}. \tag{20}$$

Proof. One can easily show that $f(x) = x \log\left(\frac{x+1}{x}\right)$ is an increasing function, by checking

$$f'(x) = \log\left(\frac{x+1}{x}\right) - \frac{1}{x+1}, \; f''(x) = -\frac{1}{x(x+1)^2} < 0$$

and $\lim_{x\to\infty} f'(x) = 0$. From $f(p) \geq f(p-1)$, one can obtain

$$p \log(1 - \frac{1}{p^2}) \geq \log(1 - \frac{1}{p}),$$

which concludes the proof. □

Lemma 2. *Let* $Z_1, \dots, Z_p \overset{\text{iid}}{\sim} N(0,1)$ *and* $\Delta = p^b$. *For* $b < 0$ *and the fixed constant* $a_n = \left(1/n_1 + 1/n_2\right)^{-1/2}$,

$$P\left(\max_{1\le j\le p} |Z_j| + a_n\Delta < \log p\right) > 1 - \frac{1}{p} \tag{21}$$

holds for large enough p.

Proof. Noting that for all $t > 0$, we have

$$\frac{t}{t^2+1}\phi(t) < 1 - \Phi(t) \le \frac{1}{t}\phi(t)$$

where ϕ, Φ denote the probability distribution function and the cumulative distribution function of standard normal distribution. Then we obtain

$$P\left(\max_{1\le j\le p} |Z_j| + a_n\Delta < \log p\right) = \left\{1 - 2\Phi(a_n\Delta - \log p)\right\}^p \ge \left\{1 - \frac{2\phi(\log p - a_n\Delta)}{\log p - a_n\Delta}\right\}^p. \tag{22}$$

Since $a_n\Delta \to 0$ and $\log p \to \infty$ as $p \to \infty$, we have $2a_n\Delta < \log p$ for large p, therefore (22) satisfies the followings:

$$\begin{aligned}\left\{1 - \frac{2\phi(\log p - a_n\Delta)}{\log p - a_n\Delta}\right\}^p &\ge \left[1 - \frac{4\phi\left\{(\log p)/2\right\}}{\log p}\right]^p \\ &= \left[1 - \frac{4}{\sqrt{2\pi}\log p}\exp\left\{-(\log p)^2/8\right\}\right]^p.\end{aligned}$$

From Lemma 1, for large p satisfying

$$\frac{4}{\sqrt{2\pi}\log p}\exp\left\{-\frac{(\log p)^2}{8}\right\} \le \frac{1}{p^2},$$

we obtain

$$P\left(\max_{1\le j\le p} |Z_j| + a_n\Delta < \log p\right) \ge \left(1 - \frac{1}{p^2}\right)^p \ge 1 - \frac{1}{p}. \quad \square$$

Lemma 3. *Let* $Z_i \sim N(0,1)$ *and* $\Delta = p^b$. *Then, for* $b < 0$ *and* $a_n = \left(1/n_1 + 1/n_2\right)^{-1/2}$, *we have*

$$P\left(|Z_i| + a_n\Delta < \sqrt{2\alpha(\log p + 1)}\right) > 1 - \frac{8}{\sqrt{2\alpha\log p}}\frac{e}{\sqrt{2\pi}p^\alpha} \tag{23}$$

for large enough p.

Proof. From the fact

$$\frac{t}{t^2+1}\phi(t) < 1 - \Phi(t) \le \frac{1}{t}\phi(t),$$

for large enough p which satisfy $a_n\sqrt{2\alpha(\log p + 1)} < p^{-b} = 1/\Delta$ and $\sqrt{2\alpha(\log p + 1)} - 1 > \sqrt{2\alpha\log p}/2$, we obtain

$$\begin{aligned}&P\left(|Z_i| + a_n\Delta < \sqrt{2\alpha(\log p + 1)}\right) \\ &= 1 - 2\Phi\left(\sqrt{2\alpha(\log p + 1)} - a_n\Delta\right) \\ &> 1 - \frac{2}{\sqrt{2\pi}}\frac{2}{\sqrt{2\alpha(\log p + 1)} - a_n\Delta}\exp\left\{-\frac{1}{2}(\sqrt{2\alpha(\log p + 1)} - a_n\Delta)^2\right\} \\ &> 1 - \frac{4}{\sqrt{2\alpha(\log p + 1)} - 1}\frac{1}{\sqrt{2\pi}p^\alpha}\exp\left\{a_n\Delta\sqrt{2\alpha(\log p + 1)}\right\} \\ &> 1 - \frac{8}{\sqrt{2\alpha\log p}}\frac{e}{\sqrt{2\pi}p^\alpha}. \quad \square\end{aligned}$$

Lemma 4. *For* $\widehat{\mu_{D,i}}^{EB}$ *defined in* (31), *we have the lower bound* W_i *such that*

$$\sum_{i=1}^{p_1}\widehat{\mu_{D,i}}^{EB} \ge \sum_{i=1}^{p_1} W_i \ge \frac{h^2}{1+h^2}a_n p^{a+b}\{1 + o(1)\} + o_p\left\{(\log p)^2\right\}$$

where

$$W_i = \begin{cases} W_{1i}, & \text{if } |Z_i| + a_n\Delta \le \sqrt{2\alpha(\log p + 1)} \\ W_{2i}, & o.w, \end{cases} \tag{24}$$

and

$$W_{1i} = \frac{h^2 Z_i^*}{1+h^2} - \frac{h^2}{1+h^2}\frac{(1+2\sqrt{2\pi}p^{1/2+\epsilon})\sqrt{2\alpha(\log p+1)}+(1+h^2)/h^4}{p^{1-\alpha}(1-p^{a-1})-2p^{1/2+\epsilon}},$$
$$W_{2i} = \Delta + \frac{1}{a_n}(1-\log p)Z_i - \frac{1}{a_n h^2}\log p.$$

Here, $Z_i^* = Z_i + a_n\Delta$.

Proof. As this proof contains several technical details, we first summarize the flow of the proof. We obtain the first inequality by introducing the auxiliary random variable V_i which satisfies $\widehat{\mu_{D,i}}^{EB} \geq V_i \geq W_i$. Note that the random variable W_i is independent. The second inequality is shown by using the weak law of large numbers (WLLN).

- First, we consider an auxiliary random variable V_i as the following one :

$$V_i = \begin{cases} V_{1i}, & \text{if } |Z_i| + a_n\Delta \leq \sqrt{2\alpha(\log p+1)} \\ V_{2i}, & \text{o.w} \end{cases} \tag{25}$$

where

$$V_{1i} = \Delta + \frac{Z_i}{a_n} - \frac{1}{h^2}\frac{\frac{1}{\sqrt{2\pi}}\frac{h^3}{(1+h^2)^{3/2}}\left\{(p_1-1)Z_iT_i + (p-p_1)Z_i^*T_i^*\right\} + 2p^{1/2+\epsilon}}{\frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}}\frac{h}{(1+h^2)^{1/2}}\left\{(p_1-1)T_i + (p-p_1)T_i^*\right\} - 2p^{1/2+\epsilon}},$$
$$V_{2i} = \Delta + \frac{Z_i}{a_n} + \frac{1}{a_n h^2}\min\left\{(Z_{ji}^-)_{1\leq j\leq p_1}, (Z_{ji}^- + a_n\Delta)_{p_1+1\leq j\leq p}\right\}.$$

Here, $Z_{ij}^-, Z_i^*, T_i, T_i^*$ are defined as follows.

$$Z_{ij}^- = Z_i - Z_j, \qquad Z_i^* = Z_i + a_n\Delta,$$
$$T_i = \exp\left\{-\frac{Z_i^2}{2(1+h^2)}\right\}, \quad T_i^* = \exp\left\{-\frac{Z_i^{*2}}{2(1+h^2)}\right\}.$$

We first notice that V_i is constructed to always satisfy $\widehat{\mu_{D,i}}^{EB} \geq V_{2i}$ (see the derivation of (35)), and $\widehat{\mu_{D,i}}^{EB} \geq V_{1i}$ on $E = \cap_{i=1}^{p_1} E_i$. E is the set of events in which Z_is satisfy a specific relationship, with $P(E) \to 1$ as $p \to \infty$. Here, we defer the detailed definition of E to Appendix B. This implies that $\widehat{\mu_{D,i}}^{EB}$ can be bounded by V_{1i} with large probability and otherwise, by V_{2i}.

- In addition, one can rewrite V_{1i} as

$$V_{1i} = \frac{h^2 Z_i^*}{1+h^2} + \frac{\frac{1}{\sqrt{2\pi}}\left(\frac{h^2 Z_i^*}{1+h^2} + \frac{h(p_1-1)a_n\Delta T_i}{(1+h^2)^{3/2}}\right) - 2p^{1/2+\epsilon}\left(\frac{h^2 Z_i^*}{1+h^2} + \frac{1}{h^2}\right)}{\frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}}\frac{h\left\{(p_1-1)T_i + (p-p_1)T_i^*\right\}}{(1+h^2)^{1/2}} - 2p^{1/2+\epsilon}}.$$

By plugging in the inequality $|Z_i| + a_n\Delta \leq \sqrt{2\alpha(\log p+1)}$ in V_{1i} and the condition from E_{5i} for V_{2i}, we obtain $W_i \leq V_i$ for W_i defined above.
Eventually, W_i serves as the lower bound of $\widehat{\mu_{D,i}}^{EB}$ which are mutually independent. Therefore, by applying WLLN for triangular arrays on W_i, we show the second inequality of the lemma.
For large enough p which satisfies

$$\log p \geq \sqrt{2\alpha(\log p+1)}, \tag{26}$$
$$|W_{1i}| \leq \frac{h^2}{1+h^2}\left\{\log p + \frac{(1+2\sqrt{2\pi}p^{1/2+\epsilon})\log p + (1+h^2)/h^4}{p^{1-\alpha}(1-p^{a-1})-2p^{1/2+\epsilon}}\right\}$$

and

$$|W_{2i}| \leq \Delta + 2(\log p)^2/a_n$$

holds on E since $\max_{1\leq i\leq n}|Z_i| < \log p$ and $1/h^2 = \log p$. Therefore,

$$|W_i| \leq 1 + 2(\log p)^2/a_n.$$

In addition, from (26), one can obtain

$$\begin{aligned} E[W_1] &= E\left[W_{1i}\,\middle|\,|Z_i| + a_n\Delta \leq \sqrt{2\alpha(\log p+1)}\right] \times P\left(|Z_i| + a_n\Delta \leq \sqrt{2\alpha(\log p+1)}\right) \\ &\quad + E\left[W_{2i}\,\middle|\,|Z_i| + a_n\Delta > \sqrt{2\alpha(\log p+1)}\right] \times P\left(|Z_i| + a_n\Delta > \sqrt{2\alpha(\log p+1)}\right) \\ &> \left\{\frac{h^2}{1+h^2}a_n\Delta - \frac{1+2\sqrt{2\pi}p^{1/2+\epsilon}+\log p}{p^{1-\alpha}(1-p^{a-1})-2p^{1/2+\epsilon}}\right\}(1-\gamma) + \left(\Delta - \frac{1}{a_n h^2}\log p\right)\gamma, \end{aligned}$$

where we define $\gamma = P\left(|Z_i| + a_n\Delta > \sqrt{2\alpha(\log p + 1)}\right)$. Note that

$$E\left[Z_i \,\middle|\, |Z_i| < c\right] = 0.$$

By applying Lemma 3 and comparing the order of p in each term, one can obtain that for b which satisfy $b > \alpha + \epsilon - 1/2$ and $b > -\alpha$,

$$E\left[W_1\right] \geq \frac{h^2}{1+h^2} a_n \Delta \{1 + o(1)\}$$

holds. Note that when $b > -1/4$, in which all $(a, b) \in \mathcal{R}_{NPEB}$ with $b < 0$ are included, we can always pick such $\alpha, \epsilon > 0$ as $\epsilon = b + 1/4$, $\alpha = 1/4$. From the WLLN for triangular arrays,

$$\frac{\sum_{i=1}^{p_1} W_i - p_1 E[W_1]}{1 + 2(\log p)^2/a_n} \xrightarrow{\text{p}} 0. \tag{27}$$

Therefore, one can write

$$\sum_{i=1}^{p_1} W_i \geq \frac{h^2}{1+h^2} a_n p^{a+b} \{1 + o(1)\} + o_p\left\{(\log p)^2\right\}. \quad \square$$

Appendix B. Proof of Theorem 1

We define $\bar{z}_D = \bar{z}_1 - \bar{z}_2$ which has $\bar{z}_D \sim N\left(\mu_D^*, a_n^{-2} I_p\right)$ where $a_n = \left(1/n_1 + 1/n_2\right)^{-1/2}$. From (13), the estimator using EB method of mean difference μ_D^* is as follows:

$$\widehat{\mu_{D,i}}^{EB} = \bar{z}_{D,i} + \frac{1}{h^2} \frac{\sum_{j=1}^{p} (\bar{z}_{D,j} - \bar{z}_{D,i})\, \phi\left\{\frac{a_n(\bar{z}_{D,i} - \bar{z}_{D,j})}{h}\right\}}{\sum_{j=1}^{p} \phi\left\{\frac{a_n(\bar{z}_{D,i} - \bar{z}_{D,j})}{h}\right\}}. \tag{28}$$

Here, $\widehat{\mu_{D,i}}^{EB}, \bar{z}_{D,i}$ denotes the ith component of the EB estimator of the mean vector $\widehat{\mu_D}^{EB}$ and $\bar{z}_D$, respectively. $\phi(\cdot)$ is the probability distribution function of standard Gaussian distribution. According to Brown and Greenshtein (2009) and Greenshtein and Park (2009), we choose the bandwidth $h = 1/\sqrt{\log p}$ in (28).

Since $\bar{z}_D \sim N(\mu_D^*, a_n^{-2} I_p)$, we have $\bar{z}_{D,i} \sim N(\mu_i^*, a_n^{-2})$ which are mutually independent. Thus, we can rewrite $\bar{z}_{D,i}$ as follows using independent $Z_i \sim N(0,1)$ for $1 \leq i \leq p$:

$$\bar{z}_{D,i} = \begin{cases} \Delta + Z_i/a_n, & \text{if } 1 \leq i \leq p_1 \\ Z_i/a_n, & \text{if } p_1 + 1 \leq i \leq p. \end{cases}$$

For simplicity in expressions, we define the followings, aligned with lemma 4 :

$$Z_{ij}^- = Z_i - Z_j, \qquad Z_i^* = Z_i + a_n\Delta, \tag{29}$$

$$T_i = \exp\left\{-\frac{Z_i^2}{2(1+h^2)}\right\}, \quad T_i^* = \exp\left\{-\frac{Z_i^{*2}}{2(1+h^2)}\right\}. \tag{30}$$

With these notations, $\widehat{\mu_{D,i}}^{EB}$ is represented as

$$\widehat{\mu_{D,i}}^{EB} = \begin{cases} \Delta + \frac{Z_i}{a_n} + \frac{\sum_{j=1}^{p_1} \frac{Z_{ji}^-}{a_n} \phi(Z_{ij}^-/h) + \sum_{j=p_1+1}^{p} \frac{Z_{ji}^- - a_n\Delta}{a_n} \phi\left\{(Z_{ij}^- + a_n\Delta)/h\right\}}{h^2\left[\sum_{j=1}^{p_1} \phi(Z_{ij}^-/h) + \sum_{j=p_1+1}^{p} \phi\left\{(Z_{ij}^- + a_n\Delta)/h\right\}\right]}, & \text{if } 1 \leq i \leq p_1 \\ \frac{Z_i}{a_n} + \frac{\sum_{j=1}^{p_1} \frac{Z_{ji}^- + a_n\Delta}{a_n} \phi\left\{(Z_{ij}^- - a_n\Delta)/h\right\} + \sum_{j=p_1+1}^{p} \frac{Z_{ji}^-}{a_n} \phi\left(Z_{ij}^-/h\right)}{h^2\left[\sum_{j=1}^{p_1} \phi\left\{(Z_{ij}^- - a_n\Delta)/h\right\} + \sum_{j=p_1+1}^{p} \phi\left(Z_{ij}^-/h\right)\right]}, & \text{if } p_1 + 1 \leq i \leq p \end{cases} \tag{31}$$

We now show that, for $(a, b) \in A \cup C \cup D$, we have the divergence of V to ∞ in probability as in (18).

For this, we consider two cases depending on the sign of b: (i) $b < 0$ for $(a, b) \in A \cup D$ and (ii) $b > 0$ for $(a, b) \in C$. First, we show that among the region that satisfies $b < 0$, the region $A \cup D$ is a subset of $\mathcal{R}_{NPEB}$.

Claim I: $A \cup D \subset \mathcal{R}_{NPEB}$

In this case, we provide the proof of $\mathcal{R}_{NPEB} \supset A \cup D$, in which (a, b) satisfies $a + 2b > 0$ and $-1/4 < b < 0$. For the proof of (18), we have the following roadmap consisting of two steps :

Step 1 : We present a lower bound of the numerator in V such as

$$\widehat{\mu_D^*}^T \mu_D^* \geq \frac{h^2}{1+h^2} a_n p^{a+2b} \left\{1 + o_p(1)\right\}.$$

Step 2 : We then propose an upper bound of the denominator in V as

$$\|\widehat{\mu_D^*}\|_2 = O_p\left\{p^{\max(2a+2b-1,\epsilon')}\right\},$$

where ϵ' is an arbitrary positive constant.

We present the proofs of (i) Step 1 and (ii) Step 2 as follows.

• *Proof of Step 1:* From the expression of $\widehat{\mu_{D,i}}^{EB}$ in (31), one can observe that

$$\widehat{\mu_{D,i}}^{EB} \equiv \Delta + \frac{Z_i}{a_n} + \frac{1}{h^2}\left(\sum_{j=1}^{p_1} w_j \frac{Z_{ji}^-}{a_n} + \sum_{j=p_1+1}^{p} w_j \frac{Z_{ji}^- - a_n\Delta}{a_n}\right) \tag{32}$$

for $1 \le i \le p_1$, where $\sum_{j=1}^{p} w_j = 1$ with

$$w_j = \phi(Z_{ij}^-/h) \Big/ \left[\sum_{j=1}^{p_1}\phi(Z_{ij}^-/h) + \sum_{j=p_1+1}^{p}\phi\left\{(Z_{ij}^- + a_n\Delta)/h\right\}\right] \tag{33}$$

for $1 \le j \le p_1$ and

$$w_j = \phi\left\{(Z_{ij}^- + a_n\Delta)/h\right\} \Big/ \left[\sum_{j=1}^{p_1}\phi(Z_{ij}^-/h) + \sum_{j=p_1+1}^{p}\phi\left\{(Z_{ij}^- + a_n\Delta)/h\right\}\right] \tag{34}$$

for $p_1 + 1 \le j \le p$. From (32), one can check that

$$\widehat{\mu_{D,i}}^{EB} \ge \Delta + \frac{Z_i}{a_n} + \frac{1}{a_n h^2}\min\left\{(Z_{ji}^-)_{1\le j\le p_1}, (Z_{ji}^- + a_n\Delta)_{p_1+1\le j\le p}\right\} \tag{35}$$

always holds for $1 \le i \le p_1$ since the weighted mean is always greater than the minimum value.

Subsequently, we define A_i, B_i, C_i, D_i as follows:

$$\begin{aligned}
A_i &= \sum_{j=1}^{p_1}\frac{Z_{ji}^-}{a_n}\phi(Z_{ij}^-/h) \equiv \sum_{j=1}^{p_1} w_{1j}\frac{Z_{ji}^-}{a_n},\\
B_i &= \sum_{j=p_1+1}^{p}\phi\left\{(Z_{ij}^- + a_n\Delta)/h\right\}\frac{Z_{ji}^- - a_n\Delta}{a_n} \equiv \sum_{j=1}^{p_1} w_{2j}\frac{Z_{ji}^- - a_n\Delta}{a_n},\\
C_i &= \sum_{j=1}^{p_1}\phi(Z_{ij}^-/h) \equiv \sum_{j=1}^{p_1} w_{1j},\\
D_i &= \sum_{j=p_1+1}^{p}\phi\left\{(Z_{ij}^- + a_n\Delta)/h\right\} \equiv \sum_{j=p_1+1}^{p} w_{2j}.
\end{aligned}$$

For Z_i and Z_j $(i \ne j)$, one can check

$$\begin{aligned}
E\left[\phi(Z_{ji}^-/h)\,\middle|\,Z_i\right] &= \frac{h}{\sqrt{2\pi}\sqrt{1+h^2}}T_i,\\
E\left[Z_{ij}^-\phi\left(Z_{ji}^-/h\right)\,\middle|\,Z_i\right] &= -\frac{h^3}{\sqrt{2\pi}\sqrt{1+h^2}^3}Z_iT_i
\end{aligned}$$

by direct calculation. A similar relation holds for Z_i^* and T_i^* as follows.

$$\begin{aligned}
E\left[\phi\left\{(Z_{ji}^- - a_n\Delta)/h\right\}\,\middle|\,Z_i^*\right] &= \frac{h}{\sqrt{2\pi}\sqrt{1+h^2}}T_i^*,\\
E\left[(Z_{ij}^- + a_n\Delta)\phi\left\{(Z_{ji}^- - a_n\Delta)/h\right\}\,\middle|\,Z_i^*\right] &= -\frac{h^3}{\sqrt{2\pi}\sqrt{1+h^2}^3}Z_i^*T_i^*.
\end{aligned}$$

In addition, using the fact that $\phi(x) \in (0, 1/\sqrt{2\pi}]$, we obtain the following result : for $1 \le i \le p_1$, Z_{ij}^- are conditionally independent on Z_i, for all $j \ne i$. Therefore, for every $\epsilon > 0$, by Hoeffding's inequality, we have

$$P\left(\left|C_i - \frac{1}{\sqrt{2\pi}} - \frac{(p_1-1)h}{\sqrt{2\pi}\sqrt{1+h^2}}T_i\right| > p_1^{1/2+\epsilon}\,\middle|\,Z_i\right) \le 2\exp(-4\pi(p_1-1)^{2\epsilon}) \tag{36}$$

since

$$E\left[C_i \mid Z_i\right] = E\left[\sum_{j=1}^{p_1}\phi\left(Z_{ij}^-/h\right)\,\middle|\,Z_i\right] = \frac{1}{\sqrt{2\pi}} + \frac{(p_1-1)h}{\sqrt{2\pi}\sqrt{1+h^2}}T_i.$$

By taking the expectation of Z_i in (36), we have

$$P\left(\left|C_i - \frac{1}{\sqrt{2\pi}} - \frac{(p_1-1)h}{\sqrt{2\pi}\sqrt{1+h^2}}T_i\right| > p_1^{1/2+\epsilon}\right) \le 2\exp\left\{-4\pi(p_1-1)^{2\epsilon}\right\}. \tag{37}$$

Similarly, we also have

$$P\left(\left|D_i - \frac{(p-p_1)h}{\sqrt{2\pi}\sqrt{1+h^2}}T_i^*\right| > (p-p_1)^{1/2+\epsilon}\right) \le 2\exp\left\{-4\pi(p-p_1)^{2\epsilon}\right\}. \tag{38}$$

Also, from $|x\phi(x)| \le 1/\sqrt{2\pi e}$, we obtain

$$P\left(\left|A_i + \frac{(p_1-1)h^3 Z_i}{\sqrt{2\pi}\sqrt{1+h^2}^3}T_i\right| > p_1^{1/2+\epsilon}\right) \le 2\exp\left\{-\pi e(p_1-1)^{2\epsilon}\log p\right\},$$
$$P\left(\left|B_i + \frac{(p-p_1)h^3 Z_i^*}{\sqrt{2\pi}\sqrt{1+h^2}^3}T_i^*\right| > (p-p_1)^{1/2+\epsilon}\right) \le 2\exp\left\{-\pi e(p-p_1)^{2\epsilon}\log p\right\}.$$

Now, we define the events $E_{1i}, \ldots, E_{5i}$ as follows:

$$E_{1i} = \left\{\left|A_i + \frac{(p_1-1)h^3 Z_i}{\sqrt{2\pi}\sqrt{1+h^2}^3}T_i\right| < p_1^{1/2+\epsilon}\right\},$$
$$E_{2i} = \left\{\left|B_i + \frac{(p-p_1)h^3 Z_i^*}{\sqrt{2\pi}\sqrt{1+h^2}^3}T_i^*\right| < (p-p_1)^{1/2+\epsilon}\right\},$$
$$E_{3i} = \left\{\left|C_i - \frac{1}{\sqrt{2\pi}} - \frac{(p_1-1)h}{\sqrt{2\pi}\sqrt{1+h^2}}T_i\right| < p_1^{1/2+\epsilon}\right\},$$
$$E_{4i} = \left\{\left|D_i - \frac{(p-p_1)h}{\sqrt{2\pi}\sqrt{1+h^2}}T_i^*\right| < (p-p_1)^{1/2+\epsilon}\right\},$$
$$E_{5i} = \left\{\max_{1\le j\le p}|Z_j| + a_n\Delta < \log p\right\}.$$

Let $E_i = \cap_{j=1}^5 E_{ji}$. Applying Bonferroni's inequality on E_i^c and from Lemma 2, we obtain

$$\begin{aligned} P(E_i) &= 1 - P(E_i^c) = 1 - P(\cup_{j=1}^5 E_{ji}^c) \\ &\ge 1 - \sum_{j=1}^5 P(E_{ji}^c) \\ &= 1 - 8\exp\left\{-\pi e(p_1-1)^{2\epsilon}\right\} - 1/p. \end{aligned}$$

Therefore, when we define $E = \cap_{i=1}^{p_1} E_i$, we have

$$\begin{aligned} P(E) &= 1 - P(E^c) = 1 - P(\cup_{i=1}^{p_1} E_i^c) \\ &\ge 1 - \sum_{i=1}^{p_1}\left\{1 - P(E_i)\right\} \\ &\ge 1 - 8p_1\exp\left\{-\pi e(p_1-1)^{2\epsilon}\right\} - 1/p^{1-a} \end{aligned} \tag{39}$$

Note that for all $0 < a < 1$, $\lim_{p\to\infty} P(E) = 1$. On the event E, we can set the bounds for A_i, B_i, C_i, D_i to eventually build a lower bound of $\widehat{\mu_{D,i}}^{EB}$. Since $\lim_{p\to\infty} P(E) = 1$, our aim is to define an appropriate lower bound, W_i, on E which satisfies

$$V = \frac{\widehat{\mu_D^*}^T \mu_D^*}{\|\widehat{\mu_D^*}\|_2} = \frac{\Delta\sum_{i=1}^{p_1}\widehat{\mu_{D,i}}^{EB}}{\|\widehat{\mu_D^*}\|_2} \ge \frac{\Delta\sum_{i=1}^{p_1} W_i}{\|\widehat{\mu_D^*}\|_2} \xrightarrow{\text{p}} \infty. \tag{40}$$

for $(a,b) \in \mathcal{R}_{NPEB}$.

We directly apply the result in Lemma 4 and obtain $\Delta\sum_{i=1}^{p_1}\widehat{\mu_{D,i}}^{EB} \ge \Delta\sum_{i=1}^{p_1} W_i$ where W_i is defined in Lemma 4 which leads to

$$\widehat{\mu_D^*}^T \mu_D^* = \Delta\sum_{i=1}^{p_1}\widehat{\mu_{D,i}}^{EB} \ge \Delta\sum_{i=1}^{p_1} W_i \ge \frac{h^2}{1+h^2}a_n p^{a+2b}\left\{1 + o_p(1)\right\}. \tag{41}$$

This inequality holds on E, and E has the probability larger than

$$1 - 8p_1 \exp\left\{-\pi e(p_1-1)^{2\epsilon}\right\} - p^{-1+a},$$

which goes to 1 for all $0 < a < 1$ as $p \to \infty$.

• *Proof of Step 2:* The denominator in (18) can be bounded using the equation (45) in Brown and Greenshtein (2009). For $\nu = 1 + h^2$, let

$$\begin{aligned}\delta_i &:= \frac{\sum_{j=1}^{p_1} \Delta\phi\left\{(a_n\bar{z}_{D,i} - a_n\Delta)/\sqrt{\nu}\right\}}{\sum_{j=1}^{p_1} \phi\left\{(a_n\bar{z}_{D,i} - a_n\Delta)/\sqrt{\nu}\right\} + \sum_{j=p_1+1}^{p} \phi\left(a_n\bar{z}_{D,i}/\sqrt{\nu}\right)} \\ &= \begin{cases} \frac{p_1\Delta\phi(Z_i/\sqrt{\nu})}{p_1\phi(Z_i/\sqrt{\nu})+(p-p_1)\phi\left\{(a_n\Delta+Z_i)/\sqrt{\nu}\right\}}, & \text{if } 1 \le i \le p_1 \\ \frac{p_1\Delta\phi\left\{(Z_i-a_n\Delta)/\sqrt{\nu}\right\}}{p_1\phi\left\{(Z_i-a_n\Delta)/\sqrt{\nu}\right\}+(p-p_1)\phi\left(Z_i/\sqrt{\nu}\right)}, & \text{if } p_1+1 \le i \le p. \end{cases}\end{aligned}$$

Then the equation (45) in Brown and Greenshtein (2009) implies

$$E\left[\sum_{i=1}^{p}(\delta_i - \widehat{\mu_{D,i}}^{EB})^2\right] = o(p^{\epsilon}) \tag{42}$$

for all $\epsilon > 0$. Using this, one obtain the probabilistic upper bound of $\|\widehat{\mu}_D^{EB}\|_2^2$ from

$$\|\widehat{\mu}_D^{EB}\|_2^2 \le 2\sum_{i=1}^{p}(\delta_i - \widehat{\mu_{D,i}}^{EB})^2 + 2\sum_{i=1}^{p}\delta_i^2. \tag{43}$$

Since Z_i are i.i.d., $E\left[\sum_{i=1}^{p}\delta_i^2\right]$ is simplified as follows :

$$\begin{aligned}E\left[\sum_{i=1}^{p}\delta_i^2\right] &= p_1 E\left[\frac{p_1\Delta\phi(Z_i/\sqrt{\nu})}{p_1\phi(Z_i/\sqrt{\nu}) + (p-p_1)\phi\left\{(a_n\Delta+Z_i)/\sqrt{\nu}\right\}}\right]^2 \\ &\quad + (p-p_1)E\left[\frac{p_1\Delta\phi\left\{(Z_i-a_n\Delta)/\sqrt{\nu}\right\}}{p_1\phi\left\{(Z_i-a_n\Delta)/\sqrt{\nu}\right\} + (p-p_1)\phi\left(Z_i/\sqrt{\nu}\right)}\right]^2 \\ &= p_1 E\left[\frac{p_1\Delta}{p_1 + (p-p_1)\exp\left\{-a_n\Delta(a_n\Delta+2Z_i)/(2\nu)\right\}}\right]^2 \\ &\quad + (p-p_1)E\left[\frac{p_1\Delta\exp\left\{-a_n\Delta(a_n\Delta-2Z_i)/(2\nu)\right\}}{p_1\exp\left[-a_n\Delta(a_n\Delta-2Z_i)/(2\nu)\right\} + (p-p_1)}\right]^2 \\ &=: J_1 + J_2.\end{aligned}$$

From $e^{-x} \ge 1 - x$ for all x, under the event E, J_1 and J_2 satisfy

$$J_1 \le p_1\left\{\frac{p_1\Delta}{p-(p-p_1)a_n\Delta(a_n\Delta+2\log p)/2}\right\}^2,\ J_2 \le (p-p_1)\left\{\frac{p_1\Delta}{p_1+(p-p_1)}\right\}^2,$$

which implies $E\left[\sum_{i=1}^{p}\delta_i^2\right] \le p^{2a+2b-1}\{1+o(1)\}$. As a result, from (42) and (43), we have

$$\|\widehat{\mu_D^{EB}}\|_2^2 = O_p\left\{p^{\max(2a+2b-1,\epsilon')}\right\} \tag{44}$$

for all $\epsilon' > 0$.

Therefore, we prove the results in *Step 1* and *Step 2* leading to

$$\begin{aligned}V &= \frac{\widehat{\mu_D^*}^T\mu_D^*}{\|\widehat{\mu_D^*}\|_2} \ge \frac{h^2}{1+h^2}\frac{a_n p^{a+2b}(1+o_p(1))}{\sqrt{p^{\max(2a+2b-1,\epsilon')}O_p(1)}} \\ &= \frac{a_n}{\log p}p^{a+2b-\max(a+b-1/2,\epsilon'/2)}\frac{1+o_p(1)}{\sqrt{O_p(1)}} \xrightarrow{\text{p}} \infty\end{aligned}$$

for $-1/4 < b < 0$ and $a + 2b > \epsilon'/2$ since $a + 2b - \max(a + b - 1/2, \epsilon'/2) > 0$ for the given region of (a, b).

Subsequently, we show that the region C is a subset of $\mathcal{R}_{NPEB}$ among the region that satisfies $b > 0$.

Claim II: $C \subset \mathcal{R}_{NPEB}$

Here, we show (18) is satisfied for all $(a, b) \in C$ under the NPEB method. Since $\widehat{\mu_D^*}^T \mu_D^* = \Delta \sum_{i=1}^{p_1} \widehat{\mu_{D,i}}^{EB}$ and $\widehat{\mu_{D,i}}^{EB} \geq V_{2i}$ always holds,

$$\widehat{\mu_D^*}^T \mu_D^* \geq \Delta \left(\sum_{i=1}^{p_1} \left[\Delta + \frac{1}{a_n} Z_i + \frac{1}{a_n h^2} \min \left\{ (Z_{ji}^-)_{1 \leq j \leq p_1}, (Z_{ji}^- + a_n \Delta)_{p_1+1 \leq j \leq p} \right\} \right] \right) \tag{45}$$

$$\geq \Delta \left[\sum_{i=1}^{p_1} \left\{ \Delta + \left(\frac{1}{a_n} - \frac{1}{a_n h^2} \right) Z_i - \frac{1}{a_n h^2} \min_{1 \leq j \leq p} Z_j \right\} \right]. \tag{46}$$

Therefore, from Lemma 2 and the WLLN, one can obtain $\widehat{\mu_D^*}^T \mu_D^* \geq p^{a+2b} \left\{ 1 + o_p(1) \right\}$.
In addition, by using $E\left[\exp(cZ_i)\right] = \exp(c^2/2)$ for any constant $c > 0$, we have

$$\begin{aligned} E \sum_{i=1}^{p} \delta_i^2 &\leq p_1 \Delta^2 + (p - p_1) E \left[\frac{p_1 \Delta \exp\left\{ -a_n \Delta (a_n \Delta - 2Z_i) / (2v) \right\}}{p - p_1} \right]^2 \\ &\leq p^{a+2b} + \frac{p^{2a+2b}}{p - p_1} = p^{a+2b} + p^{2a+2b-1}(1 + o(1)) \\ &= p^{a+2b}(1 + o(1)) \end{aligned}$$

where the last equality is from $a < 1$. Therefore, from (42) and (43), we obtain

$$\|\widehat{\mu_D^*}\|_2^2 = O_p(p^{a+2b}). \tag{47}$$

leading to

$$V = \frac{\widehat{\mu_D^*}^T \mu_D^*}{\|\widehat{\mu_D^*}\|_2} \geq \frac{p^{a+2b} \left\{ 1 + o_p(1) \right\}}{O_p \left\{ p^{(a+2b)/2} \right\}} \xrightarrow{\mathrm{p}} \infty$$

for all $a, b > 0$.

Therefore, our main Theorem 1 is proved by Claims I and II. □

References

Anderson, T., 2003. An Introduction to Multivariate Statistical Analysis. In: Wiley Series in Probability and Statistics, Wiley.
Bickel, P.J., Levina, E., 2004. Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 10 (6), 989–1010.
Bodnar, T., Okhrin, O., Parolya, N., 2019. Optimal shrinkage estimator for high-dimensional mean vector. J. Multivariate Anal. 170, 63–79.
Brown, L.D., Greenshtein, E., 2009. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. Ann. Statist. 1685–1704.
Cai, T., Liu, W., Luo, X., 2011. A constrained L_1 minimization approach to sparse precision matrix estimation. J. Amer. Statist. Assoc. 106 (494), 594–607.
Chételat, D., Wells, M.T., 2012. Improved multivariate normal mean estimation with unknown covariance when p is greater than n. Ann. Statist. 40 (6), 3137–3160.
Dicker, L.H., Zhao, S.D., 2016. High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. Biometrika 103 (1), 21–34.
Efron, B., 2009. Empirical Bayes estimates for large-scale prediction problems. J. Am. Stat. Assoc. 29 (487), 1015–1028.
Efron, B., 2014. Two modeling strategies for empirical Bayes estimation. Statist. Sci. Rev J. Inst. Math. Stat. 29 (2), 285.
Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors–an empirical Bayes approach. J. Amer. Statist. Assoc. 117–130.
Fan, J., Fan, Y., 2008. High dimensional classification using features annealed independence rules. Ann. Stat. 36 (6), 2605.
Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.
Greenshtein, E., Park, J., 2009. Application of non parametric empirical Bayes estimation to high dimensional classification.. J. Mach. Learn. Res. 10 (7).
Hyodo, M., Yamada, T., Himeno, T., Seo, T., 2012. A modified linear discriminant analysis for high-dimensional data. Hiroshima Math. J. 42 (2), 209–231.
Ignatiadis, N., Wager, S., 2022. Confidence intervals for nonparametric empirical Bayes analysis. J. Amer. Statist. Assoc. 1–18.
James, W., 1961. Estimation with quadratic loss. In: Proc. 4th Berkeley Symp. on Math. Statist. and Prob., 1961.
Jiang, W., Zhang, C.-H., 2009. General maximum likelihood empirical Bayes estimation of normal means. Ann. Statist. 37 (4), 1647–1684.
Koenker, R., Mizera, I., 2014. Convex optimization in r. J. Stat. Softw. 60, 1–23.
Lam, C., 2016. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. Ann. Statist. 44 (3), 928–953.
Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88 (2), 365–411.
Ledoit, O., Wolf, M., 2012. Nonlinear shrinkage estimation of large-dimensional covariance matrices. Ann. Statist. 40 (2), 1024–1060.
Ledoit, O., Wolf, M., 2020. Analytical nonlinear shrinkage of large-dimensional covariance matrices. Ann. Statist. 48 (5), 3043–3065.
Ledoit, O., Wolf, M., 2022. Quadratic shrinkage for large covariance matrices. Bernoulli 28 (3), 1519–1547.
Li, M., Wang, C., Yin, Y., Zheng, S., 2025. High-dimensional scale invariant discriminant analysis. Statist. Sinica in press.
Mariathasan, S., Turley, S.J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., Kadel III, E.E., Koeppen, H., Astarita, J.L., Cubas, R., et al., 2018. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of t cells. Nature 554 (7693), 544–548.
Park, H., Baek, S., Park, J., 2022. High-dimensional linear discriminant analysis using nonparametric methods. J. Multivariate Anal. 188, 104836.
Saha, S., Guntuboyina, A., 2020. On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. Ann. Statist. 48 (2), 738–762.
Sifaou, H., Kammoun, A., Alouini, M.-S., 2020. High-dimensional linear discriminant analysis classifier for spiked covariance model. J. Mach. Learn. Res. 21 (1), 4508–4531.
Wang, C., Jiang, B., 2018. On the dimension effect of regularized linear discriminant analysis. Electron. J. Stat. 12, 2709–2742.
Wang, C., Tong, T., Cao, L., Miao, B., 2014. Non-parametric shrinkage mean estimation for quadratic loss functions with unknown covariance matrices. J. Multivariate Anal. 125, 222–232.
Xu, P., Brock, G.N., Parrish, R.S., 2009. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. Comput. Statist. Data Anal. 53 (5), 1674–1687.

서범석 교수

논문

- Seo, Beomseok* (2025). Econometric Forecasting Using Ubiquitous News Text: Text-enhanced Factor Model. International Journal of Forecasting, 41(3), 1055-1072.

- Yoo, Jaewon, Seo, Beomseok* (2025). Deposit Competition and Depository Institution Risk: Do Interest Margins Matter? Emerging Markets Finance and Trade. 1-21.

- Yoo, Jaewon, Seo, Beomseok*, & An, Junyoung (2025). Who Pays When Zombie Firms Persist? Asymmetric Debt Conditions by Credit Rating. . Finance Research Letters, 108616.

- Chang, Yu-Ling, Seo, Beomseok* (2026). Using Explainable Artificial Intelligence to Envision the Future of Work and the Workforce – From Prediction to Explanation. Technology in Society. 84.

International Journal of Forecasting 41 (2025) 1055–1072

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Econometric forecasting using ubiquitous news text: Text-enhanced factor model

Beomseok Seo

Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 04310, Republic of Korea

ARTICLE INFO

Keywords:
Dynamic factor model
Text mining
Machine learning
Economic forecasting
Nowcasting

ABSTRACT

News text is gaining increasing attention as a novel source for econometric forecasting. This paper revisits how narrative information is incorporated into econometric forecasting by effectively quantifying sector-specific textual information without requiring training data. We propose *Theme Frequency Indices* (TFIs), which utilize domain-specific subject-predicate patterns to measure public perception about the economy. TFIs for 15 sectors, including production, inflation, employment, capital investment, stock and house prices, and others, were examined and integrated into the *Text-enhanced Factor Model* (TFM), using latent factor structures. Empirical analysis based on over 18 million news articles from Korea reveals that TFM improves the accuracy of near-term GDP forecasts, demonstrating that simple text-mining techniques combined with domain knowledge can effectively leverage qualitative information in the news without costly training. The proposed method is applicable to a wide range of subjects for utilizing narrative information on the economy, offering a rapid and cost-effective approach.

© 2024 International Institute of Forecasters. Published by Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Over the last decade, nowcasting has gained popularity among central banks and economists due to the increasing accessibility of big data.[1] Many recent works point out that text data is crucial for improving the prediction accuracy of nowcasting models (Babii, Ghysels, & Striaukas, 2022; Barbaglia, Consoli, & Manzan, 2022; Kalamara, Turrell, Redl, Kapetanios, & Kapadia, 2022; Thorsrud, 2020). Various approaches have been studied to incorporate narrative information from text into quantitative research. However, most of these works rely on topic modeling, i.e., text clustering (Babii et al., 2022; Ferreira et al., 2021; Thorsrud, 2020; Zheng, Fan, Jin, & Fang, 2024), using unsupervised methods like latent Dirichlet allocation (LDA) or composite textual indices (Kalamara et al., 2022), such as the news sentiment index (NSI) (Seo, Lee, & Cho, 2024; Shapiro, Sudhof, & Wilson, 2022) and economic policy uncertainty (EPU) (Baker, Bloom, & Davis, 2016; Lee, Cho, & Jo, 2020). Despite their popularity, however, unsupervised approaches have inherent drawbacks. One major challenge is that it can be difficult to verify the results, as each cluster may not necessarily connect to a specific economic topic.[2] On the other hand, supervised methods using machine learning (Seo et al., 2024) can be expensive for developing even a single index due to the need for labeled training data.

To address these challenges, we utilize a text-mining approach to extract quantified information from news text and incorporate it into an econometric forecasting model. We generate text-based economic indicators, called *Theme Frequency Indices* (TFI), for 15 sectors that are important for economic forecasts, including macroeconomic variables such as production, inflation, employment, capital investment, and stock and house prices, as

E-mail address: bsseo@sookmyung.ac.kr.

[1] The numerical results presented in this manuscript were reproduced by the Editor-in-Chief on the 1st of December 2024.

[2] Additionally, in unsupervised methods, information that appears with low frequency in the text is often ignored, as small clusters tend to merge with larger clusters, resulting in clustering outcomes being vulnerable to the addition or removal of a few data points (Li, Seo, & Lin, 2019).

https://doi.org/10.1016/j.ijforecast.2024.11.001
0169-2070/© 2024 International Institute of Forecasters. Published by Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

well as industry-related microeconomic variables such as semiconductors and wholesale & retail. We then propose a forecasting model, *Text-enhanced Factor Model* (TFM), that employs the textual information to forecast economic variables.

Specifically, we utilize a lexical approach to extract quantified information from news text based on predefined domain-specific *subject-predicate patterns* of the 15 sectors. This approach does not require any training data for computation and can be applied in a timely manner. The method generates essential sector indices for the economy, making it practical and verifiable by comparing the results to official statistics. Our empirical analysis has shown a strong correlation between the text indices and the corresponding official statistics for up to 10 months, with the text indices leading the official statistics. Additionally, the proposed model, which incorporates textual information, has been demonstrated to improve the generalized testing accuracy for near-term forecasting compared to the method that uses official statistics alone. The model leverages qualitative information through texts and takes advantage of the latest information ubiquitous in the public when official statistics are unavailable.

Textual information is particularly valuable for economic forecasting as it addresses the challenge of limited data availability for statistical models. Economic judgment heavily relies on expert opinions, which qualitatively evaluate various information (Mæhle, HlÃ, Pranovich, Selander, et al., 2021). The limitations of statistical models are primarily due to the difficulty of immediately acquiring sufficient numeric information. Firstly, outside of the financial sector, few high-frequency economic indicators are available daily or weekly. Additionally, official statistics used as quantitative indicators often have a time lag in publication, making them difficult to use in rapidly changing economic conditions. Moreover, integrating qualitative information into quantitative models poses further difficulties, as variable importance varies depending on the economic situation and cannot be easily quantified.

Textual information, however, can address the abovementioned limitations, as vast amounts of qualitative information are conveyed swiftly through texts. News text, in particular, delivers a wide variety and volume of information quickly. Therefore, it is expected that faster economic judgment is possible when appropriately incorporating news information into forecasting models (Babii et al., 2022; Bybee, Kelly, Manela, & Xiu, 2021; Thorsrud, 2020). News articles present factual information, quote statistics, and convey expert opinions and insights through interviews and journalists' interpretations of the economy. Additionally, news stories that capture public interest are more likely to be widely distributed, reflecting the evolving interests of the public over time. This contrasts with official statistics, which rely on fixed metrics.

Thus, text mining-generated indices supplement existing information and enrich economic monitoring, particularly when there are delays in releasing official statistics. Moreover, these indices can improve econometric forecasting by integrating future opinions and outlooks found within news texts.

The information gleaned from news texts can be likened to that obtained through surveys as outlined in Fig. 1. The news texts have a two-fold effect on the public's perception of the economy: news journalists capture the public's perception of the economy, while their reporting also influences it. Surveys, an experimental study, are a reliable method for gauging public opinion and are managed under stringent quality control. However, they are costly, time-consuming, and not frequently available. In contrast, text mining, an observational study, can extract similar information as surveys with significantly lower costs and time investment at any time, given that news texts are associated with the public's perception of the economy. Furthermore, news texts include more information than surveys because they convey the degree of public attention to a given topic, with coverage increasing as the topic gains more attention.

Therefore, our proposed method has the potential to contribute to several areas of future research. Firstly, the proposed text-mining technique is applicable to a wide range of subjects to measure the narrative information rapidly and cost-effectively. TFIs can complement existing survey indices and serve as supplementary indices based on observational studies. Secondly, the proposed framework for forecasting near-term gross domestic product (GDP) provides an approach to developing a real-time statistical model that leverages timely news information for econometric predictions. This framework can effectively combine narrative information with existing sources as it is automated through a web-scraping program and text-mining algorithm, enabling timely predictions.

The remainder of this paper is organized as follows: Section 2 provides an overview of relevant literature. Section 3 presents the proposed method in detail, covering data collection, the text-mining method used to create text indicators, and the text-enhanced factor model. Section 4 analyzes the empirical results of the proposed indices and the model incorporating text information. Finally, Section 5 summarizes the implications of this paper and provides avenues for future research.

2. Related works

Economic analysis using text has been actively conducted since the mid-2010s, utilizing sources such as news texts (Baker et al., 2016; Lee, Kim, & Park, 2019; Seo et al., 2024; Shapiro et al., 2022) that reflect press opinions, as well as social media and search information (Sun, Lachanski, & Fabozzi, 2016) that convey public views, and corporate accounting and evaluation reports (Lewis & Young, 2019; Seo, 2023) that address industrial information. Most studies analyze texts using sentiment analysis and topic modeling based on a lexical approach or statistical models.

Early works using the lexical approach employed topic keywords to classify each article rather than analyzing sentences. Baker et al. (2016) used keywords for three categories—economy, policy, and uncertainty—to count the articles that include at least one word from all categories. This method is not appropriate for measuring sentiments for a specific topic because it does not consider the predicates associated with topic words.

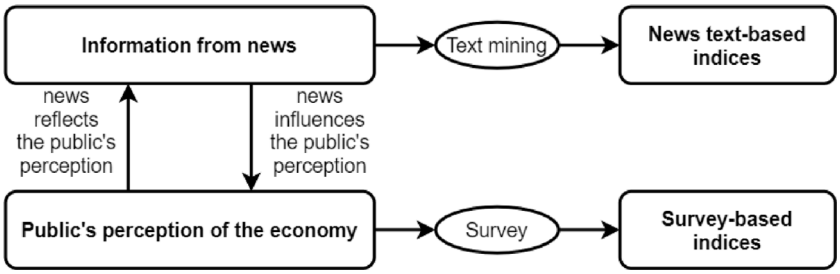


Fig. 1. The relationship between news and the public's perception of the economy.
Notes: News text-based indices can complement survey-based indices with significantly lower costs and time investment, given that news texts are associated with the public's perception of the economy.

Gardner, Scotti, and Vega (2022) suggested using topic words and modifiers to determine a sentence's sentiment. They used the topic-modifier pairs to evaluate FOMC statements regarding five topics: labor, output, inflation, financial condition, and future monetary policy. This method is similar to our approach, which utilizes subject-predicate pairs, although the scoring algorithms differ. In Gardner et al. (2022), each modifier takes on values of 1, 0, and −1 based on whether the modifier's connotation is good, neutral, or bad. Then, relative counts of each topic's sentiment are computed by dividing them by the square root of the total number of words. The scoring rule is applied to each of the 131 FOMC statements. This method is appropriate for the meticulously written FOMC statements but may not be suitable for news articles that contain multiple opinions in an article and have irregular length.

In contrast, we propose an algorithm to compute frequency indices that help estimate official statistics at the desired frequency, such as daily, weekly, and monthly. The algorithm computes patterns based on combinations of multiple grammatical categories, such as subject, tense, and verbs. The patterns are then investigated based on sentences, and frequency is computed based on articles that include these patterns for the desired time frequency. This approach allows us to efficiently compute specific sector indices for a much larger dataset of 18.3 million news articles.

In recent years, research has advanced to include applying more sophisticated statistical models, such as neural networks, for text analysis (Hájek & Olej, 2013; Seo et al., 2024). Furthermore, the advent of ChatGPT has enabled studies (Korinek, 2023) to leverage publicly accessible large language models, eliminating the necessity of training individual models.

Text, as unstructured data, can convey information on a broad range of subjects. Existing studies have analyzed text to predict market prices such as inflation and stock prices (Kalamara et al., 2022; Li, Wu, & Wang, 2020), identify new economic information such as crisis indicators and uncertainty indicators (Baker et al., 2016; Caldara & Iacoviello, 2022; Li, Liang, Li, Wang, & Wu, 2009), detect fraud (Humpherys, Moffitt, Burns, Burgoon, & Felix, 2011), classify corporate sustainability (Te Liew, Adhitya, & Srinivasan, 2014), and conduct credit scoring (Yap, Ong, & Husain, 2011).

The Bank of Korea is at the forefront of using text data for economic analysis in Korean text. Lee et al. (2020) calculated the monetary policy surprise index through news text analysis, and Seo et al. (2024) developed a news sentiment index for Korea using transformer neural networks. Besides news data, various text analyses, such as using analyst reports (Seo, 2023), are also being attempted in Korea.

Indeed, new data and models are crucial for generating unprecedented insights in economic forecasting. Alongside new data such as text, recent advancements in statistical models and machine learning techniques have played a pivotal role in overcoming various challenges, such as enabling the utilization of ultra-high-dimensional data in forecasting.

The Dynamic Factor Model (DFM) (Bańbura, Giannone, Modugno, & Reichlin, 2013; Bok, Caratelli, Giannone, Sbordone, & Tambalotti, 2018) has been gaining popularity for nowcasting, as it can handle more variables than samples by utilizing unsupervised factor analysis. Using factor analysis and the EM algorithm, DFM can predict a large number of variables simultaneously and incorporate mixed-frequency data (Mariano & Murasawa, 2010).

Real-time prediction has also been explored using a range of machine-learning models, including k-nearest neighbor regression and Support Vector Regression (SVR) for GDP (Richardson, van Florenstein Mulder, & Vehbi, 2021), and neural networks for global merchandise trade exports (Hopp, 2022).

3. Data and methods

3.1. Preparing texts

Using web-scraping techniques, we collected economic news articles posted on internet news portal sites in Korea after January 2005. Our web-scraping method automatically downloads publicly available news articles and constructs a news database every morning. The method was built using *Python* programming language to automate the entire process of collecting news articles and generating TFI textual information without human intervention. The collected data consists of approximately 4,000 articles per weekday, corresponding to about 1 million articles per year (a total of 18, 759, 775 articles between January 2005 and December 2023) from 120

media companies. Overall, the dataset encompasses about 227 million sentences and 3.8 billion words.[3]

3.2. Topic selection

It is crucial to define specific topics that need to be extracted and to select an appropriate method to extract information from text effectively. If the topic is too general, the extracted information may lose its value; conversely, if it is too specific, the information may not be present in the text.

We selected text topics across 15 sectors, covering major macroeconomic variables and industry-related micro-variables (as shown in Table 1). These topics were selected based on their predictive importance for Korea's GDP, with the understanding that news information can be collected at any time and used as a supplementary source when official statistics are unavailable. "Production" is included to capture industrial production, which is widely used as a high-frequency proxy for economic activity. "Wholesale and Retail" reflect private consumption, constituting the largest share of Korea's GDP (47.8% as of the 2020 base year). "Government Expenditure" is also included to capture government consumption, accounting for 17% of GDP. "Facilities Investment" and "House Construction" are used to estimate investments, which comprise 31.5% of GDP, including facilities investment (9.4%) and construction investment (15.2%). "World Trade" is added to reflect exports and imports, representing a significant portion of Korea's GDP at 65.5%. Key industries in Korea's manufacturing sector, such as semiconductors, automobiles, and shipbuilding, are included. To capture labor market conditions closely linked to production, unemployment, recruitment, and job search are separately considered. Lastly, the outlook for key market price indicators, such as inflation, stock prices, and house prices, is also included. These 15 selected sectors are frequently referenced in the media when evaluating Korea's economic situation.

In addition to the 15 sector text indicators, we also utilized the news sentiment index (NSI) of Korea (Seo et al., 2024) and the economic policy uncertainty (EPU) index. The NSI was obtained from the Bank of Korea's Economic Statistics System (ECOS). The monthly EPU indices were calculated using the same news database used for the 15 text indices, following the methods of Baker et al. (2016) and Lee et al. (2020).

[3] The news articles were scraped from the internet news portal NAVER using the Python package *BeautifulSoap*. The configuration of the database is as follows.

Provider	Num. of articles	Examples
Economic newspapers	10.8 m (59.0%)	Money Today, Asia Economics, EDAILY, etc.
News agencies	4.8 m (26.1%)	Yonhap News, Newsis, News1, etc.
Regular news	2.5 m (13.8%)	Segye Ilbo, Donga Ilbo, Kyunghyang Shinmun, etc.
Others	0.2 m (1.1%)	Shindonga, Jungang Sunday, etc.

Given that news texts convey the latest information and expert opinions, integrating topics related to key variables in the forecasting model could enhance performance, particularly if these topics frequently appear in news articles and effectively complement the existing variables.

3.3. Choice of information extraction method

There are two primary information extraction methods: stochastic approaches that utilize statistical models (such as machine learning) and rule-based approaches that rely on human-defined extraction criteria. The choice of method depends on the nature of the text data source and the specific topic of information being extracted.

Statistical models generally exhibit better accuracy than rule-based methods when handling unstructured texts and abstract topics. This is because individuals can express abstract concepts in countless ways within unstructured texts, such as capturing economic sentiment from informal social media posts. Conversely, for more formal texts, such as news articles addressing specific topics, rule-based methods can achieve competitive accuracy (Seo et al., 2024). For instance, describing the direction of inflation is relatively more straightforward to define in narrative patterns because news journalists frequently use formalized expressions. Inflation is often associated with verbs like rise, increase, surge, spike, escalate, and climb. These patterns are even more apparent in Korean, which heavily utilizes verbs derived from Chinese characters in formal writing.

A significant advantage of the rule-based method is that it eliminates the need to create expensive labeled datasets and minimizes preprocessing steps, leading to computationally efficient text analysis. Typically, statistical models necessitate text standardization through tokenization, normalization, removal of stopwords, and integer embedding. This standardization is essential because statistical models require inputting all text characters and transforming the text tokens into a uniform format to understand their meaning, even when words have different conjugations. In contrast, by focusing on the root form of words,[4] the rule-based model can quickly check for the presence of selected patterns without extensive preprocessing. Consequently, we analyze the news text by applying human-defined rules to identify the subject and predicate components of word patterns.

3.4. News sentence-based theme frequency index

The news sentence-based *Theme Frequency Index* (TFI) is defined by the relative frequency of a specific topic appearing in news articles. Unlike article-based approaches, TFI is computed at the sentence level. If any sentence

[4] In Korean, the root part of a predicate usually retains a fixed form as a stem and is combined with various endings for conjugation. This is especially evident in formal writing, where words derived from Chinese characters are more prevalent. Similarly, in English, this approach can be applied by utilizing the invariant parts of a word for conjugation, considering exceptions for short words like 'rise' and 'rose,' where the meaningful invariant parts cannot be separated.

Table 1
Subjective-predicate patterns used to compute 15 sector text indices.

Topic		Topic Words
Industry	Production	|(production 생산), (rise 상승, increase 늘, skyrocket 급등, grow 증가, improve 개선, accelerate 가속)| − |(production 생산), (fall 하락, decrease 줄, drop 급락, shrink 감소, worsen 악화, slowdown 둔화)|
	Shipbuilding	|(ship 선박), (order 수주), (rise 상승, increase 늘, skyrocket 급등, grow 증가, improve 개선) | − |(ship 선박), (order 수주), (fall 하락, decrease 줄, drop 급락, shrink 감소, worsen 악화) |
	Automobiles	|(car 자동차, automobile 승용차), (rise 상승, skyrocket 급등, grow 증가, improve 개선, accelerate 가속)| − |(car 자동차, automobile 승용차), (fall 하락, drop 급락, shrink 감소, worsen 악화, slowdown 둔화) |
	Semiconductors	|(semiconductor 반도체), (rise 상승, skyrocket 급등, grow 증가, improve 개선, accelerate 가속)| − |(semiconductor 반도체), (fall 하락, drop 급락, shrink 감소, worsen 악화, slowdown 둔화) |
Facilities Investment		|(facility investment 설비투자, r&d R&D), (rise 상승, skyrocket 급등, grow 증가, improve 개선, accelerate 가속, increase 늘, expand 확대)| − |(facility investment 설비투자, r&d R&D), (fall 하락, drop 급락, shrink 감소, worsen 악화, slowdown 둔화, decrease 줄, reduce 감축) |
House Construction		|(house 주택, apartment 아파트), (construct 건축, build 건설, construction 착공, construction 시공) |
Employ-ment	Unemployment	|(unemployment 실업), (rise 상승, increase 늘, grow 증가, worsen 악화)| − |(unemployment 실업), (fall 하락, decrease 줄, shrink 감소, improve 개선) |
	Recruitment	|(recruitment 채용, hiring 고용), (rise 상승, increase 늘, grow 증가, improve 개선)| − |(recruitment채용, hiring 고용), (fall 하락, decrease 줄, shrink 감소, worsen 악화) |
	Job Search	|(employment 취업, job search 구직), (rise 상승, increase 늘, grow 증가)| − |(employment 취업, job search 구직), (fall 하락, decrease 줄, shrink 감소) |
Wholesale and Retail		|(wholesale 도매, retail 소매, wholesale and retail 도소매), (rise 상승, skyrocket 급등, grow 증가, improve 개선, accelerate 가속, increase 늘)| − |(wholesale 도매, retail 소매, wholesale and retail 도소매), (fall 하락, drop 급락, shrink 감소, worsen 악화, slowdown 둔화, decrease 줄) |
Government Expenditure		|(government 정부), (support 지원, subsidy 보조, expenditure 지출) |
Inflation Outlook		|(inflation 물가), (forecast 전망, predict 예측, expect 예상), (rise 상승, skyrocket 급등, go up 올라, high 높)| − |(inflation 물가), (forecast 전망, predict 예측, expect 예상), (fall 하락, drop 급락, go down 내려, low 낮) |
Stock Price Outlook		|(kospi 코스피, kosdaq 코스닥, stock price 주가, stock 주식), (forecast 전망, predict 예측), (rise 상승, grow 증가, improve 개선, increase 늘, high 높)| − |(kospi 코스피, kosdaq 코스닥, stock price 주가, stock 주식), (forecast 전망, predict 예측), (fall 하락, shrink 감소, worsen 악화, decrease 줄, low 낮) |
House Price Outlook		|(house 주택, apartment 아파트), (price 가격, selling price 매매가, lease price 전세가, presale price 분양가), (forecast 전망, predict 예측), (rise 상승, skyrocket 급등, expand 확대, improve 개선, accelerate 가속, high 높)| − |(house 주택, apartment 아파트), (price가격, selling price 매매가, lease price 전세가, presale price 분양가), (forecast 전망, predict 예측), (fall 하락, drop 급락, reduce 축소, worsen 악화, slowdown 둔화, low 낮) |
World Trade		|(global 글로벌, world 세계), (trade 교역, trade 무역, export 수출, import 수입), (rise 상승, skyrocket 급등, grow 증가, improve 개선, accelerate 가속, increase 늘, expand 확대)| − |(global 글로벌, world 세계), (trade 교역, trade 무역, export 수출, import 수입), (fall 하락, drop 급락, shrink 감소, worsen 악화, slowdown 둔화, decrease 줄, reduce 축소)

Notes: it is recommended to use different groups of words to apply this method to English, as the English words used here are direct translations of the corresponding Korean words. The elements in $|\cdot|$, denoted by a $(\cdot)$, indicate each word group, $W^{[k]}$, for $k = 1, ..., K$, while those in $(\cdot)$ are the topic words, $w_l^{[k]}$, for $l = 1, \cdots, L_k$. The word groups before and after $-$ indicate the keyword patterns for $TFI^{(rise)}$ and $TFI^{(fall)}$, respectively.

within an article matches a given theme, the article is counted toward that theme for frequency calculation. This process is carried out as follows.

Step 1: Let $\Omega_t = \{A_1^{(t)}, \ldots, A_{N_t}^{(t)}\}$ be the set of N_t news articles at time t for articles $A_i^{(t)}$, $i = 1, \ldots, N_t$, and let each news article $A_i^{(t)} = \{S_{i1}^{(t)}, \ldots, S_{iM_i}^{(t)}\}$ be the set of sentences, $S_{im}^{(t)}$, $m = 1, \ldots, M_i$, for M_i being the length of sentence in ith article.

Step 2: Then, the word groups for a certain topic, $W^{[k]} = \{w_1^{[k]}, \ldots, w_{L_k}^{[k]}\}$ for $k = 1, \ldots, K$, and the topic words belonging to group k, $w_l^{[k]}$ for $l = 1, \ldots, L_k$, are determined based on domain knowledge.

Step 3: The TFI is computed as follows.

$$C_{im}^{(t)} = \prod_{k=1}^{K} \bigvee_{l=1}^{L_k} I_{S_{im}^{(t)}}(w_l^{[k]}), \tag{1}$$

$$\hat{A}_i^{(t)} = \bigvee_{m=1}^{M_i} C_{im}^{(t)}, \tag{2}$$

$$TFI_t = \frac{\sum_{i=1}^{N_t} \hat{A}_i^{(t)}}{N_t} \tag{3}$$

where $I_S(w) = \begin{cases} 1, & \text{if } w \in S \\ 0, & \text{o.w.} \end{cases}$ w∈ S is an indicator function, and $\bigvee_{m=1}^{M} C_m = max(C_1, \ldots, C_M)$ is an elementwise max function.

Step 4: If it is necessary to account for directional words in the computation of the TFI, the TFIs associated with 'rise' and 'fall' are computed separately, and the final TFI is then calculated as the difference between the two directional TFIs.

$$TFI_t = TFI_t^{(rise)} - TFI_t^{(fall)} \tag{4}$$

Table 1 describes the subject-predicate word groups used for each of the 15 sectors. These word groups were initially selected based on common expressions in news articles describing each sector's economic trends. The corresponding TFIs were computed using the selected word groups and compared to official statistics for each sector (as shown in Table 3 and Fig. 2) to determine whether to use directionless or directional words. The final word groups were chosen based on Pearson's correlation between the TFIs and official statistics. For "House Construction" and "Government Expenditure," directionless word groups were used, while directional word groups were applied to the other sectors. This distinction is logical, as news articles typically report the initiation of house construction and government expenditure when they occur, whereas, for other topics, articles focus on the direction of change rather than merely the occurrence.

3.5. Forecasting models utilizing textual information

3.5.1. Text-enhanced factor model

We utilize the Dynamic Factor Model (DFM) structure, which is capable of handling mixed-frequency data, and incorporate TFIs into each factor of the DFM that corresponds to macroeconomic segments such as production, the external market, investment, labor market, and others, as shown in Table 2. This model setup enables us to utilize textual information at the ragged edge, where official statistics are not yet available for each sector.

For q latent factors at time t, denoted as f_t, and a p-dimensional vector of the observed variables, X_t, the *Text-enhanced Factor Model* (TFM) is defined by the usual dynamic factor model with two linear functions.

$$X_t = \Lambda f_t + \epsilon_t \tag{5}$$

$$f_t = \psi(L) f_{t-1} + \eta_t \tag{6}$$

$$\epsilon_{i,t} = \rho_i e_{i,t-1} + e_{i,t} \tag{7}$$

$$e_{i,t} \sim N(0, \sigma_i^2) \tag{8}$$

$$\eta_t \sim N(0, \Sigma) \tag{9}$$

where L is a lag operator, $\Lambda \in \mathbb{R}^{p \times q}$ is the matrix of factor loadings, $\psi(L) \in \mathbb{R}^{q \times q}$ represents the rth order polynomial coefficients, $\epsilon_t \in \mathbb{R}^p$ is the idiosyncratic disturbance with $\rho_i \in \mathbb{R}$ and $\sigma_i^2 \in \mathbb{R}$ for $i = 1, \ldots, p$, and $\eta_t \in \mathbb{R}^q$ is the factor innovation with $\Sigma \in \mathbb{R}^{q \times q}$.

In this research, we consider two overall factors that combine all variables and 13 sector-specific factors composed of corresponding variables (see Table 2). A total of 83 variables are utilized, comprising 69 monthly and 14 quarterly variables. These include 52 official statistics, 14 financial indices, and 17 text indices, which comprise 15 TFIs, NSI, and EPU. The financial and textual indices have no lags and are readily available, while the availability of official indices varies depending on the forecasting time. To capture the characteristics of the variables, we used year-over-year and month-over-month growth rates. For variables available in both seasonally adjusted (SA) and non-seasonally adjusted (NSA) forms, we applied month-over-month transformation for SA variables and year-over-year transformation for NSA variables. For variables released without seasonal adjustment, we applied X-13 ARIMA SEATS from the Census Bureau (Sax & Eddelbuettel, 2018) and transformed them into month-over-month growth rates before using them. Consistent with previous studies, we employed an AR(4) structure for the two overall factors and an AR(1) structure for the other 13 sector-specific factors. We used an AR(1) structure for the observed variables for the idiosyncratic disturbance.

It is crucial to organize the training data properly by utilizing the available information employing vintage data to train the TFM, which contains statistics with different release cycles and lags. The missing values are imputed using the Kalman filter (Seabold & Perktold, 2010) to resolve the mixed frequency and ragged edge problems in the training of the TFM. Then, the TFM is estimated using Maximum Likelihood Estimation (MLE) with the Expectation-Maximization (EM) algorithm. This method has been implemented in the Python package, *statsmodels*(Seabold & Perktold, 2010). Please refer to Mariano and Murasawa (2010) for more details. To assess the performance of the TFM, we utilized three datasets: official statistics alone, news texts alone, and both combined. These three-factor models, each incorporating different information sets, are compared for their predictive

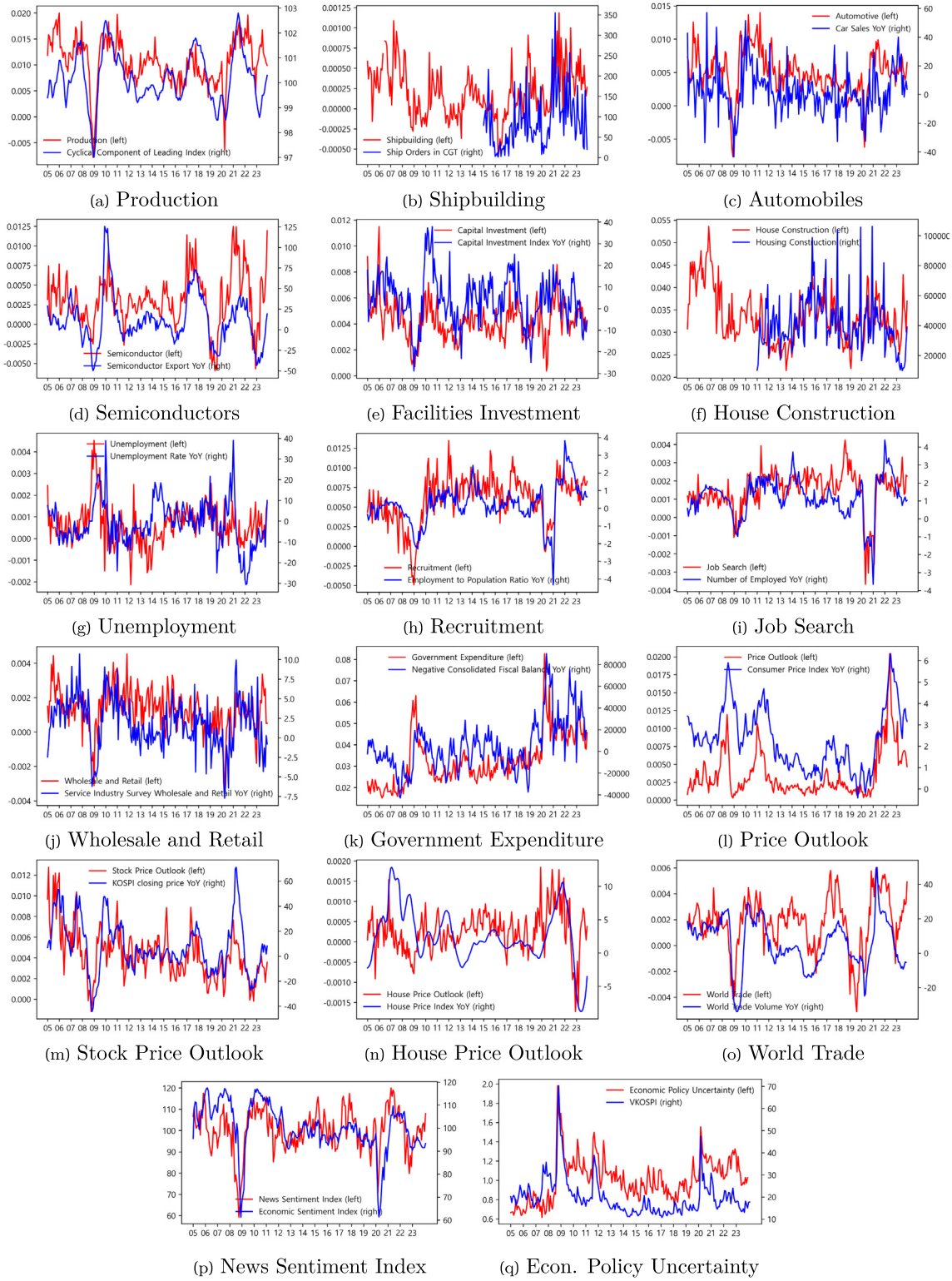


Fig. 2. Time series plots of 15 sector text indices.
Notes: Text-based economic indicators (red lines) and their corresponding official statistics (blue lines) between January 2005 and December 2023 exhibit similar trends. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
All variables and factors for Text-enhanced Factor Model.

	External	Production	Labor	Consumption	Facilities Inv.	Construction	Government	Exports	Imports	Prices	Finance	Real estate	Sentiments
GDP(SA)(Q)		V											
GDP(NSA)(Q)		V											
Private consumption(SA)(Q)				V									
Government consumption(SA)(Q)							V						
Construction(SA)(Q)						V							
Facilities investment(SA)(Q)					V								
Exports of goods and services(SA)(Q)	V							V					
Imports of goods and services(SA)(Q)	V								V				
Private consumption(NSA)(Q)				V									
Government consumption(NSA)(Q)							V						
Construction(NSA)(Q)						V							
Facilities investment(NSA)(Q)					V								
Exports of goods and services(NSA)(Q)	V							V					
Imports of goods and services(NSA)(Q)	V								V				
Unemployment rate			V										
Employment to population ratio			V										
Number of employed people			V										
Monthly goods exports	V							V					
Monthly goods imports	V								V				
Export price index	V							V		V			
Import price index	V								V	V			
Producer price index		V								V			
Consumer price index				V						V			
Price index excluding agricultural product & oil										V			
Price index excluding food & energy										V			
Consumption & Retail sales index(SA)				V									
Service industry production index(SA)		V		V									
Consumption & Retail sales index(NSA)				V									
Service industry production index(NSA)		V		V									
Manufacturing industry production index(SA)		V											
Manufacturing industry shipment index(SA)		V											
Manufacturing inventory index(SA)		V											
Manufacturing industry production index(NSA)		V											
Manufacturing industry shipment index(NSA)		V											
Manufacturing inventory index(NSA)		V											
Facility investment index(SA)					V								
Construction completed(SA)						V							
Facility investment index(NSA)					V								
Construction completed(NSA)						V							
Manufacturing business performance BSI(SA)		V											V
Manufacturing business performance BSI(NSA)		V											V
All industries performance BSI		V											V
Service industry performance BSI		V											V
All industries sales BSI		V											V
Manufacturing export BSI								V					V
Manufacturing domestic demand sales BSI		V											V
Manufacturing new orders BSI		V											V
Manufacturing operation rate BSI		V											V
Economic sentiment index													V
Current economic judgment CSI													V
Consumer sentiment index				V									V
Consolidated fiscal balance							V						
Housing sales price index(HSPI)-Seoul												V	
Housing sales price index(HSPI)-National												V	
Housing lease price index(HLPI)-Seoul												V	
Housing lease price index(HLPI)-National												V	
Call rate											V		
CD rate											V		
KTB 3-year rate											V		
KRW exchange rate	V										V		
EUR exchange rate	V										V		
KOSPI											V		
KOSDAQ											V		
Dubai crude oil											V		
WTI futures											V		
Gold futures											V		
News sentiment index(T)													V
Economic policy uncertainty(T)													V
Production(T)		V											
Shipbuilding(T)		V											
Automobiles(T)		V											
Semiconductors(T)		V											
Facilities investment(T)					V								
Construction(T)						V							
Unemployment(T)			V										
Recruitment(T)			V										
Job search(T)			V										
Wholesale & retail(T)				V									
Government expenditure(T)							V						
Price outlook(T)										V			
Stock price outlook(T)											V		
House price outlook(T)												V	
World trade(T)	V												

Notes: Variables marked with (T) indicate that they are computed from text data, while (Q) denotes quarterly variables. In total, 83 variables are used, consisting of 69 monthly and 14 quarterly variables, including 52 official statistics, 14 financial indices, and 17 text indices.

Table 3
Max leading lags and cross-correlation in the max lag between TFIs and official statistics.

Topic		Related Official Stat.	Max Leading Lag	Max Leading Corr.
Industry	Production	Cyclical component of leading index (KOSTAT)	−1	0.63
	Shipbuilding	Ship orders in CGT (Clarkson Research)	4	0.52
	Automobiles	Car sales in units (KAA)	1	0.48
	Semiconductors	Semiconductor exports (IITP)	−2	0.61
Facilities Investment		Equipment investment index (KOSTAT)	1	0.42
House Construction		Housing construction (MLIT)	0	0.41
Employment	Unemployment	Unemployment rate (KOSTAT)	1	0.31
	Recruitment	Employment to population ratio (KOSTAT)	−1	0.59
	Job Search	Number of employed (KOSTAT)	1	0.55
Wholesale and Retail		Service industry survey - W&R (KOSTAT)	1	0.40
Government Expenditure		Government fiscal balance (MOEF)	−3	0.69
Inflation Outlook		Consumer price index (KOSTAT)	−2	0.78
Stock Price Outlook		KOSPI closing price (KRX)	−4	0.67
House Price Outlook		House price index (Korea real estate board)	−6	0.43
World Trade		World merchandise trade volume (OECD)	−3	0.65
News Sentiment Index		Economic sentiment index (BOK)	−2	0.60
Econ. Policy Uncertainty		VKOSPI index (KRX)	0	0.53

Notes: The cross-correlation between TFIs and related official statistics is computed for data spanning from 2016 to 2023. Negative leading lags indicate the maximum correlation between the text index and corresponding official statistics occurs when the text index leads. In contrast, positive lags indicate that the official statistics lead.
KOSTAT: Korean Statistics, KAA: Korea Automobile Association, IITP: Institute of Information & communications Technology Planning & Evaluation, MLIT: Ministry of Land Infrastructure and Transport in Korea, MOEF: Ministry of Economy and Finance, KRX: Korea Exchange, OECD: Organization for Economic Cooperation and Development, BOK: Bank of Korea.

accuracy in Section 4.2.1. We did not investigate variable selection among the textual and official statistics data, as suggested by Giannone, Lenza, and Primiceri (2021), which implies a preference for models with many predictors over a single sparse model.

3.5.2. Convolutional recurrent neural networks (CRNNs) with TFIs

It is important to choose an appropriate forecasting model that takes into account the characteristics of text data. Text indicators, which quantify various nuances, may contain more noise than official statistics and can be challenging to interpret regarding their trends and periodicity. Although they are quantitative measures of media interest and do not exhibit obvious non-stationarity, such as upward trends (see Fig. 2), they show high volatility, making distinguishing noise from economic signals challenging. Therefore, we examine the predictive effectiveness of both linear and non-linear models to incorporate text indicators.

In addition to using the TFM, a linear model, we also constructed a *Convolutional Recurrent Neural Network* (CRNN) as a non-linear model to compare its predictive accuracy using textual information. The CRNN consists of 64 convolutional filters applied in 3-month intervals, followed by a *Long Short-Term Memory* (LSTM) layer consisting of 32 units. The convolutional layer helps estimate relationships among the multiple observed variables, while the LSTM layer incorporates time-series characteristics into the model. The CRNN model was also designed to handle mixed-frequency data by aggregating monthly and quarterly data points into quarterly intervals using convolutional filters. Unlike the TFM, which uses an unsupervised setup with latent factors and an autoregressive structure to predict all observed variables simultaneously, the CRNN was designed for a supervised setup specifically to predict GDP using all other variables.

For the p-dimensional vector of input variables, X_t, and the target variable, y, CRNN consists of the following layers:

- Convolutional Layer:

$$Z_{t\in T_q} = \sigma(W_k X_{t\in T_q} + b_k). \tag{10}$$

- Long Short-Term Memory Layer

$$f_q = \sigma(V_f Z_q + U_f h_{q-1} + e_f), \tag{11}$$
$$i_q = \sigma(V_i Z_q + U_i h_{q-1} + e_i), \tag{12}$$
$$o_q = \sigma(V_o Z_q + U_o h_{q-1} + e_o), \tag{13}$$
$$\tilde{c}_q = \tilde{\sigma}(V_c Z_q + U_c h_{q-1} + e_c), \tag{14}$$
$$c_q = f_q \circ c_{q-1} + i_q \circ \tilde{c}_q, \tag{15}$$
$$h_q = o_q \circ \sigma(c_t). \tag{16}$$

- Feed-forward Layer

$$Z'_q = \sigma(W'_j c_q + b'_j), \tag{17}$$
$$y_q = W'' Z'_q + b''. \tag{18}$$

where $\sigma()$ is the ReLU function, i.e., element-wise $max(0, x)$, $\tilde{\sigma}$ is hyperbolic tangent function, and $\circ$ is the Hadamad product, i.e., element-wise product. $W_k \in \mathbb{R}^{p\times 3}$ and $b_k \in \mathbb{R}$ are the kernel weights and bias for the convolutional layer for $k = 1, \ldots, \tilde{k}$ with the number of convolutional filters $\tilde{k} = 64$. $V \in \mathbb{R}^{h\times\tilde{k}}, U \in \mathbb{R}^{h\times h}$ and $e \in \mathbb{R}^h$ are the weights and bias for the LSTM layer, where $h = 32$ indicates the number of hidden units. $W' \in \mathbb{R}^{h\times d}$, $W'' \in \mathbb{R}^d$, $b' \in \mathbb{R}^d$, and $b'' \in \mathbb{R}$ are the weights and bias for the feed-forward (FF) layers, with the number of hidden units in the FF layer being $d = 4$. Within the

Table 4
Granger causality tests between TFIs and official statistics.

Topic		Offi. Stat.→Text		Text→Offi. Stat.		Direction
Industry	Production	4.574	(0.004)**	3.013	(0.031)*	Stat.→Text
	Shipbuilding	4.152	(0.008)**	1.040	(0.379)	Stat.→Text
	Automobiles	4.828	(0.003)**	7.135	(0.000)**	Bidirectional
	Semiconductors	2.404	(0.122)	16.065	(0.000)**	Text→Stat.
Facilities Investment		4.551	(0.004)**	1.945	(0.123)	Stat.→Text
House Construction		0.061	(0.805)	2.741	(0.100)	Not sig.
Employment	Unemployment	6.848	(0.010)*	2.849	(0.093)	Stat.→Text
	Recruitment	3.093	(0.047)*	10.142	(0.000)**	Text→Stat.
	Job Search	14.800	(0.000)**	0.854	(0.466)	Stat.→Text
Wholesale and Retail		7.047	(0.001)**	1.307	(0.273)	Stat.→Text
Government Expenditure		0.204	(0.893)	7.729	(0.000)**	Text→Stat.
Inflation Outlook		0.440	(0.508)	33.727	(0.000)**	Text→Stat.
Stock Price Outlook		0.199	(0.656)	13.437	(0.000)**	Text→Stat.
House Price Outlook		0.208	(0.649)	56.517	(0.000)**	Text→Stat.
World Trade		2.681	(0.048)*	7.296	(0.000)**	Text→Stat.
News Sentiment Index		1.851	(0.139)	9.787	(0.000)**	Text→Stat.
Econ. Policy Uncertainty		0.093	(0.964)	4.861	(0.003)**	Text→Stat.

Notes: Granger causality tests between TFIs and related official statistics are conducted on data from 2006 to 2023. Several TFIs, such as semiconductors, recruitment, and inflation outlook, exhibit a significant unidirectional Granger causality relationship with the corresponding official statistics.
* Indicate that the Granger statistics are significant at $\alpha = 0.05$.
** Indicate that the Granger statistics are significant at $\alpha = 0.01$.

LSTM layer, f, i, and o represent the forget, update, and output gates, respectively. c and h denote the cell and hidden state vectors.

$T_q = \{t_{i-2}, t_{i-1}, t_i\}$ is the index set of the quarter q that indicates 3 continuous months in quarter q. That is, for $\mathbb{X} \in \mathbb{R}^{p \times T}$, the kernel W_k is applied every 3 month while sliding through the input $\mathbb{X}$ and striding with 3 month time windows. These convolutional filters produce distinct bridge equations. To address the issue of ragged edges, we impute missing values that were unavailable during the analysis with the most recently available data.

CRNN is estimated using the gradient descent method. We employed the *TensorFlow* Python package for estimation, with training optimized using ADAM (Kingma & Ba, 2014). The training dataset was used repeatedly and evaluated for 5 epochs, each consisting of 500 iterations. To evaluate the performance of CRNN, we used the same data inputs as those for TFM, including official statistics alone, news texts alone, and both combined. These inputs are fed into the same CRNN structure, and their predictive accuracy is compared in Section 4.2.1.

However, we do not explore end-to-end language models in this study. As discussed in Seo et al. (2024), the recently prevalent transformer architecture could be utilized for nowcasting models. Note that relying on a single complex architecture may not suffice to generate multiple sector indices. Additionally, such models require costly training data. Therefore, we defer this modeling approach to future research endeavors.

4. Results

4.1. Validation of theme frequency indices

One advantage of computing TFIs by sector is that their validity can be verified by comparing them to relevant official statistics. As shown in Table 3 and 4, TFIs have a high correlation with, and tend to lead, the corresponding official statistics. Some TFIs exhibit interesting characteristics in Fig. 2 and Table 3, as outlined below:

- Some TFIs are found to lead official statistics based on cross-correlation and Granger causality tests. Whether the text indices lead varies by sector. One hypothesis is that text indices precede official statistics in sectors where expert opinions are frequently addressed rather than merely conveying released statistics.
  These sectors often cover broader topics and attract significant public interest.
- Text-based TFIs tend to fluctuate in short periods, indicating they include noise, yet no apparent outliers appear because the index quantifies public attention. In contrast, for house construction, the official statistics based on surveys show more apparent outliers.
- For the government expenditure TFI, the time lag showing the maximum correlation between the TFI and consolidated fiscal balance is 3 months, with the TFI leading. This is noteworthy because the official statistic is released with a 3-month delay.
- For the inflation outlook TFI, data analysis from 2006 to 2023 reveals a maximum correlation of 0.78 between the TFI and the CPI, with the TFI leading by 2 months. However, when the data is restricted to 2006–2021, the maximum correlation shifts to a 5-month time lag, with a value of 0.75. This change may be attributed to the increased difficulty in predicting inflation in 2022. Additionally, when computing the index without using terms such as 'forecast 전망', 'predict 예측', or 'expect 예상', the time lag remains 2 months for both analyses—whether including or excluding data from 2022 to 2023—with the TFI still leading. The correlations are slightly higher, at 0.81 and 0.76, respectively.

Table 5
Publication lags of variables used in this study.

Publication lags	Information Variables
2 months	Retail sales index
1 month	Unemployment rate, Employment to population ratio, Number of employed people, Export price index, Consumer price index, Price index excluding agricultural product & oil, Price index excluding food & energy, Producer price index, Facilities investment index, Manufacturing production index, Manufacturing industry shipment index, Manufacturing inventory index, Service industry production index, National accounts (GDP, Private consumption, Construction, Facilities investment, Exports of goods and services)
Available at the end of the month	All industries sales BSI, All industries performance BSI, Manufacturing export BSI, Manufacturing operation rate BSI, Manufacturing new orders BSI, Manufacturing domestic demand sales BSI, Manufacturing business performance BSI, Economic sentiment index, Current economic judgment CSI, Consumer sentiment index, Housing sales price index (Seoul, National), Housing lease price index (Seoul, National)
Available in real time	Financial indices (Call rate, CD rate, 3-year Korean treasury bond rate, Exchange rates, KOSPI, KOSDAQ, VKOSPI, Dubai crude oil, WTI futures, gold futures), 15 sector TFIs, News sentiment index, and Economic uncertainty index

Notes: The availability of variables for forecasting fluctuates depending on when the forecasting is conducted. For instance, the manufacturing production index becomes available at the end of the following month, typically resulting in a one-month publication lag. However, if forecasting is conducted in the early or middle of a month, this lag extends to two months. In contrast, textual data from news sources remains consistently accessible daily until the time of forecasting.

- To compute the inflation, stock price, and house price outlook TFIs, we included the words 'forecast 전망' and 'predict 예측'. These TFIs have time lags, with the maximum correlation at two, four, and six months. This finding aligns with our expectations for the forecasting horizons of interest, which are relatively short (around one month to one quarter) for stock prices and inflation and relatively long (around half a year to a year) for housing prices. This outcome is reasonable, considering that TFIs reflect the opinions and forecasts of sector experts in news articles.

Table 3 and 4 demonstrate that TFIs for each topic are useful as leading indicators for the corresponding sector and can be used independently to identify trends in the sector. Therefore, one potential application of TFIs is to use them as instrumental variables (IV) to measure public interest in each topic. Fig. 2 provides detailed time series comparisons between TFIs and official statistics, and Fig. 3 shows the ±6 month cross-correlation plots between them.

4.2. Forecasting results with the text-enhanced factor model

4.2.1. Validation of predictive accuracy improvement through textual information

This subsection examines how incorporating textual information in economic forecasting models affects predictive accuracy. In this study, we focus on predicting GDP for the current and following quarters, with predictions made at the end of each quarter. Table 5 presents all variables used in this study. It is important to note that the availability of these variables for forecasting varies depending on the timing of the forecasting process. Nevertheless, textual data from news sources remain consistently accessible up to the point of conducting the forecasting, enabling the utilization of TFIs even in the middle of the month.

To evaluate the testing error of predicting GDP growth, we computed the mean absolute error (MAE) and the root mean square error (RMSE) for out-of-sample predictions from the first quarter of 2016 to the third quarter of 2023. Specifically, we made recursive predictions for the seasonally adjusted GDP quarter-over-quarter growth and non-seasonally adjusted GDP year-over-year growth rates for the current and following quarters, using the available vintage data at the end of the current quarter. To distinguish the impact of adding textual information from the impact of model selection on accuracy improvement, we computed the testing accuracy in a cross-tabulation of both data and models, as presented in Table 6. To use TFIs in statistical models, we alleviate the noisy fluctuations of TFIs by employing a 3-month backward moving average series and transforming them into year-over-year and month-over-month growth rates.

Based on the findings in Table 6, including textual information enhances the predictive accuracy of both linear and nonlinear models for the current and following quarters. The effect of textual information on GDP(SA) QoQ predictions for the following quarter is less obvious. In contrast, GDP(NSA) YoY growth rates for both the current and following quarters and GDP(SA) QoQ growth for the current quarter show that the inclusion of textual information decreases predictive errors. TFM estimates its factor by utilizing both the observed information, including textual information, and the autoregressive structure of the model. Hence, if the text indicators are suitable as complementary variables for each factor, the model's fitness improves, and the standard error of the predictions decreases. The Superior Predictive Ability(SPA) test (Hansen, 2005) in Table 7

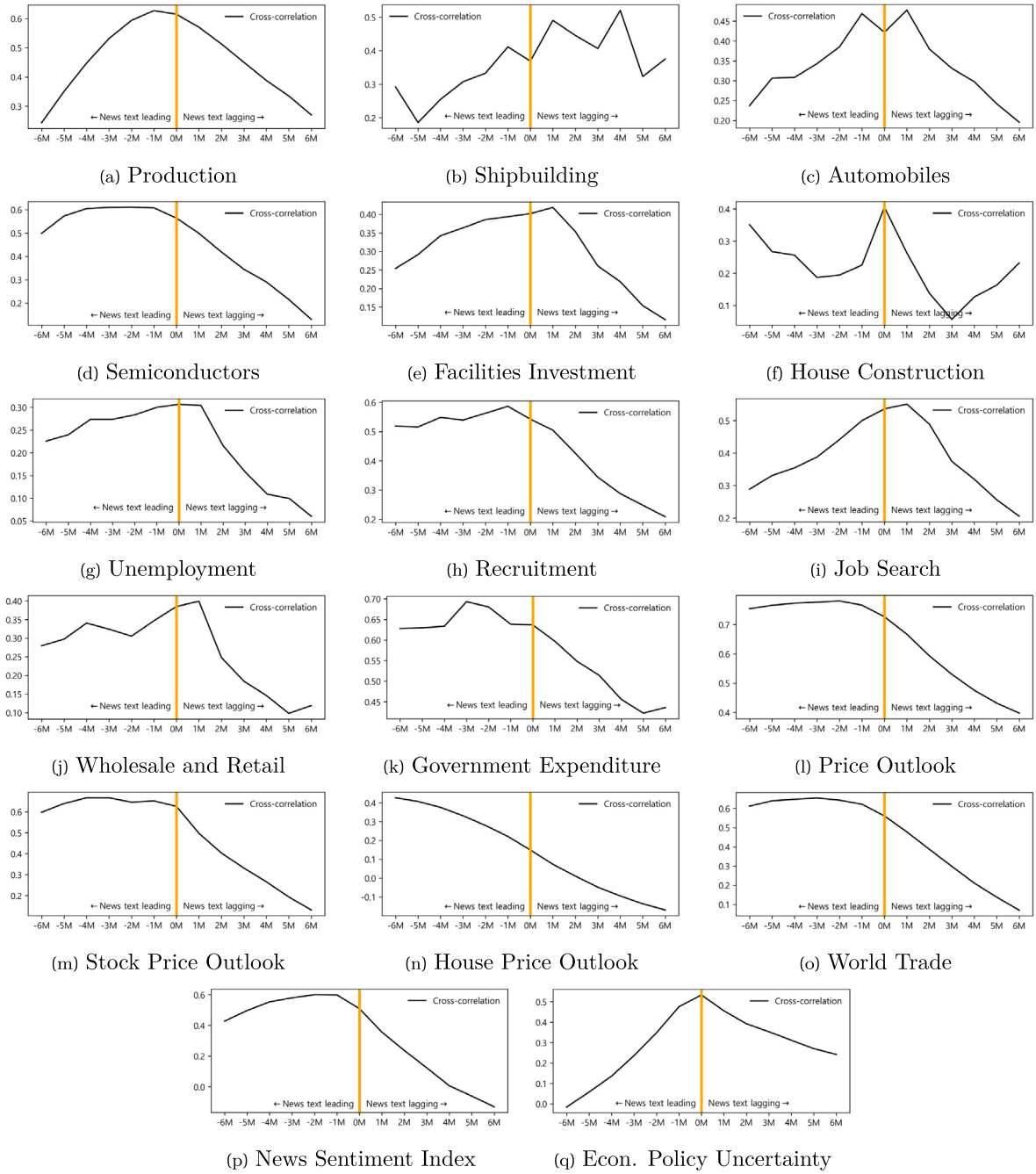


Fig. 3. Cross-correlation plots between sector text indices and their corresponding official statistics.
Notes: Cross-correlation between two series in Fig. 2 is illustrated with up to a six-month lead time. If the peak appears on the negative side of the x-axis, the maximum correlation occurs when the text index leads.

shows that TFM, augmented with textual information, outperforms the standard DFM.[5]

Additionally, in every instance, TFM and CRNN consistently outperform a benchmark autoregressive (AR) model by a significant margin. TFM mostly outperforms

[5] In comparable studies of GDP nowcasting in Korea during the COVID-19 pandemic, Lee, Kim, and You (2022) achieved an MAE of 0.513 for GDP(SA) using an XGBoost model with fixed variable selection and an MAE of 0.429 with dynamic optimization of variable selection based on 145 variables. Additionally, Jung, Pyo, and Heo (2022) achieved an RMSE of 0.99 and 0.87 for GDP(SA) using state-of-the-art LSTM and GRU neural network models with recursive training and an RMSE of 0.92 and 0.98 using rolling window training with these models based on 36 variables.

Table 6
Out-of-sample testing error with its standard deviation for predicting GDP(SA) and GDP(NSA) (Diebold & Mariano, 2002; Newey & West, 1987).

	Data / Model	Nowcasting			1Q Ahead Forecasting		
		Offi. Stat.	News text	OS+NE	Offi. Stat.	News text	OS+NE
• Prediction on GDP seasonally adjusted (SA) QoQ growth							
MAE	AR	0.686	–	–	0.657	–	–
		(.821)	–	–	(.793)	–	–
	CRNN	0.611	0.602	0.588	0.696	0.796	0.718
		(.696)	(.699)	(.542)	(.864)	(.881)	(.706)
	TFM[a]	0.460*	0.563	**0.448***	0.653	**0.619**	0.657
		(.415)	(.558)	(.415)	(.643)	(.614)	(.686)
RMSE	AR	1.060	–	–	1.020	–	–
		(2.841)	–	–	(2.841)	–	–
	CRNN	0.917*	0.914	0.794	1.098	1.177	0.999
		(2.131)	(1.887)	(1.137)	(3.014)	(3.272)	(2.015)
	TFM[a]	0.615	0.787	**0.606***	0.909	**0.865**	0.942
		(.618)	(1.204)	(.583)	(1.686)	(1.534)	(1.731)
• Prediction on GDP non-seasonally adjusted (NSA) YoY growth							
MAE	AR	0.996	–	–	1.302	–	–
		(1.040)	–	–	(1.374)	–	–
	CRNN	1.021	0.984	0.975	0.954**	1.037**	**0.898****
		(1.030)	(.873)	(.845)	(1.012)	(1.100)	(.992)
	TFM[a]	0.836	0.992	**0.723**	1.145	1.341	0.997
		(.735)	(.838)	(.673)	(.770)	(1.140)	(.818)
RMSE	AR	1.428	–	–	1.877	–	–
		(4.663)	–	–	(7.965)	–	–
	CRNN	1.439	1.306	1.281	1.379	1.499*	1.326**
		(4.337)	(3.154)	(2.976)	(4.031)	(5.203)	(4.657)
	TFM[a]	1.106	1.290	**0.980**	1.373	1.748	**1.281**
		(1.947)	(2.715)	(1.697)	(2.137)	(5.589)	(2.580)

Notes: Predictive accuracy is compared for data spanning from 2016.1Q to 2023.3Q for nowcasting and from 2016.2Q to 2023.4Q for one-quarter-ahead forecasting. MAE refers to mean absolute error, while RMSE indicates root mean square error. In most cases, the Text-enhanced Factor Model (TFM), which incorporates official statistics with news information, demonstrates higher predictive accuracy. AR: Autoregressive (baseline model). CRNN: Convolutional-Recurrent Neural Network. TFM: Text-enhanced Factor Model. Bold numbers indicate the smallest error for the corresponding forecasting horizon and error criterion.

[a] For the official-statistics-alone model, the results have been obtained from a standard DFM.

* Stand for the rejection of the (Newey & West, 1987) adjusted (Diebold & Mariano, 2002) test with the null hypothesis of equal predictive power to AR at the significant levels of 10%.

** Stand for the rejection of the (Newey & West, 1987) adjusted (Diebold & Mariano, 2002) test with the null hypothesis of equal predictive power to AR at the significant levels of 5%.

Table 7
Test statistics for the Superior Predictive Ability (SPA) test comparing TFM against the standard DFM.

Target \ Model	Nowcasting		1Q Ahead Forecasting	
	News text	OS+NE	News text	OS+NE
GDP(SA, QoQ%)	−0.103	0.013**	0.034	−0.005
	(1.00)	(.00)	(.67)	(1.00)
GDP(NSA, YoY%)	−0.156	0.113**	−0.196	0.148*
	(1.00)	(.03)	(1.00)	(.06)

Notes: The test statistics are computed using a rolling window method of 24 quarters, as outlined in the study by Hansen (2005).

* Denote rejection of the test under the null hypothesis of equal predictive ability at significance levels of 10%.

** Denote rejection of the test under the null hypothesis of equal predictive ability at significance levels of 5%.

CRNN, except for GDP(NSA) YoY growth in one-quarter-ahead forecasting. One possible explanation for TFM's superiority over the more complex CRNN in predicting GDP lies in the nature of GDP itself, which is essentially a linear sum of subcategories. Thus, the nonlinear structure of the CRNN model may not be necessary and could potentially lead to overfitting errors, particularly when complete information is available for nowcasting. Conversely, neural networks, in general, offer advantages when the target variable cannot be expressed as a simple linear combination of inputs and when new features synthesized from the network supplement the insufficient information provided by the original inputs.

For the CRNN in the current quarter's nowcasting, the model using official statistics shows inferior performance compared to the model using text indicators alone. This can be attributed to imputing missing values at ragged edges. The CRNN extrapolates each missing observation using each time series separately. Hence, the CRNN using official statistics rely on outdated information for extrapolation. In contrast, the CRNN using text indicators alone and that using text-enhanced data can incorporate more

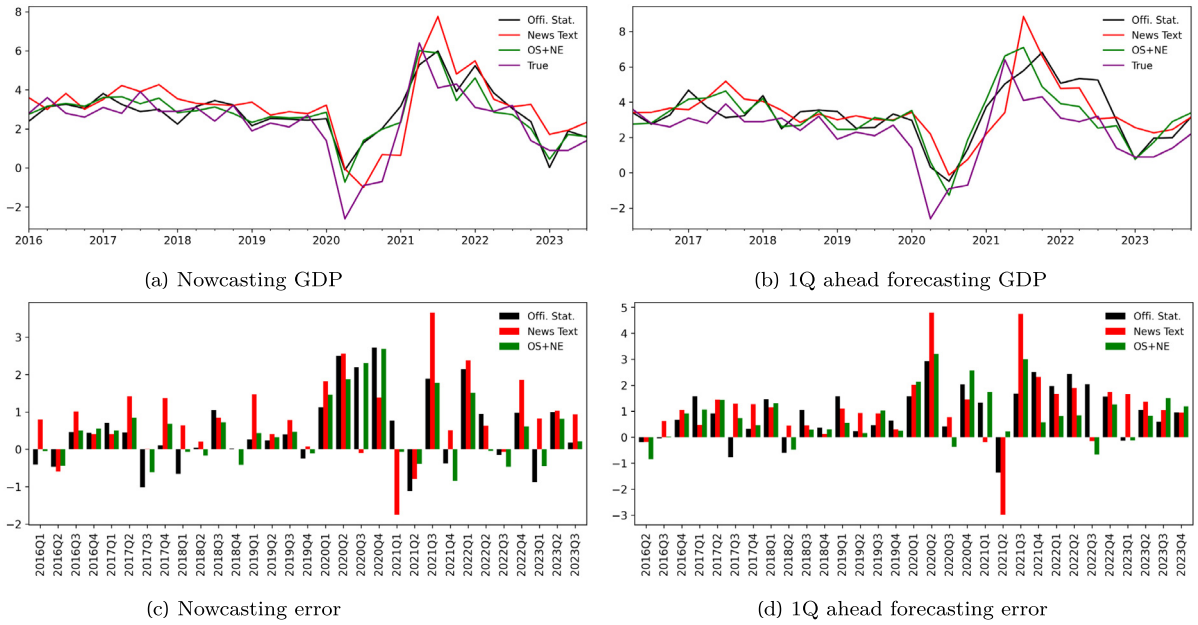


Fig. 4. Predicted GDP (NSA) year-over-year growth rates and their errors by quarter.
Notes: The predicted GDP from the TFM (green line) exhibits smaller out-of-sample errors compared to the standard DFM (black line) for both the current and following quarters. This difference is especially evident after 2021 when economic uncertainty was high due to consumption contraction stemming from the resurgence of COVID-19 and disruptions in the supply of automotive semiconductors caused by policy changes.

recent information using text. On the other hand, the TFM and DFM improve their predictions by imputing missing values with other observed variables, such as text and financial data, using an EM algorithm.

Our analysis demonstrated that when utilizing solely text indicators, the TFM's forecast error for GDP(SA) nowcasting was slightly higher than that of the DFM using only official statistics. This outcome can be attributed to the limited number of text indicators, which consist of only 17 text indices. However, despite this limitation, the TFM's accuracy remains comparable to that of the DFM. It outperforms the AR by a significant margin, suggesting that news sources possess additional information for economic forecasting. Fig. 4 illustrates the errors from three different information sets over time. The gap between the TFM utilizing both official statistics and textual information and the standard DFM is especially evident after 2021 when economic uncertainty was high in the Korean economy due to consumption contraction stemming from the resurgence of COVID-19 and disruptions in the supply of automotive semiconductors caused by policy changes. Furthermore, the prediction accuracy of major monthly economic variables in the TFM is presented in Table 8. The improvement in forecasting accuracy varies across variables. For monthly variables that are relatively timely and have ample available information, incorporating text indicators has a lower impact on improving predictive power compared to quarterly GDP.

The investigation extended to the DFM model, which was trained with a topic series generated by LDA. Due to the computational intensity involved in text normalization and LDA model training, we restricted the data to 876,600 sample articles, randomly selecting 300 news articles per day from 2015 to 2022. Despite integrating text information, the DFM combined with the LDA topic series did not yield significant improvements in predictive accuracy (refer to Section 1 in supplementary materials). This limitation may arise from the difficulty in effectively classifying sentiment in news articles through topic modeling. Alternatively, LDA could potentially be employed to analyze evolving issues over time, as it does not impose topic restrictions in advance, as suggested by Zheng et al. (2024). Nevertheless, utilizing unsupervised methods poses challenges in establishing connections between obtained indices and specific economic themes and in compiling information on less frequently mentioned issues. In contrast, TFIs are tailored to promptly capture real-time opinions on designated macroeconomic topics, focusing at the sentence level, as a single news article often presents various perspectives.

4.2.2. Exploring the characteristics of the text-enhanced factor model

In this subsection, the forecast accuracy of the TFM was examined for the GDP year-over-year growth rate by the end of March 2020, when the COVID-19 outbreak began. Fig. 5 demonstrates that the TFM enhanced the prediction accuracy of GDP in March 2020 compared to using official statistics alone. This highlights the value of textual information, particularly in scenarios where the economic landscape changes rapidly, and inadequate quantitative data is available. Furthermore, the figure illustrates that incorporating textual information may alter the mid-term forecasting trajectory for GDP by up to one and a half years within the considered TFM.

Table 8
Out-of-sample testing error with its std. dev. of DFM and TFM for predicting monthly variables.

Variable \ Data	Nowcasting		1M ahead FC		2M ahead FC	
	OS	OS+NE	OS	OS+NE	OS	OS+NE
Unemployment rate	3.436	3.408	4.378	4.314	5.861	5.884
	(2.855)	(2.838)	(2.779)	(2.85)	(4.598)	(4.642)
Consumer price index	0.173	0.169	0.235	0.203	0.243	0.239
	(0.129)	(0.126)	(0.156)	(0.159)	(0.165)	(0.201)
Manufacturing production ind.(SA)	2.315	2.22	1.712	1.753	2.06	2.037
	(2.207)	(1.762)	(1.372)	(1.21)	(1.625)	(1.498)
Service industry prod. ind.(SA)	0.737	0.772	0.667	0.73	0.885	0.904
	(0.973)	(1.048)	(0.646)	(0.562)	(0.813)	(0.828)
Facility investment ind.(SA)	3.786	3.613	3.034	3.093	3.691	3.772
	(2.71)	(2.599)	(2.16)	(2.301)	(2.254)	(2.389)
Construction completed(SA)	2.212	2.159	1.985	1.984	2.864	2.913
	(2.067)	(2.095)	(1.991)	(1.989)	(1.941)	(2.089)
Housing sales price ind.-Nat.	0.182	0.143	0.118	0.121	0.193	0.163
	(0.159)	(0.146)	(0.086)	(0.093)	(0.157)	(0.166)

Notes: The calculations are performed between March 2016 and December 2023, measured by mean absolute error (MAE).

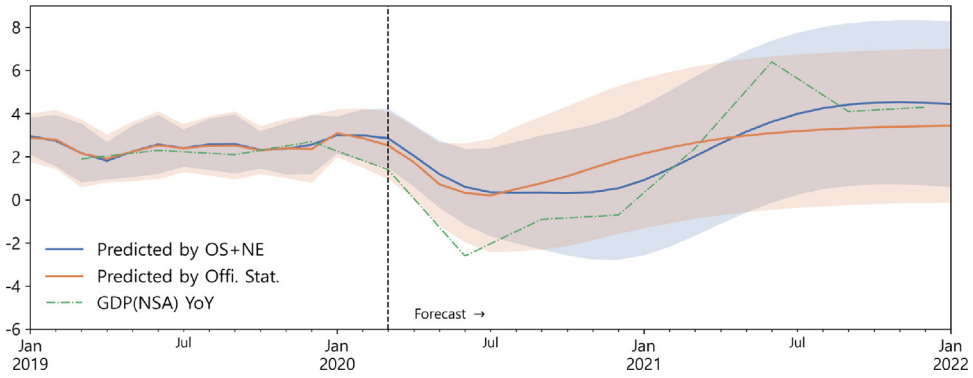


Fig. 5. The GDP forecasting comparison between TFM and DFM during the COVID-19 outbreak.
Notes: The comparison was made in March 2020, when the COVID-19 outbreak occurred. The DFM is based solely on official statistics, while the TFM is enhanced with textual information.

However, it is worth noting that evaluating machine learning models based on their long-term forecasting performance is unreasonable, as it is challenging to assume that the inputs contain information about the long-term forecast horizon, even if a particular model shows lower prediction errors. In Fig. 5, the superior forecasting performance of the TFM for the one-year horizon compared to the DFM can be attributed to the TFM model anticipating a deeper trough in the following quarter, June 2020. The autoregressive structure of the linear factor model then determined the subsequent forecast path.

In Fig. 6, we can see the coefficient of determination, R^2, for each factor with respect to each variable, x, in the vintage data as of March 2020.

$$R_x^2 = \frac{\sum(x_t - \hat{x}_t)^2}{\sum(x_t - \bar{x})^2} \tag{19}$$

$$\hat{x}_t = \hat{\Lambda}\hat{f}_t \tag{20}$$

The figure allows us to analyze how each factor relates to the observed variables by computing the explanatory power of each factor for each observed variable. The overall-1 factor has the strongest relationship with most text indicators, and the coefficient of determination of the text indicators for individual factors varies depending on their respective factors.

Tables 9 display the variables that had the greatest impact on the change in GDP forecasts between December 2019 and March 2020. For the current quarter (2020.1Q), the change in the forecast is primarily attributed to officially released statistics. In contrast, for the subsequent quarter forecasting (2020.2Q), textual information has the greatest influence on the change.

4.3. Exploring potential biases in news data for economic forecasting

An essential issue when using news data for forecasting is the various potential biases inherent in news reporting. Individual news articles may intentionally or unintentionally convey false information due to factors such as sensationalism or the media's pro-government or pro-business stance. Even if individual news articles are assumed to be truthful, the aggregation of news can still lead to opinion imbalances, where dominant views are overrepresented, and minority perspectives or less publicly prominent issues are underreported, causing distortions in collective information.

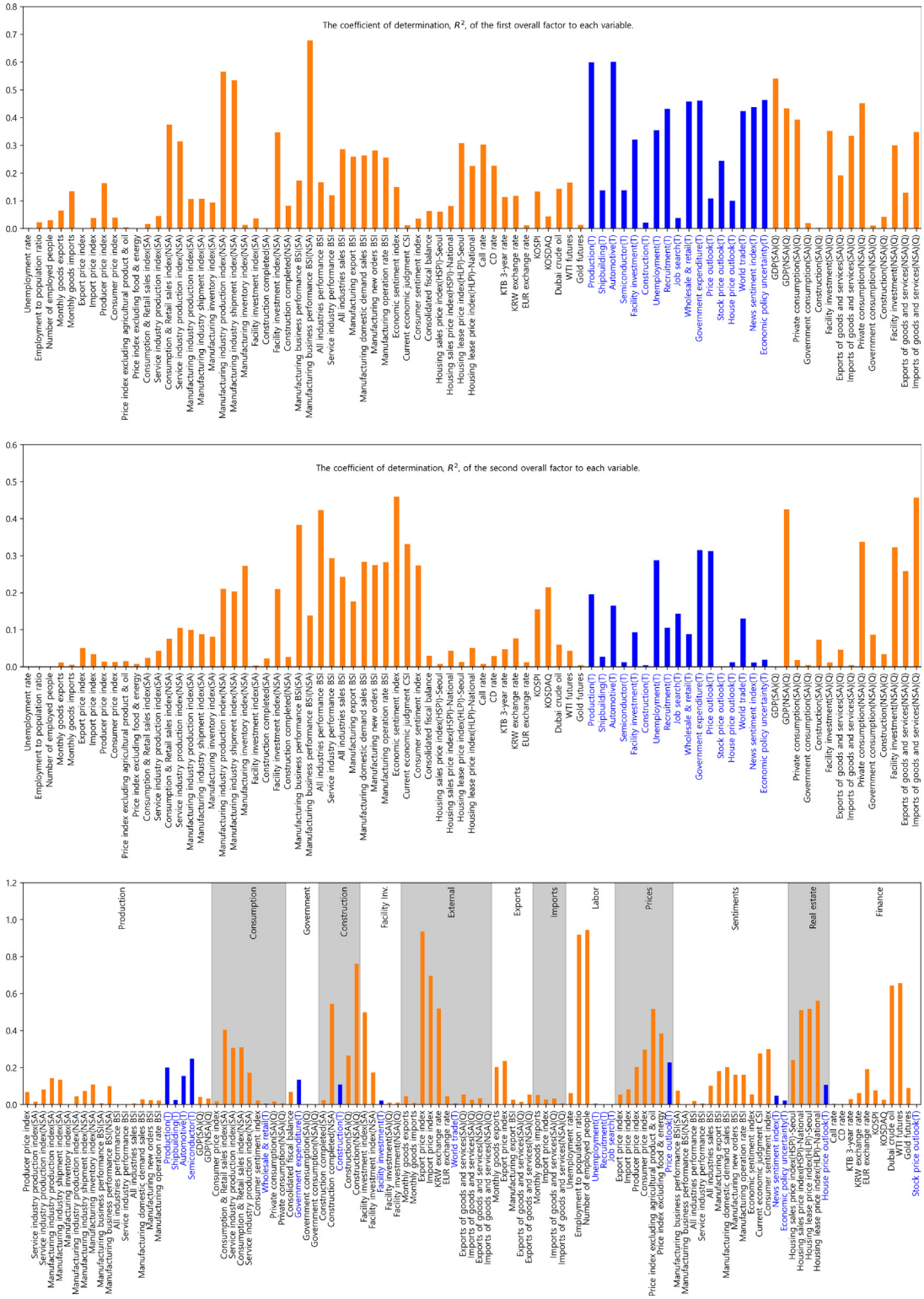


Fig. 6. The explanatory power, R^2, that each factor has for each observed variable.
Notes: The R^2 values are computed using the vintage data as of March 2020. In the figures, blue indicates variables computed from textual data, while orange represents variables obtained from other sources. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 9
The variables affecting the GDP(NSA) forecasts for the first and second quarters of 2020.

		Observed	Prev. forecast	News	Weight	Impact on GDP
•For forecasting the 1st quarter of 2020						
2020–01	Manufact. ind. shipment(NSA)	−6.89	2.72	−9.61	0.03	−0.27
	Manufact. ind. production(NSA)	−4.62	3.77	−8.39	0.02	−0.16
	Manufact. ind. shipment(SA)	−4.06	0.33	−4.39	0.03	−0.12
2020–02	Manufact. ind. shipment(SA)	−4.30	0.51	−4.82	0.02	−0.10
2019–12	Manufact. ind. shipment(NSA)	5.67	1.37	4.30	0.02	0.09
2020–01	Manufact. ind. production(SA)	−3.62	0.59	−4.21	0.02	−0.09
2020–02	Manufact. ind. production(NSA)	10.01	5.23	4.78	0.02	0.09
	All industries performance BSI	−10.92	1.31	−12.23	−0.01	0.08
	Manufact. ind. production(SA)	−3.78	0.53	−4.31	0.02	−0.08
2019–12	Manufact. ind. production(NSA)	7.30	1.80	5.51	0.01	0.07
•For forecasting the 2nd quarter of 2020						
2020–03	Government expenditure(T)	1.42	−0.20	1.62	1.47	2.38
	Production(T)	−0.24	0.24	−0.48	−2.47	1.19
2020–02	Government expenditure(T)	0.43	−0.14	0.58	−1.85	−1.07
	World trade(T)	2.45	0.17	2.27	0.34	0.77
	Production(T)	0.00	0.22	−0.22	2.99	−0.67
	Semiconductors(T)	0.67	0.10	0.57	−1.04	−0.59
2020–03	Semiconductors(T)	0.61	0.09	0.52	0.87	0.45
	Automobiles(T)	0.00	0.28	−0.28	−1.53	0.42
	KOSPI	−16.45	1.44	−17.89	0.02	−0.36
	Recruitment(T)	−0.23	0.02	−0.25	−1.24	0.31

Notes: At the end of March 2020, the variables affecting the predicted GDP (NSA) year-over-year growth rate for the first and second quarters of 2020 were compared to the values predicted in December 2019. The majority of the new information used for nowcasting the current quarter was derived from officially released statistics, while textual information was primarily relied upon for forecasting the subsequent quarter.

These inherent biases are difficult to eliminate in advance. However, utilizing a diverse range of news sources can help reduce the issue of false information from inaccurate reports, as suggested by the theory of collective intelligence (Pröllochs & Feuerriegel, 2023). Nevertheless, the distortion of collective information, which persists even after aggregating individual news articles, is not easily resolved by simply increasing the number of news sources.

Despite this, such information can still be useful for forecasting, even if it does not fully reflect reality. For instance, between 2020 and 2021, COVID-19 case numbers were crucial quantitative data for forecasting private consumption. However, as public sensitivity to the pandemic decreased after 2022, case numbers no longer significantly influenced private consumption. Yet, incorporating news mentions of COVID-19 into models allows the development of quantitative indicators that reflect diminishing public interest, aligning with the forecaster's intent.

This paper aims to incorporate the broadest possible range of publicly accessible news sources in Korea into the models rather than restricting sources to only a few reputable media outlets. Alternatively, to mitigate potential biases, news data can be pre-processed using various false information detection (FDI) methods (Guo, Ding, Yao, Liang, & Yu, 2020; Tchakounté, Faissal, Atemkeng, & Ntyam, 2020) before being incorporated into the models.

5. Conclusion

This paper proposes a new method for generating news-text-based economic indicators called *Theme Frequency Indices* (TFI), which reflect economic trends across sectors. TFIs are computed using simple subject-predicate word group patterns based on domain knowledge without requiring labeled data. Because these text indicators are computed for specific economic fields, they can be easily verified and used as timely alternatives to corresponding official statistics. Empirical analysis demonstrates that TFIs, computed using simple text-mining techniques combined with domain knowledge, offer highly correlated insights and often precede official statistics. This is crucial because TFIs can help circumvent costly training procedures. Furthermore, this paper investigates suitable econometric models for incorporating textual information into a forecasting model for Korea's gross domestic product (GDP). A model structure incorporating TFIs into separate factors, referred to as the *Text-enhanced Factor Model* (TFM), is proposed, with empirical examinations of its predictive accuracy and characteristics. The empirical study shows that TFM improves near-term economic forecasting accuracy for the GDP of Korea compared to models relying solely on official statistics.

The findings of this study contribute to related research in two folds. First, the proposed text-mining technique using subject-predicate patterns can be applied to numerous topics for analyzing textual information without needing expensive labeled data. By utilizing ubiquitous news sources and automated algorithms, this approach can complement traditional surveying methods at significantly lower costs and less time. Moreover, this study illustrates the potential of textual information for economic forecasting, a field still in its early stages. The results of this study emphasize that information from news sources is valuable by providing big-data-driven insights into public perceptions of the economy.

Acknowledgments

The author sincerely appreciates the support provided by the members of the Office of Economic Modeling and Policy Analysis at the Bank of Korea throughout this research.

Data and code availability

Supplementary materials pertaining to this article are available online. All code and data are available at https://github.com/seo-beomseok/econometric_forecasting_via_news.Additionally, the TFI dataset depicted in Fig. 2 can be accessed through the Text-Hub web application via https://seo-beomseok.github.io.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2024.11.001.

References

Babii, A., Ghysels, E., & Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, *40*(3), 1094–1106.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, *131*(4), 1593–1636.

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. *Handbook of Economic Forecasting*, *2*, 195–237.

Barbaglia, L., Consoli, S., & Manzan, S. (2022). Forecasting with economic news. *Journal of Business & Economic Statistics*, 1–12.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, *10*, 615–643.

Bybee, L., Kelly, B., Manela, A., & Xiu, D. (2021). Business news and business cycles. *The Journal of Finance*.

Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, *112*(4), 1194–1225.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Ferreira, L. N., et al. (2021). *Forecasting with VAR-teXt and DFM-teXt models: exploring the predictive power of central bank communication*: *2021*, Banco Central do Brasil.

Gardner, B., Scotti, C., & Vega, C. (2022). Words speak as loudly as actions: Central bank communication and the response of equity prices to macroeconomic announcements. *Journal of Econometrics*, *231*(2), 387–409.

Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, *89*(5), 2409–2437.

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2020). The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys*, *53*(4), 1–36.

Hájek, P., & Olej, V. (2013). Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In *International conference on engineering applications of neural networks* (pp. 1–10). Springer.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, *23*(4), 365–380.

Hopp, D. (2022). Economic nowcasting with long short-term memory artificial neural networks (LSTM). *Journal of Official Statistics*, *38*(3), 847–873.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585–594.

Jung, S., Pyo, D.-J., & Heo, D.-Y. (2022). *Vol. 2022–8*, 딥러닝 방법론을 활용한 *gdp* 예측모형 연구 *[study of GDP prediction models using deep learning methods]*. National Assembly Budget Office of Korea.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2022). Making text count: economic forecasting using newspaper text. *Journal of Applied Econometrics*, *37*(5), 896–919.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, *61*(4), 1281–1317.

Lee, G.-H., Cho, J.-H., & Jo, J.-G. (2020). New economic policy uncertainty indexes for South Korea. *Korean Journal of Applied Statistics*, *33*(5), 639–653.

Lee, Y. J., Kim, S., & Park, K. Y. (2019). Measuring monetary policy surprises using text mining: The case of Korea. *2019–11*, BOK Working Paper.

Lee, Y., Kim, Y., & You, T. (2022). 딥러닝 시계열 모형을 이용한 당분기 GDP 예측 성능 분석 [Analysis of QoQ GDP prediction performance using deep learning time series model]. *Journal of KIISE (JOK)*, *49*(10), 873–883.

Lewis, C., & Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, *49*(5), 587–615.

Li, N., Liang, X., Li, X., Wang, C., & Wu, D. (2009). Network environment and financial risk using machine learning and sentiment analysis. *Human and Ecological Risk Assessment*, *15*(2), 227–252.

Li, J., Seo, B., & Lin, L. (2019). Optimal transport, mean partition, and uncertainty assessment in cluster analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *12*(5), 359–377.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management*, *57*(5), Article 102212.

Mæhle, N., HlÃ, T., Pranovich, M., Selander, C., et al. (2021). Taking stock of IMF capacity development on monetary policy forecasting and policy analysis systems.

Mariano, R. S., & Murasawa, Y. (2010). A coincident index, common factors, and monthly real GDP. *Oxford Bulletin of Economics and Statistics*, *72*(1), 27–46.

Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777–787.

Pröllochs, N., & Feuerriegel, S. (2023). Mechanisms of true and false rumor sharing in social media: collective intelligence or herd behavior? *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–38.

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), 941–948.

Sax, C., & Eddelbuettel, D. (2018). Seasonal adjustment by x-13arima-seats in r. *Journal of Statistical Software*, *87*, 1–17.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th python in science conference*: *vol. 57*, (61), (pp. 10–25080). TX: Austin.

Seo, B. (2023). AI알고리즘을 이용한 산업모니터링: 증권사 리포트 텍스트 분석 [industry analysis using AI algorithm: Analyst reports text mining]. *2023–5*, BOK Issue Note.

Seo, B., Lee, Y., & Cho, H. (2024). Measuring news sentiment of Korea using transformer. *Korean Economic Review*, *40*, 149–176.

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, *228*(2), 221–243.

Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, *48*, 272–281.

Tchakounté, F., Faissal, A., Atemkeng, M., & Ntyam, A. (2020). A reliable weighting scheme for the aggregation of crowd intelligence to detect fake news. *Information*, *11*(6), 319.

Te Liew, W., Adhitya, A., & Srinivasan, R. (2014). Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry*, *65*(3), 393–400.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, *38*(2), 393–409.

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, *38*(10), 13274–13283.

Zheng, T., Fan, X., Jin, W., & Fang, K. (2024). Words or numbers? Macroeconomic nowcasting with textual and macroeconomic data. *International Journal of Forecasting*, *40*(2), 746–761.

Emerging Markets Finance and Trade

ISSN: 1540-496X (Print) 1558-0938 (Online) Journal homepage: www.tandfonline.com/journals/mree20

Deposit Competition and Depository Institution Risk: Do Interest Margins Matter?

Jaeweon You & Beomseok Seo

To cite this article: Jaeweon You & Beomseok Seo (2025) Deposit Competition and Depository Institution Risk: Do Interest Margins Matter?, Emerging Markets Finance and Trade, 61:15, 4893-4913, DOI: 10.1080/1540496X.2025.2520886

To link to this article: https://doi.org/10.1080/1540496X.2025.2520886

Published online: 16 Jul 2025.

Submit your article to this journal

Article views: 165

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=mree20

EMERGING MARKETS FINANCE AND TRADE
2025, VOL. 61, NO. 15, 4893–4913
https://doi.org/10.1080/1540496X.2025.2520886

Routledge
Taylor & Francis Group

Check for updates

Deposit Competition and Depository Institution Risk: Do Interest Margins Matter?

Jaeweon You [a] and Beomseok Seo [b]

[a]Financial Stability Department, Bank of Korea, Seoul, Republic of Korea; [b]Department of Statistics, Sookmyung Women's University, Seoul, Republic of Korea

ABSTRACT

We investigate how intensified deposit competition affects the stability of depository institutions in South Korea. Using novel and granular institution-level data on deposit interest rate spreads from 2016 to 2023, we find that elevated deposit competition significantly reduces profitability and exacerbates earnings volatility, particularly during periods of financial market instability amid the monetary tightening phase between 2021 and 2023. Importantly, the adverse effects are more pronounced among depository institutions with narrower interest margins, which also experience declines in capital adequacy. Our findings provide stronger support for the competition-fragility hypothesis over the competition-stability view, indicating that intensified deposit competition under market stress disproportionately undermines the stability of institutions with limited pricing power. The results underscore the necessity of closer integration of deposit competition metrics into liquidity risk surveillance, alongside enhanced liquidity management and funding diversification by depository institutions.

KEYWORDS

Deposit competition; non-banking depository institutions; financial stability

JEL

G21; G28

Introduction

Deposits constitute a critical funding source for depository institutions, offering not only stable financing but also a foundational tool for liquidity management (Craig and Dinger 2013; Demirgüç-Kunt and Huizinga 2010). Sustaining robust deposit inflows is essential to support lending operations, preserve profitability, and maintain financial stability. However, when competition for deposits intensifies, institutions are often compelled to raise deposit interest rates, thereby increasing funding costs and compressing net interest margins. This margin compression erodes internal capital accumulation and constrains the institution's ability to absorb financial shocks, ultimately weakening its resilience (Anginer, Demirgüç-Kunt, and Zhu 2014). Although such dynamics have traditionally been associated with monetary tightening, recent evidence from emerging markets, including South Korea, suggests that financial market instability itself, often compounded by policy rate hikes, can serve as a more immediate trigger for intensified deposit competition (Arping 2017; Egan, Hortaçsu, and Matvos 2017; Fu, Lin, and Molyneux 2014).

The South Korean case offers a compelling illustration. During the recent monetary tightening cycle (2021–2024), the domestic financial market experienced acute liquidity stress stemming from disruptions in the commercial paper (CP) and repurchase agreement (RP) markets.[1] These dislocations significantly intensified funding pressures on depository institutions, prompting both commercial banks and non-banking depository institutions (NBDIs) to engage in aggressive deposit competition.[2] Notably, while banks can draw on more diversified funding sources, NBDIs remain

CONTACT Beomseok Seo bsseo@sookmyung.ac.kr Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Seoul 04310, Republic of Korea

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2025 Taylor & Francis Group, LLC

heavily reliant on deposits, rendering them particularly susceptible to destabilization amid episodes of financial market turmoil. Reflecting the severity of this pressure, the deposit interest rate spread, measured as the difference between deposit interest rates and benchmark CD rates, widened sharply in Q4 2022, reaching 83 basis points for banks and 142 basis points for NBDIs, the highest levels observed since 2014.

From an asset management perspective, intensified deposit competition amid financial instability can heighten risk-taking incentives for depository institutions (Delis and Kouretas 2011). As elevated deposit interest rates drive up funding costs and compress profit margins, institutions may respond by reallocating assets toward riskier exposures to preserve earnings. This behavioral shift is consistent with the competition-fragility hypothesis (Keeley 1990), which argues that heightened competition erodes franchise value, thereby encouraging excessive risk-taking and increasing systemic vulnerability. In contrast, the competition-stability hypothesis (Boyd and De Nicolo 2005) posits that greater competition can enhance financial stability by lowering lending rates, improving borrower solvency, and reducing default risk. Furthermore, competitive pressure may foster operational efficiency and stimulate financial innovation. While both theoretical frameworks offer plausible mechanisms, their empirical relevance remains unsettled, particularly in emerging markets where institutional settings, such as funding structures, regulatory environments, and pricing power (e.g., interest margins), exhibit considerable heterogeneity. Recent developments, including the South Korean experience, further suggest that financial market instability itself, beyond monetary tightening, can serve as a proximate trigger for intensified deposit competition.

Against this backdrop, this paper investigates how deposit competition affects the risk profiles of South Korean depository institutions, focusing on the unique episode in which financial instability, coinciding with a period of monetary tightening, served as the key driver. Utilizing novel and institution-level granular data, we construct a measure of deposit competition based on the spread between weighted-average deposit interest rates and corresponding certificate of deposit (CD) rates. This measure captures institution-specific funding pressure relative to market benchmarks, allowing us to assess how competitive dynamics influence risk-taking behavior at an institution level. Particularly, we examine how this relationship varies depending on institutions' pricing power, as proxied by the level of interest margins, and balance sheet characteristics.

Our empirical findings reveal that intensified deposit competition significantly reduces return on assets (ROA) and increases earnings volatility, consistent with the competition-fragility hypothesis (Keeley 1990). These effects indicate that higher funding costs, induced by elevated deposit interest rates, compress profit margins and prompt destabilizing shifts in portfolio risk profiles. Notably, institutions with narrower interest margins, those structurally less able to pass through rising funding costs to lending rates, are disproportionately affected. For these institutions, deposit competition not only deteriorates profitability and heightens earnings volatility but also weakens capitalization, as reflected in declining equity ratios.[3] These results underscore the critical role of institutional characteristics in mediating the risk effects of competition. Pricing power, shaped by interest margins, emerges as a key determinant of institutional resilience to funding shocks. The inability to adjust loan rates in line with deposit costs intensifies financial fragility, particularly under conditions of market-wide stress. This finding echoes the concerns raised by Egan, Hortaçsu, and Matvos (2017), who document how heightened competition, especially when combined with limited margin buffers, can amplify systemic vulnerabilities.

This paper makes three key contributions to the literature on deposit competition and financial stability. First, we show that financial market instability—rather than policy rate hikes alone—can serve as a direct and immediate trigger of intensified deposit competition in emerging economies, a dynamic that has received limited attention in prior studies (Craig and Dinger 2013; Egan, Hortaçsu, and Matvos 2017). Second, to our knowledge, this paper is the first to construct an institution-level measure of deposit competition by calculating deposit interest rate spreads at the individual institution level. This granular measure enables a more precise identification of institution-specific competitive pressures and uncovers significant heterogeneity in vulnerability

to funding cost shocks, particularly as a function of interest margin levels. Lastly, we contribute by empirically confronting the competition-fragility and competition-stability hypotheses within a unified empirical framework. By directly testing which mechanism is more dominant in the recent South Korean case, we demonstrate that deposit competition under financial market instability aligns more closely with the competition-fragility view, thereby advancing the literature on financial competition and systemic risk.

Our findings also carry important policy implications for both regulators and financial institutions. Regulators should closely monitor both liquidity and solvency risks, particularly among institutions with narrow interest margins, which are more vulnerable to destabilizing funding shocks and procyclical risk-taking. In this regard, temporary regulatory flexibility, such as calibrated adjustments to deposit interest rate caps, may help alleviate excessive competition without compromising market discipline. Furthermore, given the potential for deposit competition to trigger broader contagion effects through liquidity shortages, it is essential to incorporate systemic risk perspectives into regulatory frameworks. Lastly, depository institutions themselves should enhance their internal liquidity risk management capabilities and diversify their funding sources, thereby reducing over-reliance on deposit-based funding during periods of financial market stress.

The remainder of the paper is organized as follows. Section 2 documents recent trends in deposit competition among South Korean depository institutions. Section 3 presents the theoretical framework and hypotheses and Section 4 details the data, variable construction, and empirical methodology, including strategies to address potential endogeneity concerns. Section 5 discusses the empirical results. Section 6 concludes with policy implications and recommendations.

Deposit Competition Episode in South Korea in 2022

This section examines the episode of deposit competition in South Korea with a particular focus on deposit interest rate dynamics across depository institutions. We document institutional responses to rising funding pressures during the recent tightening cycle and highlight cross-sectoral variation in pricing strategies.

In response to wholesale funding constraints and tightening liquidity conditions between Q2 2022 to Q1 2023,[4] commercial banks in South Korea expanded their deposits bases significantly. This development triggered intensified deposit competition across the financial sector, as NBDIs, including mutual credit cooperatives and mutual savings banks, responded by sharply raising deposit interest rates to secure liquidity. Figure 1 illustrates the trends in deposit interest rates for banks and NBDIs alongside the BOK base rate, highlighting two distinct periods of monetary tightening. During the first tightening phase (2017.4Q ~ 2019.2Q), both banks and NBDIs gradually raised deposit interest rates, but the spread between them remained relatively narrow. However, in the second tightening phase (2021.3Q ~ 2023.2Q), deposit rates surged more steeply, with NBDIs showing a particularly pronounced rise. Specifically, Figure 1(B) presents the deposit interest rate spread by sector of depository institutions, defined as the difference between the weighted average 1-year deposit interest rates on new transactions and the corresponding certificates of deposit (CD) rates, segmented by sector. Importantly, the escalation of deposit interest rates in this study is measured in relative rather than absolute terms. This approach examines how a depository institution's deposit interest rates increased relative to prevailing market rates, rather than merely comparing rate increases across institutions. This analysis seeks to determine the extent to which each sector competitively raised deposit interest rates to attract deposits during the sample period.

Following the latter half of 2021, a distinct divergence in deposit interest rate spreads has emerged between commercial banks and regional banks. Prior to this period, the spreads exhibited relative parity; however, a sharp increase in deposit interest rates by commercial banks has triggered a notable widening of the gap compared to their regional counterparts.[5] Notably, sector-specific divergent trends in deposit interest rate spreads have also been evident between NBDIs and commercial banks. From Q1 to Q3 2022, NBDIs maintained relatively stable deposit interest rate spreads, while

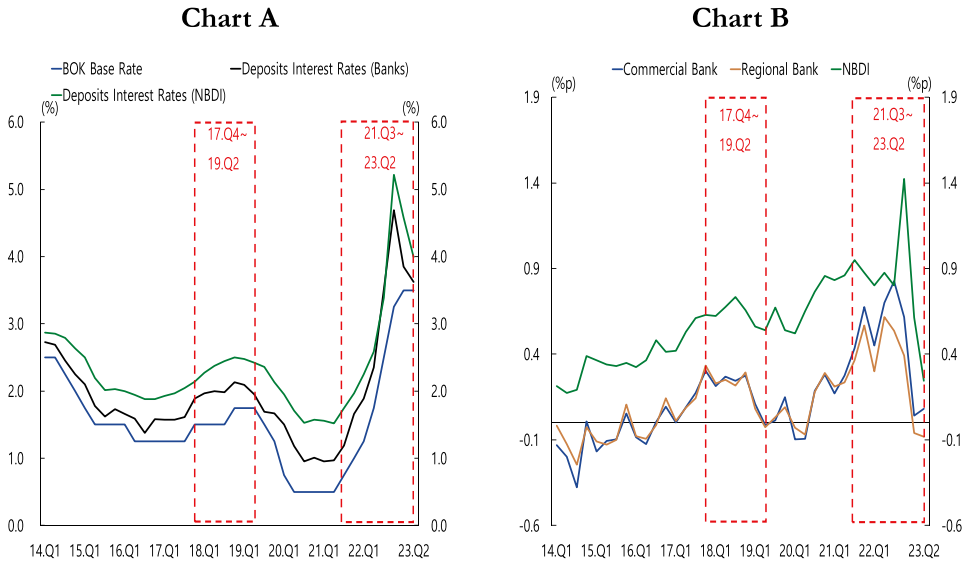


Figure 1. Deposit competition between banks and NBDIs. *Notes:* Chart A compares the Bank of Korea (BOK) base rate with the deposit interest rates offered by banks and non-bank depository institutions (NBDIs) over the period from 2014.Q1 to 2023.Q2. Chart B presents the average spread between the deposit interest rates of commercial banks, regional banks, and NBDIs and the market benchmark rate, proxied by the certificate of deposit (CD) rate. The charts highlight two distinct periods of monetary tightening: 2017.Q4–2019.Q2 and 2021.Q3–2023.Q2.

commercial banks experienced a consistent upward trend. By Q3 2022, commercial banks reached a spread of 83 basis points, the highest level since 2014. Subsequently, in Q4 2022, NBDIs reacted by raising deposit interest rates more aggressively, pushing their spread to 142 basis points.

This shift reflects a strategic response by NBDIs, which face more binding funding constraints than commercial banks. While banks have access to diversified funding sources, such as bank debentures and interbank borrowing, NBDIs remain heavily reliant on retail deposits. This reliance renders NBDIs particularly vulnerable to competitive pressures during periods of liquidity stress. As deposit competition intensified, NBDIs adopted more aggressive pricing behavior to retain deposit inflows and maintain liquidity[6] under the intensified inter-institutional competition.[7]

Theoretical Background and Hypotheses Development

Competition-Fragility or Competition-Stability Theory

The relationship between market competition and financial stability has long been central to banking theory, shaped by two competing frameworks: the competition-fragility and competition-stability hypotheses. The competition-fragility view posits that intensified competition undermines banks' charter values, weakening incentives for prudent risk management (Keeley 1990; Vives 2016). In this framework, heightened deposit competition may compress interest margins, limit internal capital accumulation, and push institutions toward riskier asset-side behaviors in pursuit of profitability, thereby increasing financial fragility. Particularly, margin compression not only affects solvency but also undermines liquidity resilience. As institutions aggressively raise deposit interest rates to attract funding, they may erode profitability and destabilize their funding bases (Craig and Dinger 2013; Egan, Hortaçsu, and Matvos 2017). Weakened earnings reduce the ability to maintain precautionary liquidity buffers, while reliance on price-sensitive deposit inflows increases exposure to abrupt withdrawals during market stress. These dual vulnerabilities, deteriorating solvency, and weakened liquidity, serve as key transmission channels through which deposit competition elevates institutional risk.

Conversely, the competition-stability hypothesis offers a fundamentally different perspective. The view suggests that competition leads to lower lending interest rates, improves borrower solvency, and disciplines excessive risk-taking by eroding monopoly rents (Boyd and De Nicolo 2005). Competitive pressures may encourage institutions to strengthen operational efficiency, improve borrower screening, and adopt more prudent risk management frameworks. From this standpoint, competition does not inherently destabilize financial institutions; under certain conditions, this mechanism can enhance resilience and promote systemic stability (Claessens and Laeven 2004). Interestingly, reconciling these divergent views, Martinez-Miera and Repullo (2010) propose a non-linear, U-shaped relationship, suggesting that competition initially strengthens stability by disciplining banks and enhancing borrower quality. However, once competition intensifies beyond a threshold, the resulting margin compression outweighs the benefits, reigniting risk-taking incentives and weakening institutional stability. This framework highlights that the effect of competition is not uniform, but depends critically on its intensity and on the institution's ability to absorb profitability and funding pressures.

Adding further nuance, the strategic motivations behind deposit competition also matter. When driven by balance sheet expansion or long-term growth objectives, deposit competition tends to be measured and profitability-aware (Kashyap, Rajan, and Stein 2002). However, when fueled by external pressures, such as liquidity constraints or volatile funding conditions, institutions may respond with aggressive deposit pricing, intensifying their exposure to margin compression and liquidity stress. Such vulnerabilities are more likely to be pronounced during periods of financial market instability under monetary tightening, when rising deposit costs and sluggish loan repricing amplify margin pressures.

Lastly, while the theoretical link between deposit competition and institutional risk has been extensively explored in advanced economies, its relevance in emerging market contexts remains less understood. In such markets, depository institutions tend to rely more heavily on deposit funding due to underdeveloped capital markets and limited access to diversified funding channels (Beck, Demirgüç-Kunt, and Levine 2013; Demirgüç-Kunt and Huizinga 2010). This structural dependence heightens institutional sensitivity to funding shocks and constrains their ability to reprice loans or restructure liabilities in response to market volatility. Despite these distinctive characteristics, relatively few studies have empirically examined how deposit competition affects financial stability in emerging markets, leaving a significant gap in the literature. In this regard, El-Chaarani, Abraham, and Azzi (2023) provide relevant insights into institutional resilience in emerging markets. Analyzing banks in the MENA region, El-Chaarani, Abraham, and Azzi (2023) show that stronger liquidity creation is associated with higher profitability and reduced credit risk during the COVID-19 crisis, highlighting the role of internal liquidity buffers in absorbing funding shocks. Although the study does not directly examine deposit competition, it highlights the critical role of funding structure and internal financial capacity in shaping risk outcomes of financial institutions in emerging markets. More importantly, this underscores the need for further research that explicitly investigates how deposit competition interacts with institutional risk in such contexts.

Given the competing theoretical views and the potential for deposit competition to be driven by external financial shocks, its implications for institutional stability remain unclear. In emerging markets, where related empirical evidence remains limited, this motivates the following open hypothesis:

H1: Deposit competition significantly affects the risk profiles of depository institutions, but the direction of the effect—stabilizing or destabilizing—depends on institutional characteristics such as profitability buffers and risk-taking incentives.

Interest Margin Heterogeneity

A critical yet often underexplored dimension in understanding how deposit competition translates into institutional risk is the degree of interest margin heterogeneity across depository institutions. Interest

margins, defined as the spread between lending and deposit interest rates, may serve as a key internal buffer that shields institutions from funding cost shocks and shapes their incentives for risk-taking (Bikker and Vervliet 2018; Claessens, Coleman, and Donnelly 2018; Chortareas, Garza-García, and Girardone 2012).

From the competition-fragility perspective, institutions with narrower interest margins have structurally thinner profitability cushions, reducing their ability to absorb rising funding costs typically associated with aggressive deposit competition (Keeley 1990; Vives 2016). When competition compresses margins further, these institutions face greater pressure to preserve earnings, potentially resorting to riskier lending practices or reducing liquidity buffers—both of which heighten solvency and liquidity risks. Margin compression, in this context, not only deteriorates current profitability but also exacerbates future vulnerability by weakening capital bases and increasing reliance on volatile funding. In contrast, institutions with wider interest margins are often assumed to be better positioned to maintain stability under competitive pressure. From the competition-stability view, higher margins afford more flexibility to absorb funding shocks without compromising prudential standards (Boyd and De Nicolo 2005). These institutions may continue to invest in robust screening mechanisms, maintain conservative lending portfolios, and avoid excessive risk-taking even amid rising deposit costs. In addition, stronger profitability may enable them to preserve liquidity buffers, thereby limiting their exposure to destabilizing outflows.

However, neither low nor high interest margins guarantee a uniform response to deposit competition (Laliberte and Sengupta 2024). For example, institutions with narrow margins, while generally considered more vulnerable, may respond to competitive pressures with more conservative strategies, such as improving operational efficiency or strengthening credit screening, particularly when survival incentives are acute. Conversely, institutions with high margins may interpret margin room as a license to expand risk exposures or pursue volume-based lending to defend market share. That is, low-margin institutions may adopt stabilizing strategies under appropriate regulatory or market conditions. Likewise, while high-margin institutions are better equipped to weather competition, they are not immune to destabilizing incentives. In this way, interest margins do not deterministically dictate institutional behavior but shape the strategic channels through which competitive pressure is internalized. Thus, interest margins can be viewed as moderating variables, interacting with competitive intensity and institutional incentives to produce heterogeneous risk outcomes.[8]

Taken together, the impact of deposit competition on institutional fragility or resilience is likely to vary across the interest margin distribution. This differentiated interpretation supports a conditional hypothesis:

H2: The effect of deposit competition on the risk profiles of depository institutions is moderated by the level of interest margins; institutions with different margin levels may exhibit heterogeneous risk responses to competition, either reinforcing or mitigating its destabilizing effects.

Data and Empirical Strategy

Data

The primary data for this study is sourced from two key datasets, both collected on a quarterly basis. First, we utilize granular deposit interest rate data for individual banks, obtained through a proprietary survey conducted by the Bank of Korea. This dataset is unique in that it provides institution-level information on deposit interest rates, allowing for a precise assessment of competitive pressures across institutions. Second, we draw on quarterly institution-specific data from the "Financial Institution Business Reports," which are submitted directly to the Bank of Korea by each institution. These reports offer detailed information on depository institutions' characteristics, including total assets, deposits, loans, and equity, which are crucial for assessing their scale and financial condition. By

integrating these two complementary datasets, we construct a panel data that enables a comprehensive analysis of how deposit competition influences institutional risk.

Descriptive statistics and Pearson correlations of the variables are presented in Tables 1 and 2. Our sample comprises 17 depository institutions in South Korea, including 6 commercial banks, 6 regional banks, 4 mutual credit cooperatives, and the aggregated mutual savings bank sector, which consolidates data for all 79 savings banks.[9] This comprehensive sample captures nearly all depository institutions in South Korea, excluding only for 5 specialized banks[10] and 3 internet-only banks[11] whose business models differ substantially from traditional banking institutions. These exclusions are justified to maintain sample consistency and analytical validity, as their funding structures and strategic priorities diverge sharply from those of conventional depository institutions. Focusing on institutions with comparable operational characteristics ensures the internal coherence of the analysis and strengthens the interpretability of the results. This sample composition allows us to examine the dynamics of deposit competition and institutional risk across a relatively homogenous group of systemically relevant depository institutions.

The analysis period spans from 2016, following the restructuring of the mutual savings bank sector and significant M&A activity involving several Domestic Systemically Important Banks (D-SIBs), to Q2 2023. This timeframe encompasses two distinct phases of monetary tightening in South Korea: the first from 2017.Q4 to 2019.Q2, characterized by gradual and moderate rate increases,[12] and the second from 2021.Q3 to 2023.Q2, marked by more rapid and substantial hikes.[13] Notably, the latter phase overlaps with the emergence of financial market instability in Q2 2022, including the "Legoland incident," which triggered a significant liquidity shock. These differences offer a valuable context for comparing how monetary tightening and financial instability jointly influence deposit competition and institutional risk. By examining both periods, the analysis investigates whether the effects of tightening cycles and market disruptions are period-specific or structurally consistent across different macro-financial environments.

Table 1. Descriptive statistics.

	Variable	# of Obs.	Mean	Median	Std. Dev.	1%	99%
Dependent Variables	Ln(Z-score)	450	3.644	3.654	0.714	1.540	5.479
	ROA volatility	450	0.304	0.212	0.392	0.066	2.477
	ROA	450	0.547	0.535	0.409	−0.463	1.930
	Equity ratio	450	8.134	7.880	1.553	5.060	12.71
Institution-specific Characteristics	Deposit interest rate spread	450	0.354	0.300	0.267	−0.050	1.160
	Ln (Total assets)	450	4.393	4.177	1.232	1.583	6.094
	Deposit-to-total debt	450	79.35	80.81	14.12	47.06	98.39
	Loan-to-total assets	450	72.77	75.32	9.30	45.04	82.69
	Delinquency ratio	450	0.918	0.511	1.212	0.124	5.825
	Interest income share	450	89.73	91.09	7.55	68.20	99.77
Macroeconomic conditions	Nominal GDP growth	450	4.09	4.30	2.53	−1.10	9.10
	Inflation (CPI)	450	1.98	1.50	1.61	0.00	5.80
	Credit spread (bp)	450	69.73	54.85	30.39	41.10	150.20

Notes: The table presents the descriptive statistics of the variables used in this study. The dependent variable, Ln(Z-score), is calculated as the natural logarithm of the 4-quarter moving average of the sum of return on total assets and equity ratio, divided by the 4-quarter standard deviation of return on total assets. ROA volatility represents the 4-quarter standard deviation of quarterly return on total assets. ROA is calculated as the 4-quarter moving average of the ratio of quarterly net income to total assets. The equity ratio is computed as the 4-quarter moving average of the ratio of quarterly equity to total assets. The explanatory variable, deposit interest rate spread, measures the difference between the quarterly interest rate of 1-year new term deposits and the 91-day CD yield. Ln(Total assets) represents the natural logarithm of total assets (in trillion won). The deposit-to-total debt and loan-to-total assets ratios serve as proxies for the funding and operational structures of depository institutions, respectively. The delinquency rate denotes the proportion of overdue loans (outstanding for more than one month) to total loans. The interest income share is calculated as the proportion of interest income to the total of interest and non-interest income. In addition, to control for macroeconomic conditions, we include the nominal GDP growth rate, inflation (CPI), and the credit spread, defined as the yield spread between 3-year AA-rated corporate bonds and 3-year government bonds, to capture market-wide credit risk.

Table 2. Correlation matrix.

	Ln (Z-score)	ROA volatility	ROA	Equity ratio	Deposit interest rate spread	Ln (Total assets)	Deposit-to-total debt	Loan-to-total assets	Delinquency ratio	Interest income share	Nominal GDP growth	Inflation
ROA volatility	**−0.837**											
	(0.000)											
ROA	**0.128**	**−0.231**										
	(0.004)	(0.000)										
Equity ratio	**0.358**	**−0.163**	**0.107**									
	(0.000)	(0.000)	(0.016)									
Deposit interest rate spread	**−0.095**	**0.227**	**0.175**	**0.303**								
	(0.043)	(0.000)	(0.000)	(0.000)								
Ln (Total assets)	−0.015	−0.079	0.028	**−0.287**	0.075							
	(0.743)	(0.076)	(0.533)	(0.000)	(0.111)							
Deposit-to-total debt	**−0.115**	0.018	**0.312**	**0.099**	**0.269**	**−0.296**						
	(0.009)	(0.689)	(0.000)	(0.026)	(0.000)	(0.000)						
Loan-to-total assets	0.013	**−0.136**	**0.438**	**−0.265**	−0.060	−0.054	**0.644**					
	(0.775)	(0.002)	(0.000)	(0.000)	(0.205)	(0.226)	(0.000)					
Delinquency ratio	**−0.299**	**0.355**	**0.385**	**0.279**	**0.536**	**−0.165**	**0.540**	0.077				
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.082)				
Interest income share	**0.357**	**−0.316**	0.034	**0.367**	0.028	**−0.153**	**0.423**	**0.296**	−0.007			
	(0.000)	(0.000)	(0.441)	(0.000)	(0.553)	(0.000)	(0.000)	(0.000)	(0.868)			
Nominal GDP growth	−0.039	0.070	−0.011	−0.005	−0.040	−0.008	0.008	0.041	−0.070	0.076		
	(0.383)	(0.112)	(0.800)	(0.908)	(0.400)	(0.851)	(0.856)	(0.354)	(0.116)	(0.087)		
Inflation	0.021	0.031	−0.018	**−0.102**	**0.464**	0.083	−0.013	0.044	−0.068	**0.124**	**0.314**	
	(0.641)	(0.491)	(0.678)	(0.021)	(0.000)	(0.063)	(0.779)	(0.319)	(0.127)	(0.005)	(0.000)	
Credit spread	**0.153**	**−0.099**	−0.007	−0.059	**0.282**	0.070	0.001	0.009	−0.027	−0.017	**−0.428**	**0.239**
	(0.000)	(0.025)	(0.869)	(0.182)	(0.000)	(0.115)	(0.983)	(0.835)	(0.542)	(0.696)	(0.000)	(0.000)

Notes: Table 2 presents the Pearson correlation coefficients between the variables used in the analysis. The table summarizes the linear relationships among institution-specific variables, including the deposit interest rate spread, log of total assets, deposit-to-debt ratio, loan-to-asset ratio, delinquency rate, and interest income share, as well as macroeconomic variables such as nominal GDP growth, consumer price index (CPI), and the credit spread, defined as the yield differential between 3-year AA-rated corporate bonds and 3-year government bonds. P-values are reported in parentheses to indicate the statistical significance of each correlation.

Research Design

We first employ a fixed-effects panel regression to analyze the nexus between deposit competition and the risk of depository institutions across phases of monetary policy (Borio et al. 2017; Jiménez et al. 2014).

$$\mathbf{Y}_{i,t} = \alpha_i + \beta_1 \text{ Deposit interest rate spread}_{i,t-1} + \beta_2 \text{ Interest rate hike period }_{t-1} + \beta_3 \text{ Deposit interest rate spread}_{i,t-1}{}^{*} \text{ Interest rate hike period }_{t-1} + \beta_4 \, BC_{i,t-1} + \beta_5 \, MC_{t-1} + \gamma F_i + \delta Y_t + e_{i,t}$$

In the equation, $\mathbf{Y}_{i,t}$ represents the log-transformed **Z-score** for depository institution i in quarter t, calculated as the four-quarter moving average of the sum of return on total assets and the equity ratio, divided by the four-quarter standard deviation of return on total assets. This calculation serves as a measure[14] of the financial health and stability of the depository institution over the specified period (Lepetit and Strobel 2015). In essence, a higher Z-score indicates a lower probability of insolvency, while a lower Z-score reflects a higher likelihood of financial distress.

The primary explanatory variables is the Deposit interest rate spread, defined as the difference between the 1-year term deposit interest rate and the certificate of deposit (CD) yield, at the individual institution level. The tightening period is a binary variable capturing two monetary policy phases: 1) Q4 2017 to Q2 2019 (7 quarters of moderate tightening), and 2) Q3 2021 to Q2 2023 (8 quarters of rapid tightening and market instability). The interaction term between the spread and the tightening period allows us to test whether the effect of deposit competition on risk is conditional on the macro-financial environment.

The variable BC includes a range of institution-specific characteristics to account for differences in the size, business model, and stability of depository institutions. These characteristics encompass log-transformed total assets, the ratio of deposits to total liabilities (funding structure), the ratio of loans to total assets (asset structure), delinquency rates (loan risk), and the proportion of interest income (profit structure). In addition, the MC controls for the macroeconomic and financial market conditions using nominal GDP growth, consumer price inflation (CPI), and the credit spread, which is defined as the yield differential between 3-year AA-rated corporate bonds and 3-year government bonds at a quarterly basis.

To control for unobserved time-invariant characteristics, institution-specific fixed effects (F_i) are included, while year dummies (Y_t) are used to account for time-specific factors. To mitigate potential endogeneity concerns, all continuous variables are lagged by one quarter, and are winsorized at the 1st and 99th percentiles to reduce the influence of outliers. Lastly, standard errors are clustered at the institution-level to account for serial correlation within entities over time.

Empirical Result

Baseline Result

Table 3 presents the baseline regression results examining the relationship between deposit competition and the risk of depository institutions, as measured by the Z-score, after accounting for institutions-specific characteristics and macroeconomic conditions. A key feature of the empirical strategy is the distinction between two discrete monetary tightening episodes: the earlier phase spanning from Q4 2017 to Q2 2019, and the more recent phase from Q3 2021 to Q2 2023, the latter of which coincides with heightened financial market instability. By isolating these distinct tightening periods, the analysis aims to capture not only average effects, but also period-specific dynamics in the link between competitive funding pressures and institutional risk. Particularly, the recent tightening cycle, characterized by both aggressive policy rate hikes and market-wide funding stress, provides a critical context to examine whether intensified deposit competition under stress amplifies institutional vulnerabilities.

Across all specifications, the results reveal substantial heterogeneity. Under stable or accommodative monetary conditions, the deposit interest rate spread exhibits a statistically insignificant relationship with Z-scores at a 5% level, indicating that deposit competition does not meaningfully erode institutional stability during periods of monetary policy stability. Similarly, during the earlier tightening cycle (2017–2019), the interaction between deposit competition and monetary tightening remains

Table 3. Deposit competition and the risk of depository institutions.

	(1) Ln (Z-score)	(2) Ln (Z-score)	(3) Ln (Z-score)
Deposit interest rate spread $_{t-1}$	**−1.499***	−0.791	−0.756
	(0.746)	(0.528)	(0.556)
Deposit interest rate spread $_{t-1}$ × Interest rate hike $_{21Q3\text{-}23Q2}$		**−0.907****	**−0.968****
		(0.369)	(0.401)
Deposit interest rate spread $_{t-1}$× Interest rate hike $_{17Q4\text{-}19Q2}$	0.028		−0.244
	(0.455)		(0.497)
Interest rate hike $_{21Q3\text{-}23Q2}$		**0.405*****	**0.425****
		(0.131)	(0.166)
Interest rate hike $_{17Q4\text{-}19Q2}$	0.023		0.068
	(0.229)		(0.241)
Ln (Assets) $_{t-1}$	−0.075	0.308	0.264
	(0.674)	(0.704)	(0.677)
Deposit/Debt $_{t-1}$	−0.027	−0.023	−0.024
	(0.056)	(0.054)	(0.054)
Loan/Assets $_{t-1}$	−0.005	−0.002	0.001
	(0.048)	(0.048)	(0.049)
Delinquency Rate $_{t-1}$	**−0.298****	**−0.315****	**−0.323*****
	(0.104)	(0.116)	(0.108)
Interest Income Share$_{t-1}$	0.008	0.008	0.008
	(0.008)	(0.008)	(0.008)
Nominal GDP growth $_{t-1}$	−0.028	−0.035	−0.036
	(0.025)	(0.030)	(0.023)
Inflation $_{t-1}$	0.054	0.039	0.044
	(0.058)	(0.062)	(0.060)
Credit spread $_{t-1}$	0.003	0.002	0.002
	(0.002)	(0.002)	(0.002)
Constant	5.999	3.864	4.007
	(3.959)	(3.878)	(3.942)
Institution FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	450	450	450
R-squared	0.177	0.192	0.192

Notes: This table reports the results of a panel fixed-effects regression analyzing the impact of deposit competition on the risk of depository institutions during periods of interest rate hikes from 2016.Q1 to 2023.Q2. The dependent variable, Ln(Z-score), is calculated as the natural logarithm of the 4-quarter moving average of the sum of return on assets and the equity ratio, divided by the 4-quarter standard deviation of return on assets. The Deposit interest rate spread is measured as the difference between the quarterly interest rate on new 1-year term deposits and the 91-day CD yield. Interest rate hike 21Q3-23Q2 is a dummy variable equal to 1 for the period from 2021.Q3 to 2023.Q2, and 0 otherwise. Interest rate hike 17Q4-19Q2 is a dummy variable equal to 1 for the period from 2017.Q4 to 2019.Q2, and 0 otherwise. Ln(Total assets) denotes the natural logarithm of total assets (in trillion won). The Deposit-to-debt and Loan-to-assets ratios serve as proxies for the funding and operational structures of depository institutions, respectively. The Delinquency rate denotes the proportion of loans overdue by more than one month, and the Interest income share is calculated as the ratio of interest income to the sum of interest and non-interest income. Macroeconomic controls include Nominal GDP growth, Inflation (CPI), and the Credit spread, defined as the yield spread between 3-year AA-rated corporate bonds and 3-year government bonds. Standard errors in parentheses are clustered at the institution level. ***$p < .01$, **$p < .05$, *$p < .1$.

statistically insignificant, suggesting that moderate and gradual interest rate increases exert limited destabilizing effects. In contrast, during the recent monetary tightening phase (2021–2023) with financial market stress, a clear pattern emerges; deposit competition has a statistically significant and negative effect on financial stability. Specifically, the coefficient on the interaction term between the deposit interest rate spread and the recent tightening dummy ranges from −0.907 to −0.968, implying that a one-percentage-point increase in the deposit interest rate spread during this period is associated with a decline of approximately 0.91 to 0.97 units in the Z-score. This deterioration in risk profiles underscores the nexus between intensification of competitive pressures and institutional risk during periods of financial market stress amid the recent tightening cycle. In addition, in terms of the interest rate hike dummy for 2021–2023, we observe a positive and highly significant sign across all models, indicating that monetary tightening, on average, contributed to enhancing the financial stability of depository institutions. This finding aligns with the theoretical expectation that rising

interest rates tend to widen net interest margins by increasing loan yields more rapidly than deposit costs, thereby improving institutions' profitability and capital buffers (Altavilla, Boucinha, and Peydró 2018). In a typical environment, such widening margins foster stronger earnings, better absorption capacity for shocks, and consequently, higher Z-scores.

Importantly, the contrasting signs of the interest rate hike dummy for 2021–2023 and the interaction term highlight an important nuance; while monetary tightening generally strengthens institutional stability, it can have the opposite effect under conditions of intensified deposit competition. That is, institutions that aggressively raise deposit interest rates to secure funding face their financial stability deteriorate, despite the favorable environment for net interest margins. This duality emphasizes that the benefits of monetary tightening are not uniformly distributed across institutions; rather, they depend critically on each institution's competitive position.

These findings naturally prompt the next question of why the destabilizing effects of deposit competition are pronounced during the 2021–2023 period but mute during the 2017–2019 phase. First, the Bank of Korea raised rates by 300 basis points in a short span, intensifying liquidity pressures and prompting aggressive deposit pricing. Unlike the earlier cycle, the recent phase coincided with market instability, exemplified by the "Legoland Incident," which triggered liquidity hoarding and increased reliance on retail funding. In contrast, wholesale markets remained stable during 2017–2019, helping contain deposit competition pressures. These differences suggest that the risk impact of deposit competition is highly contingent on macro-financial conditions.

Lastly, even after controlling for macroeconomic conditions, such as GDP growth, inflation, and credit spreads, the significant negative interaction between deposit competition and financial stability persists. This finding reinforces the interpretation that institutional vulnerabilities during 2021–2023 were not merely a reflection of adverse macroeconomic trends but were fundamentally driven by intensified competition for deposits in an environment of rapid monetary tightening and market stress.

Taken together, our results provide strong support for the view that the effects of deposit competition on financial stability are highly conditional on the monetary and financial environment. While monetary tightening generally strengthens institutional stability by widening margins, aggressive deposit competition, especially under stressed market conditions, can offset these benefits and amplify institutional vulnerabilities, supporting the competition-fragility view. Policymakers should thus recognize the critical importance of liquidity conditions and competition dynamics when evaluating the systemic implications of monetary tightening cycles.

Additional Analysis

To reinforce the robustness of our empirical findings, we conduct a series of additional analyses. The detailed results of these robustness checks, including the random effects specification, the two-quarter lagged model, and the instrumental variable (IV) fixed effects model, are reported in Appendix A1 (Vallascas, Mollah, and Keasey 2017).

While the fixed effects specification employed in the baseline analysis effectively mitigates concerns related to unobserved time-invariant heterogeneity across depository institutions, it substantially limits cross-sectional variation by relying on within-institution fluctuations. To assess the robustness of our results under an alternative specification that retains both within- and between-institution variation, we re-estimate the baseline models using a random effects approach with year dummies to control for time-specific shocks. Although the random effects estimator is less stringent than fixed effects in addressing institution-specific unobserved heterogeneity, it allows for the preservation of broader cross-sectional information across institutions. As reported in Appendix A1, the results remain qualitatively consistent with those of the baseline fixed effects model, suggesting that the main findings are not sensitive to the choice of estimator. Second, although the use of one-quarter lagged explanatory variables in the baseline regressions partially mitigates concerns about reverse causality, we further strengthen this approach by extending the lag structure. Specifically, we re-estimate the models using two-quarter lagged values of the deposit interest rate spread, institution-specific characteristics, and macroeconomic variables, following Boone et al. (2007),

Faleye, Kovacs, and Venkateswaran (2014), Faleye (2015), and Vallascas, Mollah, and Keasey (2017), who advocate using longer historical lags to enhance the plausibility of predetermined regressors. While the extended lag reduces statistical precision to some extent, the estimated coefficients remain directionally and economically consistent with the baseline results, reinforcing the validity of the causal interpretation. Third, to formally address potential dynamic endogeneity, we implement an instrumental variable (IV) fixed effects model. In this specification, we treat the lagged deposit interest rate spread as endogenous and instrument it with its third- and fourth-quarter lagged values. This strategy assumes that historical deposit competition pressures influence current outcomes but are unlikely to be contemporaneously correlated with shocks to institutional risk. As reported in Appendix A1, standard diagnostic tests, including the Kleibergen-Paap rk Wald F-statistic and the Hansen J-test, confirm the relevance and validity of the instruments. Importantly, the IV estimates corroborate the baseline results, providing further assurance that our findings are not driven by reverse causality or omitted variable bias.

Collectively, these additional analyses lend strong support to the robustness of our main findings. While causal identification challenges persist in empirical banking research, the convergence of results across alternative model specifications, lag structures, and identification strategies enhances confidence in our results.

Deposit Competition and the z-Score Components

To further validate the baseline findings and assess the robustness of the underlying mechanism, we decompose the Z-score into its constituent components, namely, the standard deviation of the return on assets (ROA), ROA, and the equity ratio, and re-estimate the model using each as the dependent variable. This approach allows us to test whether the destabilizing effect of deposit competition operates through at least one of three risk channels: reduced profitability, heightened earnings volatility, or weaker capitalization. Evidence of a significant effect in any of these components reinforce our baseline results and confirm that the observed relationship reflects a structural channel of financial fragility, not a spurious regression artifact (Vallascas, Mollah, and Keasey 2017).

Table 4 presents the decomposition results of the Z-score into its three components -ROA volatility, ROA, and the equity ratio. Particularly, the analysis focuses on how deposit competition transmits risk during the recent tightening cycle (2021Q3–2023Q2), when its adverse effects on institutional stability were most pronounced. During this period, the interaction term between deposit competition and monetary tightening is both positive and highly significant for ROA volatility (coefficients between 0.727 and 0.746). Given the average volatility level of 0.304, a one-percentage-point increase in the deposit spread would more than double typical volatility, indicating a sharp deterioration in income predictability. A similar pattern is observed for ROA: the interaction effect is strongly negative (−0.617 to −0.651), nearly offsetting average profitability levels (mean ROA = 0.547). These results suggest that, under conditions of financial market stress, deposit competition undermines stability mainly through earnings volatility and profit compression, thereby weakening institutions' ability to generate capital internally. As for the equity ratio, no significant effect is observed during the recent period, reinforcing that income-based channels are the dominant path of fragility during times of acute funding pressure.

Although this decomposition is motivated by findings from Table 3 showing that deposit competition significantly affects institutional stability only during the recent tightening cycle, additional evidence from the earlier period (2017Q4–2019Q2) reveals that the effects are not entirely absent. In this earlier phase, deposit competition is associated with a modest but statistically significant decline in both ROA and the equity ratio, even though the overall Z-score is not significantly affected. Specifically, we observe a significant negative interaction with ROA, albeit with a smaller magnitude than in the recent cycle, and a negative and significant interaction (−0.624) with the equity ratio. These results indicate that even under milder monetary conditions, deposit competition can exert destabilizing effects through specific risk components—effects that may not be immediately evident at the aggregate level.

In sum, the decomposition in Table 4 results emphasize that the destabilizing effects of deposit competition are most acute when monetary tightening coincides with financial market stress, operating

Table 4. Z-score decomposition.

	(1)	(2)	(3)	(4)	(5)	(6)
	ROA volatility		ROA		Equity ratio	
Deposit interest rate spread $_{t-1}$	**0.455***	**0.487***	−0.181	−0.188	−0.429	−0.388
	(0.249)	(0.273)	(0.223)	(0.236)	(0.412)	(0.416)
Deposit interest rate spread $_{t-1}$	**0.746****	**0.727****	**−0.617****	**−0.651*****	−0.180	−0.312
× Interest rate hike $_{21Q3-23Q2}$	(0.264)	(0.249)	(0.226)	(0.218)	(0.433)	(0.413)
Deposit interest rate spread $_{t-1}$		−0.021		**−0.205****		**−0.624****
× Interest rate hike $_{17Q4-19Q2}$		(0.219)		(0.084)		(0.242)
Interest rate hike $_{21Q3-23Q2}$	**−0.220*****	**−0.234*****	**0.249*****	**0.285*****	0.135	0.212
	(0.065)	(0.057)	(0.064)	(0.062)	(0.212)	(0.198)
Interest rate hike $_{17Q4-19Q2}$		−0.028		**0.099***		**0.232***
		(0.092)		(0.052)		(0.126)
Ln (Assets) $_{t-1}$	**−0.890****	**−0.884*****	−0.111	−0.160	**−2.580****	**−2.709*****
	(0.316)	(0.282)	(0.419)	(0.418)	(0.930)	(0.897)
Deposit/Debt $_{t-1}$	0.028	0.028	−0.024	−0.025	0.007	0.004
	(0.037)	(0.038)	(0.024)	(0.025)	(0.038)	(0.037)
Loan/Assets $_{t-1}$	−0.003	−0.002	0.009	0.010	**0.067***	**0.072***
	(0.019)	(0.019)	(0.008)	(0.009)	(0.037)	(0.036)
Delinquency Rate $_{t-1}$	**0.208*****	**0.208*****	**−0.144****	**−0.151****	**−0.249*****	**−0.270*****
	(0.057)	(0.051)	(0.055)	(0.054)	(0.062)	(0.060)
Interest Income Share$_{t-1}$	0.004	0.004	0.001	0.0004	0.005	0.005
	(0.006)	(0.006)	(0.004)	(0.004)	(0.006)	(0.005)
Nominal GDP growth $_{t-1}$	**0.020***	**0.023***	−0.013	−0.018	**−0.019****	**−0.028*****
	(0.010)	(0.012)	(0.008)	(0.010)	(0.009)	(0.009)
Inflation $_{t-1}$	−0.052	−0.048	0.037	0.037	−0.020	−0.014
	(0.052)	(0.049)	(0.039)	(0.036)	(0.029)	(0.026)
Credit spread $_{t-1}$	−0.001	−0.001	−0.0002	−0.0004	−0.0004	−0.0006
	(0.001)	(0.001)	(0.0003)	(0.0003)	(0.0004)	(0.0004)
Constant	1.417	1.328	2.411	2.656	13.73**	14.27**
	(3.056)	(3.135)	(2.334)	(2.412)	(4.994)	(4.898)
Institution FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	450	450	450	450	450	450
R-squared	0.284	0.285	0.218	0.222	0.468	0.476

Notes: The table reports the results of a panel fixed-effects analysis investigating the impact of deposit competition on the components of the Z-score during periods of interest rate hikes from 2016.Q1 to 2023.Q2. ROA volatility is defined as the 4-quarter standard deviation of quarterly return on assets. ROA is measured as the 4-quarter moving average of the ratio of quarterly net income to total assets. The Equity ratio is calculated as the 4-quarter moving average of the ratio of quarterly equity to total assets. The Deposit interest rate spread is measured as the difference between the quarterly interest rate on new 1-year term deposits and the 91-day CD yield. Interest rate hike 21Q3-23Q2 is a dummy variable equal to 1 for the period from 2021.Q3 to 2023.Q2, and 0 otherwise. Interest rate hike 17Q4-19Q2 is a dummy variable equal to 1 for the period from 2017.Q4 to 2019.Q2, and 0 otherwise. Ln(Total assets) denotes the natural logarithm of total assets (in trillion won). The Deposit-to-debt and Loan-to-assets ratios serve as proxies for the funding and operational structures of depository institutions, respectively. The Delinquency rate denotes the proportion of loans overdue by more than one month, and the Interest income share is calculated as the ratio of interest income to the sum of interest and non-interest income. Macroeconomic controls include Nominal GDP growth, Inflation (CPI), and the Credit spread, defined as the yield spread between 3-year AA-rated corporate bonds and 3-year government bonds. Standard errors in parentheses are clustered at the institution level. ***$p < .01$, **$p < .05$, *$p < .1$.

primarily through income-based risk channels. This highlights the need for regulatory frameworks that account not only for liquidity and credit conditions, but also for deposit market dynamics and their differential transmission under varying macro-financial environments.

Interest Margins Heterogeneity: Do Interest Margins Matter?

Building on the theoretical discussion in Section 3, interest margin heterogeneity is expected to shape how depository institutions respond to intensified deposit competition. As discussed earlier, while lower-margin institutions are generally considered more vulnerable to funding cost shocks, higher-margin institutions are often assumed to be more resilient. However, this relationship is not deterministic; low-margin institutions

may adopt conservative risk strategies under survival pressure, and high-margin institutions may still engage in risk-taking behavior. These heterogeneous responses call for empirical validation.

To investigate this aspect, we assess whether the impact of deposit competition during periods of financial market instability differs systematically across institutions with different levels of interest margins. Specifically, we divide the sample into two groups—high-margin and low-margin institutions based on the median value of interest margins over the sample period—and estimate the effects separately. This approach allows us to test whether the negative risk effects of competition are more pronounced among structurally low-margin institutions.

The results in Table 5 highlight a significant negative relationship between deposit interest rate spreads and Z-scores, with this relationship being particularly pronounced among institutions with lower interest margins. As shown in Panel A of Table 5, the negative correlation between deposit interest rate spreads and the Z-score is stronger for institutions with lower margins, suggesting that these institutions are more

Table 5. Do interest margins matter?

	(1)	(2)	(3)	(4)
	Institutions with low interest margins		Institutions with high interest margins	
Panel A: Results				
Deposit interest rate spread $_{t-1}$	0.156	0.294	0.187	0.384
	(0.824)	(0.944)	(0.796)	(0.806)
Deposit interest rate spread$_{t-1}$	**−2.707*****	**−2.847*****	−0.306	−0.315
× Interest rate hike $_{21Q3\text{-}23Q2}$	(0.425)	(0.629)	(0.397)	(0.390)
Deposit interest rate spread $_{t-1}$		0.083		0.299
× Interest rate hike $_{17Q4\text{-}19Q2}$		(1.842)		(0.348)
Interest rate hike $_{21Q3\text{-}23Q2}$	**0.964*****	**0.995*****	0.266	0.145
	(0.167)	(0.255)	(0.229)	(0.246)
Interest rate hike $_{17Q4\text{-}19Q2}$		0.063		**−0.300***
		(0.691)		(0.141)
Constant	5.473	5.336	−0.731	−1.692
	(6.451)	(6.360)	(7.116)	(7.288)
Institution Controls	Yes	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	240	240	210	210
R-squared	0.421	0.421	0.464	0.469
	(1) = (3)	(2) = (4)		
Panel B: Tests of the equality				
Deposit interest rate spread $_{t-1}$	0.03	0.08		
Deposit interest rate spread$_{t-1}$	**4.28*****	**3.53****		
× Interest rate hike $_{21Q3\text{-}23Q2}$				
Deposit interest rate spread $_{t-1}$		0.33		
× Interest rate hike $_{17Q4\text{-}19Q2}$				
Interest rate hike $_{21Q3\text{-}23Q2}$	**2.56****	**2.49****		
Interest rate hike $_{171Q4\text{-}19Q2}$		0.53		

Notes: This table reports the results of a panel fixed-effects regression analyzing the impact of deposit competition on the risk of depository institutions, segmented by interest margin levels, during periods of interest rate hikes from 2016.Q1 to 2023.Q2. The dependent variable, Ln(Z-score), is calculated as the natural logarithm of the 4-quarter moving average of the sum of return on assets and the equity ratio, divided by the 4-quarter standard deviation of return on assets. The Deposit interest rate spread is measured as the difference between the quarterly interest rate on new 1-year term deposits and the 91-day CD yield. Interest rate hike 21Q3-23Q2 is a dummy variable equal to 1 for the period from 2021.Q3 to 2023.Q2, and 0 otherwise. Interest rate hike 17Q4-19Q2 is a dummy variable equal to 1 for the period from 2017.Q4 to 2019.Q2, and 0 otherwise. Institution Controls cover Ln(Total assets), Deposit-to-debt, Loan-to-assets, Delinquency rate, and Interest income share. Ln(Total assets) denotes the natural logarithm of total assets (in trillion won). The Deposit-to-debt and Loan-to-assets ratios serve as proxies for the funding and operational structures of depository institutions, respectively. The Delinquency rate denotes the proportion of loans overdue by more than one month, and the Interest income share is calculated as the ratio of interest income to the sum of interest and non-interest income. Macroeconomic controls include Nominal GDP growth, Inflation (CPI), and the Credit spread, defined as the yield spread between 3-year AA-rated corporate bonds and 3-year government bonds. Standard errors in parentheses are clustered at the institution level. ***$p < .01$, **$p < .05$, *$p < .1$.

vulnerable to rising deposit costs. This finding aligns with the theoretical predictions of Busch and Memmel (2015) and Ho and Saunders (1981), who documented that banks with smaller net interest margins exhibit greater sensitivity to funding cost fluctuations, particularly in highly competitive environments. The heightened vulnerability of low-margin institutions is further evidenced by Panel B, which shows a statistically significant difference in the response of institutions with low versus high interest margins during the recent period of monetary tightening. Specifically, institutions with narrower interest margins experience a more pronounced deterioration in financial stability, reinforcing previous evidence that interest margin constraints intensify the destabilizing effects of deposit competition during periods of rising interest rates (Altavilla, Boucinha, and Peydró 2018; Beck, Demirgüç-Kunt, and Levine 2013; Berger, Klapper, and Turk-Ariss 2009).

The findings imply that, unlike institutions with greater pricing power in loan markets, institutions with constrained interest margins lack the flexibility to offset rising funding costs, making them more susceptible to financial distress. These results highlight the importance of interest margin heterogeneity in shaping institutional resilience, underscoring the need for targeted risk management strategies to mitigate the adverse effects of competitive funding pressures during periods of financial market instability.

To further examine the underlying mechanisms through which heightened deposit competition transmits risk to low-margin institutions, we extend our analysis by performing a Z-score decomposition, focusing on the level of interest margins. As shown in Table 6, the results confirm that sharp increases in deposit interest rates during the recent tightening period lead to greater earnings volatility, reduced profitability, and weaker capitalization among institutions with lower margins. Importantly, the negative and statistically significant effect on equity ratios further highlights the erosion of capital buffers, suggesting that

Table 6. Z-score decomposition for institutions with low interest margins.

	(1)	(2)	(3)	(4)	(5)	(6)
	ROA Volatility		ROA		Equity Ratio	
Deposit interest rate spread $_{t-1}$	0.154	0.135	−0.149	−0.175	0.933	0.713
	(0.654)	(0.866)	(0.317)	(0.456)	(0.652)	(0.562)
Deposit interest rate spread$_{t-1}$	**2.174*****	**2.210*****	**−1.418*****	**−1.410*****	**−2.402*****	**−2.117*****
× Interest rate hike $_{21Q3\text{-}23Q2}$	(0.433)	(0.364)	(0.328)	(0.255)	(0.426)	(0.482)
Deposit interest rate spread $_{t-1}$		0.164		−0.074		0.968
× Interest rate hike $_{17Q4\text{-}19Q2}$		(0.542)		(0.437)		(0.767)
Interest rate hike $_{21Q3\text{-}23Q2}$	**−0.636*****	**−0.660*****	**0.398*****	**0.414*****	**0.961*****	**0.837*****
	(0.173)	(0.170)	(0.108)	(0.091)	(0.219)	(0.205)
Interest rate hike $_{17Q4\text{-}19Q2}$		−0.064		0.051		−0.309
		(0.164)		(0.124)		(0.220)
Constant	4.312	4.310	−2.012	−1.964	22.68***	22.82***
	(6.869)	(6.964)	(3.217)	(3.371)	(5.820)	(5.596)
Institution Controls	Yes	Yes	Yes	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	240	240	240	240	240	240
R-squared	0.428	0.428	0.551	0.551	0.879	0.879

Notes: The table reports the results of a panel fixed-effects analysis investigating the impact of deposit competition with low interest margins on the components of the Z-score during periods of interest rate hikes from 2016.Q1 to 2023.Q2. ROA volatility is defined as the 4-quarter standard deviation of quarterly return on assets. ROA is measured as the 4-quarter moving average of the ratio of quarterly net income to total assets. The Equity ratio is calculated as the 4-quarter moving average of the ratio of quarterly equity to total assets. The Deposit interest rate spread is measured as the difference between the quarterly interest rate on new 1-year term deposits and the 91-day CD yield. Interest rate hike 21Q3-23Q2 is a dummy variable equal to 1 for the period from 2021.Q3 to 2023.Q2, and 0 otherwise. Interest rate hike 17Q4-19Q2 is a dummy variable equal to 1 for the period from 2017.Q4 to 2019.Q2, and 0 otherwise. Institution Controls cover Ln(Total assets), Deposit-to-debt, Loan-to-assets, Delinquency rate, and Interest income share. Ln(Total assets) denotes the natural logarithm of total assets (in trillion won). The Deposit-to-debt and Loan-to-assets ratios serve as proxies for the funding and operational structures of depository institutions, respectively. The Delinquency rate denotes the proportion of loans overdue by more than one month, and the Interest income share is calculated as the ratio of interest income to the sum of interest and non-interest income. Macroeconomic controls include Nominal GDP growth, Inflation (CPI), and the Credit spread, defined as the yield spread between 3-year AA-rated corporate bonds and 3-year government bonds. Standard errors in parentheses are clustered at the institution level. ***$p < .01$, **$p < .05$, *$p < .1$.

these institutions are less capable of absorbing financial shocks in periods of aggressive deposit competition and financial market instability.

The findings in Table 6 support the argument that institutions with lower interest margins are significantly more vulnerable to sharp increases in deposit interest rates, leading to heightened financial instability and deteriorating profitability. This vulnerability stems from their limited ability to pass on higher funding costs through loan repricing, a phenomenon well-documented in prior research on bank fragility in competitive environments (Albertazzi and Gambacorta 2009; Ho and Saunders 1981). Overall, the findings imply that deposit competition can be a source of institutional risks, particularly when interest rates rise sharply, as institutions with narrower margins lack the flexibility to absorb funding shocks without destabilizing their balance sheets.

Institutional Balance Sheet Heterogeneity

Our baseline findings suggest that intensified deposit competition undermines institutional soundness by reducing profitability and amplifying earnings volatility under conditions of financial market stress. Particularly, institutions with narrow interest margins appear more vulnerable, exhibiting not only deeper margin compression but also declines in capital adequacy. While these results provide strong support for the competition-fragility hypothesis, they also raise a critical question: are all institutions equally exposed to funding competition, or do structural differences in balance sheet composition alter the transmission of such competitive pressure?

To explore this possibility, we assess whether the relationship between deposit competition and institutional risk varies across four key institutional characteristics: loan-to-asset ratios, deposit-to-debt ratios, asset quality as proxied by non-performing loan ratios, and capital adequacy. These dimensions are theoretically motivated and frequently cited in the literature as potential determinants of institutional resilience to funding stress (Valverde and Fernández 2007). For instance, banks with high loan concentration may face increased margin compression, given the rigidity of illiquid and slow-to-reprice loan portfolios (Altavilla, Boucinha, and Peydró 2018). Similarly, institutions that rely heavily on deposit funding may be more exposed to liquidity pressures in tightening cycles, as rate-sensitive depositors become more prone to shifting funds (Kashyap, Rajan, and Stein 2002). Also, weak asset quality and insufficient capital buffers may constrain an institution's capacity to absorb earnings shocks and maintain solvency (Berger and Bouwman 2013; Delis and Kouretas 2011). To empirically examine these hypotheses, we conduct subsample regressions along each of the four balance sheet dimensions. The sample is partitioned at the median of each variable, and the baseline model is re-estimated for each subgroup. Analyzing these structural features allows us to evaluate the extent to which institutional heterogeneity conditions the impact of intensified deposit competition on institutional risk.

Interestingly, as reported in Table 7, across all four characteristics—loan concentration, deposit reliance, asset quality, and capital adequacy—we find no statistically significant difference in the effect of deposit competition on institutional risk. That is, the adverse impact of deposit competition on Z-scores appears broadly uniform across institutions, regardless of their loan structure, funding profile, credit risk, or capital position. These findings suggest that, contrary to conventional expectations, balance sheet indicators do not mediate the transmission of funding pressure into institutional fragility.

Importantly, these findings reinforce a key insight: interest margin compression—rather than traditional balance sheet indicators—is the primary channel through which deposit competition undermines institutional soundness. Ultimately, it is important for depository institutions to preserve margin buffers—not merely its structural profile—that determines its resilience to deposit competition. Recognizing this channel is essential for effective risk oversight in a persistently competitive funding environment.

Conclusion and Policy Implications

This paper provides new empirical evidence on the relationship between deposit competition and depository institutional risk, with a particular focus on how these dynamics unfold under financial

Table 7. Heterogeneity analysis by balance sheet structure.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Loan/Asset		Deposit/Debt		Equity/Asset		NPL/Loan	
	High	Low	High	Low	High	Low	High	Low
Deposit interest rate spread $_{t-1}$	−0.593	**−1.266***	−0.165	−0.655	−0.310	−0.111	−0.264	**−3.230****
	(0.842)	(0.670)	(0.813)	(1.291)	(0.511)	(0.703)	(0.776)	(1.411)
Deposit interest rate spread$_{t-1}$	−0.299	−1.022	−0.082	−1.940	−0.903	−0.373	−0.559	−1.011
× Interest rate hike $_{21Q3\text{-}23Q2}$	(0.236)	(0.634)	(0.414)	(1.288)	(0.575)	(0.830)	(0.626)	(1.011)
Deposit interest rate spread $_{t-1}$	0.619	**−1.683***	0.206	0.936	0.158	−0.296	0.210	1.889
× Interest rate hike $_{17Q4\text{-}19Q2}$	(0.363)	(0.834)	(0.342)	(2.212)	(0.679)	(0.790)	(0.639)	(1.714)
1.d_interest_2103_2302	0.111	**0.556***	0.012	**0.792***	0.292	**0.397***	0.559	0.437
	(0.139)	(0.286)	(0.179)	(0.405)	(0.398)	(0.202)	(0.573)	(0.291)
1.d_interest_174_192	−0.310	0.708	−0.196	−0.243	−0.169	0.090	−0.164	−0.347
	(0.215)	(0.510)	(0.157)	(0.718)	(0.297)	(0.387)	(0.357)	(0.474)
Constant	7.409	−1.698	3.152	−1.978	3.376	15.89*	−0.570	13.78**
	(4.741)	(5.354)	(6.598)	(6.975)	(7.213)	(8.650)	(7.011)	(4.914)
Institution Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	225	225	225	225	225	225	225	225
R-squared	0.537	0.448	0.576	0.358	0.475	0.359	0.505	0.513

Notes: Table 7 reports the results of heterogeneity analyses based on institutional balance sheet characteristics, including the deposit-to-debt ratio, loan-to-asset ratio, equity ratio, and non-performing loan (NPL) ratio, over the period from 2016.Q1 to 2023.Q2. The dependent variable, Ln(Z-score), is calculated as the natural logarithm of the 4-quarter moving average of the sum of return on assets and the equity ratio, divided by the 4-quarter standard deviation of return on assets. The Deposit interest rate spread is measured as the difference between the quarterly interest rate on new 1-year term deposits and the 91-day CD yield. Interest rate hike 21Q3-23Q2 is a dummy variable equal to 1 for the period from 2021.Q3 to 2023.Q2, and 0 otherwise. Interest rate hike 17Q4-19Q2 is a dummy variable equal to 1 for the period from 2017.Q4 to 2019.Q2, and 0 otherwise. Institution Controls cover Ln(Total assets), Deposit-to-debt, Loan-to-assets, Delinquency rate, and Interest income share. Ln(Total assets) denotes the natural logarithm of total assets (in trillion won). The Deposit-to-debt and Loan-to-assets ratios serve as proxies for the funding and operational structures of depository institutions, respectively. The Delinquency rate denotes the proportion of loans overdue by more than one month, and the Interest income share is calculated as the ratio of interest income to the sum of interest and non-interest income. Macroeconomic controls include Nominal GDP growth, Inflation (CPI), and the Credit spread, defined as the yield spread between 3-year AA-rated corporate bonds and 3-year government bonds. Standard errors in parentheses are clustered at the institution level. ***$p < .01$, **$p < .05$, *$p < .1$.

stress in an emerging market context. Drawing on institution-level data from South Korea, we find that intensified deposit competition significantly undermines financial stability by compressing profitability and amplifying earnings volatility. These adverse effects are particularly concentrated among institutions with narrower interest margins, which are structurally less capable of absorbing funding shocks. Notably, the destabilizing effects are salient during the 2021–2023 monetary tightening cycle, a period characterized by market-wide liquidity strain and aggressive deposit competition, while remaining muted during the earlier 2017–2019 monetary tightening cycle.

By jointly testing the competition—fragility and competition—stability hypotheses, this study contributes to the literature by demonstrating that the effects of deposit competition on institutional risk are conditional rather than uniform. Rather than enhancing resilience or efficiency, heightened deposit competition during periods of financial stress appears to amplify fragility, especially for institutions with limited pricing power. These findings highlight the need to consider both systemic market conditions and institution-specific characteristics when evaluating the financial stability implications of funding competition.

The results offer several actionable policy implications. First, regulatory authorities should adopt a system-wide perspective in monitoring the risks associated with deposit competition. Competitive pressures in funding markets can serve as conduits for liquidity shocks, particularly during monetary tightening cycles, raising the risk of contagion. Integrating deposit competition metrics into macroprudential surveillance and liquidity stress-testing frameworks would strengthen early warning systems and risk mitigation strategies. Second, financial institutions, especially those operating with structurally low interest

margins, must strengthen liquidity management frameworks and reduce their overreliance on aggressive deposit pricing strategies. While raising deposit rates may temporarily secure funding, it can simultaneously elevate income volatility and erode capital buffers. A more prudent approach would emphasize pricing discipline, diversification of funding sources, and forward-looking liquidity planning. Third, although the impact of digital banks on deposit competition remains limited in South Korea, with internet-only banks accounting for less than 4% of total deposits as of end-2023, this landscape is likely to evolve. As digital financial platforms continue to expand, future research should investigate how these emerging players further influence competitive dynamics and financial risk, particularly in emerging markets where institutional buffers may be weaker.

Overall, this study advances our understanding of how deposit market behavior and macro-financial conditions jointly shape the resilience of financial institutions. The evidence presented offers both analytical depth and practical insight for regulators and market participants navigating the complex trade-offs between competition, profitability, and systemic stability.

Notes

1. In the second half of 2022, South Korea experienced significant financial market instability, commonly referred to as the "Legoland Incident," which had far-reaching implications for the country's bond and short-term money markets. The crisis was triggered when the Gangwon Province government unexpectedly announced its intention to default on guaranteed debt related to a developer involved in the Legoland Korea project. This event occurring in September 2022 caught investors by surprise, as government-backed debt was widely perceived as low-risk. The announcement led to a sharp increase in risk aversion, causing a freeze in the corporate bond and commercial paper markets, particularly affecting lower-rated issuers.
2. Amid escalating financial uncertainty and impaired short-term funding markets, depository institutions, including both commercial banks and NBDIs, faced mounting pressure to secure stable liquidity. In this environment of heightened risk aversion and credit market dysfunction, competition for deposits intensified markedly, as institutions raised deposit interest rates to defend their funding bases. This deposit war reflected a strategic response to financial market instability, with institutions seeking to bolster liquidity buffers in anticipation of further funding stress.
3. However, additional tests indicate that other institutional characteristics, such as loan ratios, deposit reliance, asset quality, or capital adequacy, do not exhibit comparable heterogeneity in the relationship between deposit competition and institutional risk.
4. In response to the mounting financial uncertainty, referred to as the "Legoland Incident" explained in Section 1, depository institutions, including commercial banks and NBDIs, came under increasing pressure to secure liquidity. This environment of heightened risk aversion and frozen credit markets prompted intense competition for deposits, as institutions significantly raised deposit interest rates to attract and retain funding. The surge in deposit rates was a direct consequence of the financial instability, reflecting institutional efforts to fortify liquidity positions under tightening funding conditions.
5. The average quarterly spread between commercial and regional banks remained at 1 basis point from early 2014 to Q1 2021, but widened to 15 basis points from late 2021 through mid-2023.
6. By the end of Q2 2023, deposits accounted for 74.3% of total funding across all South Korean depository institutions. Among NBDIs, this ratio reached 86.4%, underscoring their structural reliance on deposit-based funding for their lending activities.
7. The Legoland Incident exemplifies how localized credit risk events can propagate broader market disruptions, intensifying competition for deposits as institutions scramble to secure liquidity amid funding stress. This case highlights the critical link between aggressive deposit pricing and financial instability, highlighting the importance of effective liquidity risk management during periods of systemic stress.
8. Interest rate pass-through mechanisms offer an important channel through which institutions attempt to defend interest margins under deposit competition. In practice, however, the ability to raise lending rates is often constrained in highly competitive or saturated loan markets, especially under monetary tightening. When pass-through is limited, margin compression intensifies, amplifying earnings volatility and risk exposure (Albertazzi and Gambacorta 2009; Ho and Saunders 1981; Van Leuvensteijn et al. 2013). Accordingly, pass-through efficiency, closely tied to pricing power, also may condition how interest margin levels interact with competition to shape institutional fragility.
9. While institution-level 1-year term deposit interest rate data for individual savings banks are not available, aggregate data for the mutual savings bank sector, comprising all 79 savings banks, can be obtained. We, thus, treat this sector as a single representative entity in our analysis. This approach allows us to capture the sector's

collective pricing behavior while addressing data limitations, thereby preserving comparability and consistency across different types of depository institutions.

10. Specialized banks in South Korea typically operate under unique government directives or policy-driven objectives, which can lead to differing risk and return profiles.
11. We exclude the internet-only banks from our sample for both conceptual and empirical reasons. Although the emergence of digital banks has contributed to intensifying deposit competition, their systemic relevance in South Korea remains limited. As of end-2023, internet-only banks accounted for 3.6% of total deposits, and their market share remained low even during the liquidity stress episode in 2022. Also, reliable institution-level data on their 1-year term deposit interest rates are not available in the survey dataset, which limits our ability to construct a consistent time-series.
12. During this period, the Bank of Korea raised its policy rate from 1.25% to 1.75% in two steps, first in November 2017 (1.25% to 1.50%), and then in November 2018 (1.50% to 1.75%). The gradual pace of tightening resulted in a mild increase in deposit competition, giving depository institutions more time to adjust to rising funding costs.
13. The Bank of Korea raised its policy rate from 0.50% in August 2021 to 3.50% by January 2023 in response to rising inflationary pressures. This sharp and sustained tightening led depositors to actively seek higher-yielding deposit products, intensifying deposit competition across institutions. The liquidity shock triggered by the "Legoland Incident" in late 2022 further destabilized market conditions, pushing both depositors and institutions toward more aggressive deposit behavior. The convergence of rapid rate hikes and financial market instability significantly amplified funding pressures.
14. The distress index is a widely used proxy variable that quantifies the distance of a bank from insolvency, serving as a key indicator of the bank's financial stability in relation to its potential risk of bankruptcy.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Jaeweon You http://orcid.org/0009-0000-0936-2890
Beomseok Seo http://orcid.org/0009-0009-2986-3567

Data Availability Statement

The data that support the findings of this study were provided by the Bank of Korea and are subject to confidentiality agreements. The data are not publicly available.

References

Albertazzi, U., and L. Gambacorta. 2009. Bank profitability and the business cycle. *Journal of Financial Stability* 5 (4):393–409. doi: 10.1016/j.jfs.2008.10.002.

Altavilla, C., M. Boucinha, and J. L. Peydró. 2018. Monetary policy and bank profitability in a low interest rate environment. *Economic Policy* 33 (96):531–86. doi: 10.1093/epolic/eiy013.

Anginer, D., A. Demirgüç-Kunt, and M. Zhu. 2014. How does competition affect bank systemic risk? *Journal of Financial Intermediation* 23 (1):1–26. doi: 10.1016/j.jfi.2013.11.001.

Arping, S. 2017. Deposit competition and loan markets. *Journal of Banking & Finance* 80:108–18. doi: 10.1016/j.jbankfin.2017.04.006.

Beck, T., A. Demirgüç-Kunt, and R. Levine. 2013. Financial institutions and markets across countries and over time: The updated financial development and structure database. *The World Bank Economic Review* 24 (1):77–92. doi: 10.1093/wber/lhp016.

Berger, A. N., and C. H. S. Bouwman. 2013. How does capital affect bank performance during financial crises? *Journal of Financial Economics* 109 (1):146–76. doi: 10.1016/j.jfineco.2013.02.008.

Berger, A. N., L. F. Klapper, and R. Turk-Ariss. 2009. Bank competition and financial stability. Bank competition and financial stability. *Journal of Financial Service Research* 35 (2):99–118. doi: 10.1007/s10693-008-0050-7.

Bikker, J. A., and T. M. Vervliet. 2018. Bank profitability and risk-taking under low interest rates. *International Journal of Finance & Economics* 23 (1):3–18. doi: 10.1002/ijfe.1595.

Boone, A. L., L. C. Field, J. M. Karpoff, and C. G. Raheja. 2007. The determinants of corporate board size and composition: An empirical analysis. *Journal of Financial Economics* 85 (1):66–101. doi: 10.1016/j.jfineco.2006.05.004.

Borio, C., L. Gambacorta, and B. Hofmann. 2017. The influence of monetary policy on bank profitability. *International Finance* 20 (1):48–63. doi: 10.1111/infi.12104.

Boyd, J. H., and G. De Nicolo. 2005. The theory of bank risk-taking and competition revisited. *Journal of Finance* 60 (3):1329–43. doi: 10.1111/j.1540-6261.2005.00763.x.

Busch, R., and C. Memmel. 2015. Banks' net interest margin and the level of interest rates (no. 16/2015). *Deutsche Bundesbank, Research Centre*.

Chortareas, G. E., J. G. Garza-García, and C. Girardone. 2012. Competition, efficiency, and interest rate margins in Latin American banking. *International Review of Financial Analysis* 24:93–103. doi: 10.1016/j.irfa.2012.08.006.

Claessens, S., N. Coleman, and M. Donnelly. 2018. Low-for-long interest rates and banks' interest margins and profitability: Cross-country evidence. *Journal of Financial Intermediation* 35:1–16. doi: 10.1016/j.jfi.2017.05.004.

Claessens, S., and L. Laeven. 2004. What drives bank competition? Some international evidence. *Journal of Money, credit, and Banking* 36 (3):563–83. doi: 10.1353/mcb.2004.0044.

Craig, B. R., and V. Dinger. 2013. Deposit market competition, wholesale funding, and bank risk. *Journal of Banking & Finance* 37 (9):3605–22. doi: 10.1016/j.jbankfin.2013.05.010.

Delis, M. D., and G. P. Kouretas. 2011. Interest rates and bank risk-taking. *Journal of Banking & Finance* 35 (4):840–55. doi: 10.1016/j.jbankfin.2010.09.032.

Demirgüç-Kunt, A., and H. Huizinga. 2010. Bank activity and funding strategies: The impact on risk and returns. *Journal of Financial Economics* 98 (3):626–50. doi: 10.1016/j.jfineco.2010.06.004.

Egan, M., A. Hortaçsu, and G. Matvos. 2017. Deposit competition and financial fragility: Evidence from the us banking sector. *American Economic Review* 107 (1):169–216. doi: 10.1257/aer.20150342.

El-Chaarani, H., R. Abraham, and G. Azzi. 2023. The role of liquidity creation in managing the COVID-19 banking crisis in selected MENA countries. *International Journal of Financial Studies* 11 (1):39. doi: 10.3390/ijfs11010039.

Faleye, O. 2015. The costs of a (nearly) fully independent board. *Journal of Empirical Finance* 32:49–62. doi: 10.1016/j.jempfin.2014.12.001.

Faleye, O., T. Kovacs, and A. Venkateswaran. 2014. Do better-connected CEOs innovate more? *Journal of Financial and Quantitative Analysis* 49 (5–6):1201–25. doi: 10.1017/S0022109014000714 .

Fu, X., Y. Lin, and P. Molyneux. 2014. Bank competition and financial stability in Asia Pacific. *Journal of Banking & Finance* 38:64–77. doi: 10.1016/j.jbankfin.2013.09.012.

Ho, T. S. Y., and A. Saunders. 1981. The determinants of bank interest margins: Theory and empirical evidence. *Journal of Financial and Quantitative Analysis* 16 (4):581–600. doi: 10.2307/2330377.

Jiménez, G., S. Ongena, J. L. Peydró, and J. Saurina. 2014. Hazardous times for monetary policy: What do twenty-three million bank loans say about the effects of monetary policy on credit risk-taking? *Econometrica* 82:463–505.

Kashyap, A. K., R. Rajan, and J. C. Stein. 2002. Banks as liquidity providers: An explanation for the coexistence of lending and deposit-taking. *The Journal of Finance* 57 (1):33–73. doi: 10.1111/1540-6261.00415.

Keeley, M. C. 1990. Deposit insurance, risk, and market power in banking. *American Economic Review* 80 (5):1183–200.

Laliberte, B., and R. Sengupta. 2024. Why do net interest margins behave differently across banks as interest rates rise? *Federal Reserve Bank of Kansas City Economic Review* 109 (1). doi:10.18651/ER/v109n1LaliberteSengupta.

Lepetit, L., and F. Strobel. 2015. Bank insolvency risk and Z-score measures: A refinement. *Finance Research Letters* 13:214–24. doi: 10.1016/j.frl.2015.01.001.

Martinez-Miera, D., and R. Repullo. 2010. Does competition reduce the risk of bank failure? *The Review of Financial Studies* 23 (10):3638–64. doi: 10.1093/rfs/hhq057.

Vallascas, F., S. Mollah, and K. Keasey. 2017. Does the impact of board independence on large bank risks change after the global financial crisis? *Journal of Corporate Finance* 44:149–66. doi: 10.1016/j.jcorpfin.2017.03.011.

Valverde, S. C., and F. R. Fernández. 2007. The determinants of bank margins in European banking. *Journal of Banking & Finance* 31 (7):2043–63. doi: 10.1016/j.jbankfin.2006.06.017.

Van Leuvensteijn, M., C. K. Sørensen, J. A. Bikker, and A. A. R. J. M. Van Rixtel. 2013. Impact of bank competition on the interest rate pass-through in the euro area. *Applied Economics* 45 (11):1359–80. doi: 10.1080/00036846.2011.617697.

Vives, X. 2016. Competition and stability in banking: The role of regulation and competition policy. In *Competition and stability in banking*. Princeton University Press.

Appendix

Table A1. Alternative specifications.

Panel A: Random Effect Model	(1) Ln (Z-score)	(2) Ln (Z-score)	(3) Ln (Z-score)
Deposit interest rate spread $_{t-1}$	−0.751	−0.270	−0.181
	(0.981)	(0.728)	(0.715)
Deposit interest rate spread $_{t-1}$ × Interest rate hike $_{21Q3\text{-}23Q2}$		**−0.849****	**−0.881****
		(0.393)	(0.410)
Deposit interest rate spread $_{t-1}$ × Interest rate hike $_{17Q4\text{-}19Q2}$	0.307		−0.025
	(0.499)		(0.517)
Interest rate hike $_{21Q3\text{-}23Q2}$		**0.406*****	**0.372***
		(0.147)	(0.193)
Interest rate hike $_{17Q4\text{-}19Q2}$	−0.142		−0.068
	(0.236)		(0.256)
Bank Controls	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes
Bank FE	No	No	No
Year FE	Yes	Yes	Yes
Observations	450	450	450
R-squared	0.110	0.110	0.115
Panel B: Fixed Effect Models with Two Lags	**(1) Ln (Z-score)**	**(2) Ln (Z-score)**	**(3) Ln (Z-score)**
Deposit interest rate spread $_{t-2}$	**−1.233***	−0.738	−0.656
	(0.674)	(0.555)	(0.558)
Deposit interest rate spread $_{t-2}$ × Interest rate hike $_{21Q3\text{-}23Q2}$		**−0.605***	**−0.689***
		(0.335)	(0.371)
Deposit interest rate spread $_{t-2}$ × Interest rate hike $_{17Q4\text{-}19Q2}$	−0.039		−0.248
	(0.468)		(0.514)
Interest rate hike $_{21Q3\text{-}23Q2}$		0.261	0.268
		(0.165)	(0.183)
Interest rate hike $_{17Q4\text{-}19Q2}$	−0.022		0.024
	(0.222)		(0.235)
Bank Controls	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	450	450	450
R-squared	0.183	0.189	0.190
Panel C: Instrumental Variable Analysis	**(1) Ln (Z-score)**	**(2) Ln (Z-score)**	**(3) Ln (Z-score)**
Deposit interest rate spread $_{t-1}$	**−1.374****	−0.622	−0.498
	(0.649)	(0.401)	(0.438)
Deposit interest rate spread $_{t-1}$ × Interest rate hike $_{21Q3\text{-}23Q2}$		**−0.945****	**−1.040*****
		(0.368)	(0.393)
Deposit interest rate spread $_{t-1}$ × Interest rate hike $_{17Q4\text{-}19Q2}$	0.072		−0.194
	(0.451)		(0.475)
Interest rate hike $_{21Q3\text{-}23Q2}$		0.273	**0.301***
		(0.137)	(0.156)
Interest rate hike $_{17Q4\text{-}19Q2}$	−0.029		−0.018
	(0.241)		(0.244)
Bank Controls	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Kleibergen-Paap Under-identification test (p=value)	0.005	0.005	0.009
Hansen J test (p-value)	0.843	0.470	0.547
Observations	450	450	450
R-squared	0.109	0.124	0.125

Notes: This table reports the results of a panel analysis with a series of alternative specifications to examine the impact of deposit competition on the risk of depository institutions during periods of interest rate hikes from 2016.Q1 to 2023.Q2. Panel A presents the results from random effects models, while Panel B reports fixed effects regressions incorporating two lags of the bank-level explanatory variables. Panel C displays the results of an instrumental variable (IV) fixed effects estimation. Standard errors in parentheses are clustered at the institution level. ***$p < .01$, **$p < .05$, *$p < .1$.

Finance Research Letters 86 (2025) 108616

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Who pays when zombie firms Persist? Asymmetric debt conditions by credit rating

Jaeweon You [a], Beomseok Seo [b,*], Junyoung An [c]

[a] Toss insight, 62 Sowol-ro, Yongsan-gu, Seoul, Republic of Korea
[b] Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, Republic of Korea
[c] Bank of Korea, 39 Namdaemun-ro, Jung-gu, Seoul, Republic of Korea

ARTICLE INFO

JEL Classification:
G21
G32
L25
E44

Keywords:
Zombie firms
Credit misallocation
Borrowing costs
Credit ratings
Financial stability

ABSTRACT

We examine how zombie firm prevalence relates to debt financing conditions for viable firms in a bank-based economy. Using firm-level panel data from South Korea, we find that a higher share of zombie firms within an industry is associated with higher borrowing costs for non-zombie firms, without a significant reduction in their debt growth on average. However, the relationship is highly asymmetric; low-rated non-zombie firms exhibit both elevated interest rates and restricted credit access. Moreover, the adverse effects of zombie prevalence are more pronounced in the service sector than in manufacturing and during periods of monetary easing than tightening. This pattern reflects the heightened sensitivity of such contexts to macro-financial conditions. Our findings highlight credit policy frameworks that incorporate firm-level risk sensitivity and context-dependent amplification.

1. Introduction

The rising prevalence of zombie firms, financially non-viable enterprises sustained by persistent debt rollovers, has emerged as a major concern for macro-financial stability. By absorbing labor and capital that could be deployed more productively, zombie firms hinder creative destruction, distort resource allocation, and suppress aggregate productivity (Caballero et al., 2008; McGowan et al., 2018; Banerjee and Hofmann, 2022; Acharya et al., 2022). Their survival is often enabled by forbearance lending, as creditors defer loss recognition to protect their balance sheets, compounding systemic vulnerabilities (Albuquerque and Iyer, 2024).

While macro-level consequences of zombie firms have been widely studied (e.g., Caballero et al., 2008), less is known about their micro-level effects on viable firms, particularly in the form of distorted debt financing conditions. Recent research highlights how zombie prevalence may elevate perceived industry-level credit risk, tightening financing conditions even for fundamentally sound firms (Albuquerque and Iyer, 2024; Havemeister and Horn, 2023). These effects may be especially pronounced in bank-based economies, such as South Korea, where relational lending dominates credit allocation and amplifies sector-wide contagion.

Nonetheless, important empirical gaps remain. First, most existing studies focus on aggregate productivity or investment crowding-out, overlooking how zombie firms reshape the credit environment for healthy borrowers. Second, the heterogeneity of these aspects is underexplored, specifically, whether financially weaker firms or those in credit-dependent sectors face disproportionate burdens (Chodorow-Reich, 2014; Dell'Ariccia et al., 2008; Hadlock and Pierce, 2010). Third, the macro-financial context in which these

* Corresponding author.
E-mail address: bsseo@sookmyung.ac.kr (B. Seo).

https://doi.org/10.1016/j.frl.2025.108616
Received 5 July 2025; Received in revised form 28 September 2025; Accepted 30 September 2025
Available online 1 October 2025
1544-6123/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

relationships occur remains ambiguous. While some argue that tighter monetary policy may indirectly prolong zombie survival via evergreening incentives, others emphasize that accommodative conditions more directly foster zombie persistence by weakening market discipline and reducing restructuring pressures (Albuquerque and Mao, 2023; Acharya et al., 2022; Banerjee and Hofmann, 2022).

Understanding how zombie firm prevalence is associated with debt conditions of viable firms is critical, as rising credit costs and financing constraints for productive firms can dampen investment, innovation, and ultimately economic recovery. Identifying which firms are most vulnerable to these negative nexuses is therefore essential for designing targeted financial and industrial policies. This study addresses these gaps by examining how zombie firm prevalence influences the borrowing costs and debt growth of viable firms, using granular firm-level panel data from South Korea. We find that zombie-heavy industries are associated with significantly higher borrowing costs for non-zombie firms, without an overall decline in debt volume. However, among low-rated firms, zombie exposure is linked to both costlier borrowing and curtailed debt growth, reflecting dual financial constraints. These patterns are particularly acute in the service sector and during periods of monetary easing.

Our contributions are threefold. First, we uncover a firm-level risk-pricing channel linking zombie prevalence to higher financing costs for viable borrowers. Second, the study demonstrates substantial heterogeneity in these patterns across credit ratings and industries. We also show that accommodative monetary policy exacerbates these distortions. These findings shift the focus from aggregate inefficiencies to distributional effects in credit markets and highlight the need for policy frameworks sensitive to both firm-level vulnerabilities and macro-financial regimes. Specifically, practical measures include integrating zombie-prevalence indicators into supervisory monitoring and stress tests, applying sectoral capital requirements when zombie concentrations are high, and calibrating loan-classification and support programs to protect viable but vulnerable firms without subsidizing persistently unviable ones.

2. Research design

2.1. Data

We construct an annual firm-level panel dataset using the Value Search database, which compiles standardized financial statements, credit ratings, and regulatory filings for South Korean firms. The sample includes all non-financial firms, both listed and unlisted, subject to external audits. The inclusion of privately held firms is essential in a bank-based financial system, such as South Korea, where many credit-reliant firms operate outside capital markets. The panel spans 2013 to 2022, a period selected to align with the implementation of a standardized ten-tier credit rating system, ensuring consistent comparability of credit risk across time. Importantly, to mitigate survivorship bias (Albuquerque and Iyer, 2024), we restrict the sample to firms continuously active throughout the period, yielding a balanced panel of 13,748 firms, resulting in 137,480 firm-year observations.

Importantly, we define zombie firms as those whose interest coverage ratio (EBIT/interest expenses) remains below one for three consecutive years, following El Ghoul et al. (2021) and McGowan et al. (2018). This criterion targets firms with a persistent inability to service debt from operating income, distinguishing structural unviability from temporary distress. To further refine identification, we exclude firms younger than ten years, given their typically volatile earnings and unstable capital structures. This definition is consistent with frameworks adopted by the Bank of Korea and OECD, lending both theoretical and policy relevance to our classification.

2.2. Empirical strategy

To estimate the effect of zombie firm concentration on borrowing conditions for healthy firms, we employ a panel fixed effects regression model.

$$Y_{i,n,t} = \beta_1 NonZombie_{i,n,t-1} + \beta_2 \left(NonZombie_{i,n,t-1} \times ZombieShare_{n,t-1}\right) + \beta_3 X_{i,n,t-1} + \alpha_i + \delta_{n,t} + \epsilon_{i,n,t}$$

$\mathbf{Y_{i,n,t}}$ denotes the outcome variable for firm i in industry n and year t, either the firm's debt growth rate (year-on-year change in total debt) or its average interest rate (interest expenses divided by the average of total debt at the beginning and end of year). The binary indicator $\mathbf{NonZombie_{i,n,t-1}}$ identifies firms classified as financially healthy, while $\mathbf{Zombie\ Share_{n,t-1}}$ is defined as the ratio of total debt held by zombie firms to total debt of all firms within industry n at time t, constructed at the two-digit KSIC level.

Next, we control for time-varying firm characteristics through the vector $\mathbf{X_{i,n,t-1}}$, which includes: log total assets (firm size), current ratio (liquidity), total debt-to-total capital (leverage), EBIT over total assets (profitability), ICR (EBIT/interest expenses), and firm age. Firm fixed effects ($\boldsymbol{\alpha_i}$) absorb time-invariant heterogeneity across firms, such as persistent differences in managerial quality or risk profiles. Industry–year fixed effects ($\boldsymbol{\delta_{n,t}}$) absorb contemporaneous shocks common to all firms within the same sector and year: for example, changes in sectoral demand conditions or credit supply shifts that might otherwise confound the relationship between zombie prevalence and debt financing conditions. To mitigate simultaneity concerns, all continuous firm-level regressors are measured at $t - 1$, ensuring that current outcomes are related to predetermined covariates rather than contemporaneous firm conditions. Continuous variables are also winsorized at the 1st and 99th percentiles to reduce the influence of outliers. Under this empirical design, the interaction coefficient (β_2) captures the differential association between industry-level variation in zombie prevalence and the credit outcomes of non-zombie firms, after controlling for firm and industry–year fixed effects. Lastly, standard errors are clustered at the industry-year level to address within-cluster correlation. Table 1 provides a summary statistic of the variables used in the analysis.

3. Findings

3.1. Baseline results

We first find that the non-zombie indicator is positively associated with debt growth in the baseline model, but the effect becomes insignificant when firm fundamentals are included (see Table 2). This result suggests that any borrowing advantage for viable firms dissipates after controlling for financial strength. Furthermore, the interaction between non-zombie status and zombie debt share is not statistically significant, indicating that healthy firms' debt growth does not systematically decline with greater zombie prevalence in the industry.

In contrast, borrowing costs exhibit a clear pattern. The interaction term is consistently positive and highly significant; a 1 percentage point increase in the zombie debt share raises borrowing rates for non-zombie firms by approximately 1 basis point. This relationship holds even after adjusting for firm-specific controls.[1] Theoretically, this pattern can be explained by the way banks perceive risk in zombie-dense industries. Higher zombie prevalence increases lenders' concerns about sectoral default risk and recovery values, inducing them to charge a risk premium on all firms in the sector. Information asymmetry further amplifies this tendency. When lenders cannot perfectly assess firm-specific fundamentals - especially in environments with limited collateral or intangible assets - they are more likely to price credit at the sector level. Consequently, even fundamentally sound firms are subject to higher borrowing costs simply because they operate in zombie-heavy industries.

Overall, this evidence highlights a distinct financial externality that operates through credit pricing, rather than quantity constraints. Whereas previous studies emphasize resource misallocation and productivity drag (Albuquerque and Iyer, 2024; McGowan et al., 2018), we demonstrate that zombie prevalence may distort price signals in credit markets, causing lenders to inadequately differentiate firm-specific risk in zombie-dense sectors.

3.2. Credit rating-based analysis

To examine heterogeneity in how zombie prevalence relates to debt financing outcomes, we disaggregate the baseline results by firm-level credit ratings,[2] testing whether financially weaker, low-rated firms are more exposed to adverse credit conditions.

Table 3 presents the regression results by credit rating group, from "High-grade" to "Speculative-grade." Panel A shows that the interaction between non-zombie status and zombie debt share is negative and statistically significant only for speculative-grade firms. That is, in industries with higher zombie concentration, debt growth is significantly curtailed for financially weaker firms. In contrast, high- and medium-grade firms show no meaningful change in debt volume, indicating stable credit access even in zombified sectors.[3] Panel B turns to borrowing costs. The interaction term is again significant only for speculative-grade firms, indicating that each 1 percentage point increase in zombie debt share raises their borrowing rates by 1.7 basis points. The coefficients for higher-rated firms are statistically insignificant and close to zero. These results confirm that the negative effects of zombie firms on debt financing conditions are not evenly distributed across the credit spectrum but disproportionately burden weaker but non-zombie firms.

This pattern illustrates how zombie prevalence exacerbates financial polarization. While stronger firms are shielded, weaker but viable firms could face both higher borrowing costs and reduced access to credit, a dual constraint. This divergence in credit conditions intensifies firm-level vulnerabilities and distorts market-based risk pricing. The findings underscore the need for macroprudential policies that go beyond average effects and target distributional impacts, with regulatory tools calibrated to firms' financial strength and credit access

3.3. Contextual amplification: industry sector and monetary policy stance

To explore the heterogeneity in the relationship between zombie prevalence and firms' credit conditions, we further conduct subsample analyses for speculative-grade firms, those most sensitive to credit conditions. Table 4 reports results by industry (Panel A) and monetary policy stance (Panel B).

Panel A shows that zombie firm prevalence is significantly associated with financial externalities in the service sector but not in

[1] As a robustness check, we also apply a looser definition of zombie firms, requiring only two consecutive years with ICR<1. This broader threshold captures a larger set of firms, including those with more transitory episodes of financial distress. The results reported in Panel A of Table A1 remain qualitatively unchanged, indicating that our main findings are not sensitive to the precise persistence threshold used.

[2] We classify the ten-tier Value Search credit ratings into four risk groups, High-grade (1–3), Medium-grade (4–5), Non-investment-grade (6–7), and Speculative-grade (8–10), following the supervisory framework used in the *Bank of Korea Financial Stability Report* (June 2021) (Bank of Korea, 2021). This classification reflects economically salient thresholds of credit risk and broadly corresponds to global rating agency standards: for example, "High-grade" aligns with AAA–AA (S&P) and Aaa–Aa (Moody's), while "Speculative-grade" includes speculative-grade ratings, such as B–D (S&P) and Ba–C (Moody's).

[3] Our baseline measure of Δ Debt is defined as the year-on-year change in firm debt relative to the previous year's debt stock, which captures the pace of new borrowing compared with existing leverage. As a robustness check, we alternatively scale the change in debt by lagged total assets. This alternative specification mitigates concerns that debt growth may mechanically reflect firm size and ensures that our results are not driven by differences in firm scale. That is, the two measures are conceptually distinct: while the former reflects the intensity of new borrowing relative to past debt, the latter normalizes debt changes by firm scale. As reported in Panel B of Table A1, the results under this alternative specification remain qualitatively similar to the baseline, reinforcing the robustness of our findings.

Table 1
Descriptive statistics.

Variable	N	Mean	Std. Dev	10 %	50 %	90 %
Debt Growth (%)	115,398	9.71	64.14	−40.01	−0.32	59.77
Borrowing interest rate (%)	112,137	4.13	2.90	1.71	3.50	6.64
Total assets (in USD billion)	137,480	2.19	25.42	0.15	0.38	2.23
Liquidity (CA/CL, %)	137,468	212.15	279.34	34.41	125.45	449.79
Capital structure (TL/TE, %)	137,480	225.66	484.16	14.30	110.9	523.30
Profitability (EBIT/TA, %)	137,480	4.16	7.21	−2.85	3.50	12.79
ICR (EBIT/IE, times)	122,860	45.71	155.52	−2.67	2.62	66.78
Age (years)	137,480	22.84	11.74	9.0	21.0	41.0

Table 2
Baseline results.

	(1) Δ Debt	(2) Δ Debt	(3) Δ Int. rate	(4) Δ Int. rate
Non-Zombie	**5.916*****	0.268	**−0.762*****	**−0.642*****
	(1.359)	(1.568)	(0.057)	(0.052)
Non-Zombie * Zombie share	−0.026	−0.010	**0.010*****	**0.010*****
	(0.065)	(0.072)	(0.002)	(0.002)
Total assets		**−34.88*****		**0.419*****
		(1.874)		(0.047)
Liquidity		**0.047*****		**−0.001*****
		(0.004)		(8.88e-05)
Capital structure		**−0.005*****		**0.0002*****
		(0.001)		(2.26e-05)
Profitability		**0.425*****		−0.001
		(0.062)		(0.002)
ICR		**0.081*****		**−0.0025*****
		(0.005)		(0.0002)
Firm age		**0.763***		**−0.032***
		(0.423)		(0.018)
Constant	5.198***	356.2***	4.640***	0.955
	(0.755)	(19.94)	(0.035)	(0.679)
Firm FE	Yes	Yes	Yes	Yes
Industry * Year FE	Yes	Yes	Yes	Yes
Observations	115,125	112,445	111,818	110,103
R-squared	0.001	0.049	0.005	0.019

Table 3
Results by credit rating.

Panel A: Δ Debt	(1) High-grade	(2) Medium-grade	(3) Non-investment-grade	(4) Speculative-grade
Non-Zombie	−19.06	0.394	**6.088*****	**8.675*****
	(13.32)	(4.087)	(2.173)	(1.926)
Non-Zombie * Zombie share	0.469	**0.277***	−0.043	**−0.226*****
	(0.355)	(0.147)	(0.082)	(0.086)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry * Year FE	Yes	Yes	Yes	Yes
Observations	13,059	38,613	37,585	17,529
R-squared	0.042	0.076	0.111	0.078
Panel B: Δ Int. rate	(1) High-grade	(2) Medium-grade	(3) Non-investment-grade	(4) Speculative-grade
Non-Zombie	0.771	−0.136	**−0.577*****	**−0.712*****
	(1.357)	(0.133)	(0.088)	(0.088)
Non-Zombie * Zombie share	−0.001	0.003	0.004	**0.017*****
	(0.013)	(0.004)	(0.003)	(0.003)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry * Year FE	Yes	Yes	Yes	Yes
Observations	11,807	37,963	37,475	17,354
R-squared	0.021	0.029	0.028	0.013

Table 4
Results by industry sector and monetary policy stance.

Panel A: by Industry sector	(1) Δ Debt Manufacturing	(2) Δ Int. rate	(3) Δ Debt Services	(4) Δ Int. rate
Non-Zombie	**6.330*****	**−0.590*****	**9.850****	**−0.753*****
	(2.193)	(0.102)	(4.537)	(0.199)
Non-Zombie * Zombie share	−0.024	0.006	**−0.300****	**0.019*****
	(0.136)	(0.005)	(0.136)	(0.005)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry * Year FE	Yes	Yes	Yes	Yes
Observations	9585	9524	7924	7793
R-squared	0.097	0.017	0.075	0.013
Panel B: by MP stance	**(1) Δ Debt MP tightening**	**(2) Δ Int. rate**	**(3) Δ Debt MP easing**	**(4) Δ Int. rate**
Non-Zombie	5.009	**−0.525****	**8.010*****	**−0.646*****
	(5.039)	(0.239)	(2.353)	(0.107)
Non-Zombie * Zombie share	−0.109	0.009	**−0.199***	**0.015*****
	(0.207)	(0.007)	(0.105)	(0.004)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry * Year FE	Yes	Yes	Yes	Yes
Observations	5138	5053	10,625	10,511
R-squared	0.051	0.013	0.110	0.013

manufacturing. In services, a 1 percentage point increase in zombie debt share is associated with a 0.3 percentage point decline in debt growth and a 1.9 basis point increase in borrowing costs. By contrast, effects in manufacturing are small and statistically insignificant. This sectoral gap likely reflects structural differences in asset tangibility and credit assessment practices. Service firms typically rely more heavily on intangible assets and have limited collateral value, which constrains banks' ability to differentiate risk at the firm level. Credit assessments in services therefore depend more on projected cash flows, which are harder to verify and more volatile, prompting lenders to adopt sector-wide repricing strategies when zombie prevalence rises. Manufacturing firms, by contrast, generally possess more tangible and collateralizable assets, enabling banks to tailor credit terms more precisely to firm-specific conditions. This aligns with Dell'Ariccia et al. (2008), who find that finance-dependent sectors are more susceptible to credit disruptions.

Panel B examines the role of monetary policy. Adverse effects from zombie firms are concentrated during monetary easing; a 1 percentage point rise in zombie debt share reduces debt growth by 0.199 percentage points and raises borrowing costs by 1.5 basis points. Under monetary tightening, these effects are not significant. This pattern shows that accommodative monetary conditions may amplify zombie-related distortions, possibly due to looser credit standards or forbearance incentives. These findings are consistent with Banerjee and Hofmann (2022), who argue that low interest rates and ample liquidity weaken credit allocation by sustaining unproductive zombie firms. Overall, we highlight that the negative effects of zombie prevalence on debt financing conditions are both structurally and cyclically amplified, stronger in fragile sectors and under expansionary monetary regimes.[4] Macroprudential responses should therefore consider sectoral vulnerabilities and the credit cycle to better mitigate resource misallocation risks.

4. Conclusion and policy implications

This study provides new evidence that higher zombie firm prevalence is associated with distortions in credit markets. While average effects on debt volume are limited, borrowing costs for viable but financially weaker firms are significantly higher in zombie-heavy industries. This reflects a repricing externality: lenders, perceiving elevated sectoral risk, tend to increase borrowing costs broadly, even for fundamentally sound firms. The mispricing is especially pronounced among speculative-grade borrowers, intensifying their financing constraints and amplifying financial segmentation. More interestingly, these patterns are shaped not only by firm-specific characteristics, but also by macro-financial conditions. In low-interest-rate environments, banks face reduced incentives to restructure zombie exposures, prolonging misallocation. In service sectors, where intangible assets dominate and credit risk is harder to assess, lenders are more likely to adopt sector-wide pricing adjustments rather than firm-level differentiation.[5]

The policy implications are best understood as guidance for surveillance and macroprudential calibration. Macroprudential

[4] As reported Panel C of Table A1, scaling debt growth by firm size yields results similar to the baseline, thereby reinforcing the robustness of our findings.

[5] In comparing our results with prior evidence, Havemeister and Horn (2023) find little evidence of deteriorating credit conditions for healthy firms in the Euro area, whereas Albuquerque and Iyer (2024) show that credit growth of non-zombie firms is constrained in zombie-heavy industries worldwide. Taken together, our findings are broadly consistent with these global patterns, while also underscoring the contextual factors that may amplify such effects in a bank-based financial system.

frameworks could integrate indicators of zombie prevalence into supervisory dashboards and stress tests, and tailor regulatory intensity to firms' credit sensitivity. Practical measures may include sectoral capital buffers when zombie concentration is elevated, especially during accommodative monetary phases; tighter loan-classification to limit persistent rollover of unviable firms; and targeted guarantee or liquidity programs that safeguard viable but financially weaker borrowers without subsidizing zombies. Coordinated policy responses that align enhanced sectoral supervision with liquidity-support measures are essential to prevent resource misallocation and maintain credit-market efficiency.

In sum, zombie firms not only depress productivity through direct misallocation, but may also erode the integrity of credit markets by distorting pricing for viable firms. Addressing these observed links are critical to safeguarding financial system resilience in economies, where bank lending remains the dominant source of corporate finance.

CRediT authorship contribution statement

Jaeweon You: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Beomseok Seo:** Writing – original draft, Validation, Supervision, Software, Resources. **Junyoung An:** Formal analysis, Data curation.

Acknowledgement

This Research was supported by Sookmyung Women's University Research Grants (1-2503-2010).

Appendix

Table A1.

Table A1
Robustness tests.

Panel A: ICR<1 for 2 years	(1) Δ Debt	(2) Δ Debt	(3) Δ Int. rate	(4) Δ Int. rate
Non-Zombie	**4.451*****	**−3.217****	**−0.535*****	**−0.404*****
	(1.305)	(1.465)	(0.051)	(0.051)
Non-Zombie * Zombie share	−0.056	−0.045	**0.007*****	**0.008*****
	(0.053)	(0.053)	(0.002)	(0.002)
Constant	7.375***	360.6***	4.423***	0.852
	(0.646)	(19.92)	(0.026)	(0.683)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry * Year FE	Yes	Yes	Yes	Yes
Observations	115,125	112,445	111,818	110,103
R-squared	0.002	0.050	0.003	0.017

Panel B: Δ Debt scaled by size	(1) High-grade	(2) Medium-grade	(3) Non-investment-grade	(4) Speculative-grade
Non-Zombie	−0.837	0.843	**1.692*****	**3.620*****
	(1.198)	(0.598)	(0.519)	(0.682)
Non-Zombie * Zombie share	0.048	0.015	−0.014	**−0.134*****
	(0.031)	(0.025)	(0.020)	(0.035)
Constant	31.44***	94.29***	170.7***	135.4***
	(6.771)	(5.633)	(7.466)	(14.49)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry * Year FE	Yes	Yes	Yes	Yes
Observations	17,817	42,421	38,426	17,801
R-squared	0.067	0.093	0.133	0.082

Panel C: Δ Debt scaled by size	(1) Manufacturing	(2) Services	(3) MP tightening	(4) MP easing
Non-Zombie	**2.442*****	**3.732****	2.023	**3.912*****
	(0.740)	(1.530)	(1.742)	(0.915)
Non-Zombie * Zombie share	−0.019	**−0.155*****	−0.089	**−0.149*****
	(0.039)	(0.050)	(0.0715)	(0.047)
Constant	94.58***	161.3***	125.1***	160.5***
	(13.84)	(22.32)	(29.99)	(21.29)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry * Year FE	Yes	Yes	Yes	Yes

(*continued on next page*)

Table A1 (*continued*)

Panel C: Δ Debt scaled by size	(1) Manufacturing	(2) Services	(3) MP tightening	(4) MP easing
Observations	8494	8134	5252	10,751
R-squared	0.105	0.079	0.055	0.101

Data availability

Data will be made available on request.

References

Acharya, V.V., Crosignani, M., Eisert, T., Eufinger, C., 2022. Zombie lending: theoretical, international, and historical perspectives. Annu. Rev. Financ. Econ. 14(1), 21–38.

Albuquerque, B., Iyer, R., 2024. The rise of the walking dead: Zombie firms around the world. J. Int. Econ. 152 (C). Article 104019.

Albuquerque, B. and Mao, C. (2023) 'The Zombie Lending Channel of Monetary Policy', *IMF Working Paper* No. 2023/192.

Banerjee, R., Hofmann, B., 2022. corporate Zombies: anatomy and life cycle. Econ. Policy 37 (112), 757–803.

Bank of Korea. (2021) *Financial Stability Report: June 2021*.

Caballero, R.J., Hoshi, T., Kashyap, A.K., 2008. Zombie lending and depressed restructuring in Japan. Am. Econ. Rev. 98 (5), 1943–1977.

Chodorow-Reich, G., 2014. The employment effects of credit market disruptions: firm-level evidence from the 2008–2009 financial crisis. Q. J. Econ. 129 (1), 1–59.

Dell'Ariccia, G., Detragiache, E., Rajan, R.G., 2008. The real effect of banking crises. J. Financ. Intermediation 17 (1), 89–112.

El Ghoul, S., Fu, Z., Guedhami, O., 2021. Zombie firms: prevalence, determinants, and corporate policies. Finance Res. Lett. 41, 101876.

Hadlock, C.J., Pierce, J.R., 2010. New evidence on measuring financial constraints: moving beyond the KZ index. Rev. Financ. Stud. 23 (5), 1909–1940.

Havemeister, L.K., Horn, K., 2023. Fear the Walking Dead? Zombie Firms in the Euro Area and Their Effect On Healthy Firms' Credit Conditions. European Systemic Risk Board. *ESRB Working Paper* No. 143.

McGowan, M.A., Andrews, D. and Millot, V. (2018) 'The walking dead? Zombie firms and productivity performance in OECD countries', *OECD Economics Department Working Paper* No. 1372.

Technology in Society 84 (2026) 103139

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Using explainable artificial intelligence to envision the future of work and the workforce – from prediction to explanation☆

Yu-Ling Chang [a], Beomseok Seo [b,*]

[a] Graduate Institute of International Human Resource Development, National Taiwan Normal University, Taiwan
[b] Deparment of Statistics, Research Institute of Natural Science, Sookmyung Women's University, Republic of Korea

ARTICLE INFO

Keywords:
Future of work
Explainable artificial intelligence
Interpretable machine learning

ABSTRACT

Despite extensive research on the future of work, our understanding remains limited, leaving us ill-prepared for the evolving workforce requirements. Much of the literature focuses on job displacement caused by artificial intelligence (AI) and automation, assuming future work consists merely of "leftovers" after technology takes over. This perspective overlooks which types of work will thrive and emerge in the upcoming future. To address this knowledge gap, the present study employs Explainable AI (XAI) on data from the Occupational Information Network (O*NET) and the U.S. Bureau of Labor Statistics (BLS) to identify emerging work and workforce archetypes, along with the most in-demand competencies, while contrasting them with those expected to decline by 2031. The results reveal ten rapidly growing work archetypes—*Creatives*, *Sages*, *Wizards*, *Supporters*, *Nurturers*, *Caregivers*, *Explorers*, *Hands-On*, *Data Scientists*, and *Frontliners*—in contrast to six archetypes facing declines across technical, operative, and administrative domains. The study further identifies eight fastest-growing worker archetypes, including *Strategic Tech Workers*, *Workers of Technology/Engineering Design*, *Tech-Driven Leadership and Management*, *Agile Healthcare Workers*, *Business Analytics Workers*, and *Empathetic Leaders*, emphasizing hybrid talents. Conversely, the four fastest-declining worker archetypes are highly specialized in the technical and administrative roles. Finally, the study outlines the most in-demand workforce competencies, highlighting the need for blended capabilities encompassing technical expertise (Telecommunications, Engineering and Technology, Computers and Electronics), human-centered skills (Customer and Personal Service, Active Listening, Social Perceptiveness, Concern for Others, Relationship Building), as well as analytical and adaptive capabilities (Active Learning, Troubleshooting, Fluency of Ideas, Initiative, Analytical Thinking, Adaptability/Flexibility). These findings offer significant implications for individuals, organizations, educators, and policymakers to navigate the complex interplay between technological advancement and workforce transformation, offering critical insights for managing the social and economic dynamics of our technological future.

1. Introduction

In recent years, the future of work has attracted considerable attention in both scholarly research and public media. It has been one of the top ten "Big Ideas" for the National Science Foundation's (NSF) research investments since 2016, as well as a recurring series of the World Economic Forum (WEF). However, recent events such as the COVID-19 pandemic and the rapid advancement of artificial intelligence (AI) have made forecasting the future of work more urgent yet more difficult than ever. In a world of unprecedented volatility, uncertainty, complexity, and ambiguity (VUCA), the existing studies often fall short by overwhelmingly focusing on job displacement and job loss due to AI or automation (Acemoglu & Restrepo, 2020; Autor et al., 2003; Brynjolfsson & McAfee, 2011; Frey & Osborne, 2017; Josten & Lordan, 2020, pp. 1–24; Lund et al., 2019). While these studies illuminate what will likely disappear from the future of work, they leave the more critical question of what the future of work will actually look like unanswered. A major reason for this gap is that much of the existing research on job displacement and automation rests on a flawed assumption: that the future of work consists merely of the "leftovers" from AI and automation (Berg, 2017; Goldin & Katz, 2010; Manyika et al., 2017; Paus, 2018; West, 2018). They portray the future of work as the residual tasks that

☆ We have no conflict of interest to disclose.

* Corresponding author. Sookmyung Women's University, 100 Cheongpa-ro 47-gil, Yongsan-gu, Seoul, 04310, Republic of Korea.
E-mail address: bsseo@sookmyung.ac.kr (B. Seo).

https://doi.org/10.1016/j.techsoc.2025.103139
Received 25 November 2024; Received in revised form 25 October 2025; Accepted 29 October 2025
Available online 30 October 2025
0160-791X/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

remain after automation, while overlooking the potential for new, thriving forms of work that may emerge in the future.

In addition to these philosophical flaws, the existing studies on the future of work also face methodological challenges when attempting to forecast the future of work using surveys or interviews with subject matter experts (SMEs). While both surveys and interviews are well-established methods in research, the retrospective and subjective nature of these expert-based forecasts is often influenced by present-day assumptions and heuristics and lacks empirical validation. Consequently, such approaches provide insights into what experts *believe* might happen but not necessarily what *will* happen (Cass, 2018; Jaiswal et al., 2021; Störmer et al., 2014). This "after-the-fact" orientation makes it difficult to validate projections and limits the explanatory power of the future-forecasting studies in facing the rapidly changing technological and societal environment. In the context of forecasting the future of work, the reliance on subjective judgment without validations from dynamic or empirical data sources presents a significant methodological constraint. To better understand the dynamics of the future of work through data-driven approaches, Frey and Osborne's (2017) study opened a new door to the future of work studies by using machine learning models to estimate the susceptibility of occupations to automation with variables from the Occupational Information Network (O*NET). As of today, the article has gained substantial influence, with over 18,000 citations on Google Scholar. A considerable body of subsequent research has followed Frey and Osborne's (2017) approach, applying machine learning models on datasets from O*NET and the U.S. Bureau of Labor Statistics (BLS) to predict what jobs will most likely disappear in the future. Following Frey and Osborne's (2017) line of research, Lund et al. (2019), which forecasted that nearly 40 percent of current jobs in the United States would shrink by 2030, presents a good example of creating news headlines that were both stimulating and disturbing.

However, both Frey and Osborne (2017) and Lund et al. (2019) relied on relatively simple models such as linear regression and Gaussian process classification while did not account for the non-linear interactions among O*NET variables (Williams & Rasmussen, 2006). Conversely, advanced machine learning approaches, such as neural networks, are more capable of capturing complex patterns in the data to improve predictive accuracy. Nevertheless, these complex models are frequently criticized for their "black-box" nature, which limits interpretability and makes it difficult to understand the specific factors driving their predictions. To address this gap, this study applies Explainable Artificial Intelligence (XAI) to O*NET and BLS data, offering a methodological advancement over conventional black-box approaches. Compared to opaque black-box models, XAI makes it possible to identify intricate patterns and interactions between multiple variables while maintaining the accuracy and interpretability of forecasts. This approach offers a more nuanced and actionable understanding of the future of work, providing explainable insights that help individuals, organizations, human resource professionals, educators, policymakers, and governments prepare for workforce demands and adapt to an ever-evolving world of work.

Building on Frey and Osborne's (2017) seminal work on the future of work, this research continues the use of machine learning models to link the U.S. BLS and O*NET variables in exploring the future of work and the workforce while addressing the philosophical and methodological limitations in Frey and Osborne (2017) and subsequent studies such as Lund et al. (2019). This study advances the field by integrating XAI to elucidate how models generate predictions, thereby enhancing interpretability and fostering trust in AI-driven workforce analysis. The purpose of this research is to use XAI to explore the future of work and the workforce by linking O*NET variables with BLS employment projections for 2031. Specifically, this study addresses the following research questions.

RQ1. What are the fastest-growing work archetypes by 2031?
RQ2. What are the fastest-declining work archetypes by 2031?
RQ3. What are the fastest-growing worker archetypes by 2031?
RQ4. What are the fastest-declining worker archetypes by 2031?
RQ5. What are the most in-demand workforce competencies by 2031?

2. Previous research on the future of work and the workforce

While early studies examined how globalization, demographic shifts, and social upheavals shape the future of work (Autor et al., 2003; Bakhshi et al., 2017; Blinder, 2009), contemporary research tends to situate these discussions within the context of technological advancement, particularly focusing on the impact of AI, automation, and related technologies (Acemoglu & Restrepo, 2020; Autor et al., 2020; Babashahi et al., 2024; Parker & Grote, 2022; Frey & Osborne, 2017; Lund et al., 2019; Manyika et al., 2017; Störmer et al., 2014). A thorough review of the literature reveals three major streams in technology-focused future of work research: technology-driven (1) job loss, (2) job growth, and (3) job transformation.

2.1. The future of work: the job loss perspective

First, the discourses on job loss have been the most widely publicized perspective. For example, Gao et al. (2016) projected that self-driving vehicles could replace the jobs of 2.4 million truck drivers, 340,000 taxi and Uber drivers, 660,000 bus drivers, and 445,000 auto repair workers in the United States by 2030. Chui et al. (2017) predicted job losses of 5.4 million in food services and 4 million in customer service and sales by 2036. Similarly, Frey and Osborne (2017) forecasted that 47 % of jobs in the United States are susceptible to automation. Although these projections have been challenged by subsequent research, concerns about technological unemployment remain influential and have shaped much of the discourse in follow-up studies on the future of work (Koetsier, 2019; Lund et al., 2019).

2.2. The future of work: the job growth perspective

A competing theory to the job loss is the job growth perspective, which suggests that increases in employment may offset job displacement driven by technology (Paus, 2018; West, 2018). For example, Lund et al. (2019) projected employment growth in healthcare and STEM occupations, along with rising demand for roles such as personal care aides, therapists, psychologists, fashion designers, and video editors. These occupations rely heavily on interpersonal interaction, intuition, critical thinking, problem-solving, and empathy—domains in which humans remain irreplaceable and retain a competitive advantage over AI (Chuang, 2024).

2.3. The future of work: the job transformation perspective

Nevertheless, the recent discussions about the future of work increasingly emphasize job transformation. This perspective highlights the emergence of new types of jobs rather than a net loss or increase within existing occupational categories (Berg, 2017; Goldin & Katz, 2010; Manyika et al., 2017; Paus, 2018; West, 2018). It also critiques broad predictions of job eradication, arguing that automation tends to target specific tasks rather than entire occupations (Manyika et al., 2017; Parker & Grote, 2022). As a result, moving beyond the static lists of current job titles and focusing on the lower-level occupation composition such as work activities and tasks will offer greater capability to capture fundamental transformation in the future of work (Rust & Huang, 2014). From the task-oriented perspective, Autor et al. (2003) argued that routine tasks are more susceptible to computerization, while non-routine tasks are less likely to be fully automated. This assumption was later challenged by Frey and Osborne (2017), who contended that both routine and non-routine tasks could be vulnerable to automation.

For instance, modern AI systems are capable of performing tasks that involve subtle judgment and decision-making, such as fraud detection, medical diagnostics, and functions in legal and financial services. Moreover, the automation of non-routine manual tasks has become increasingly feasible due to advances in AI-powered robotics and sensor technologies. Industrial and surgical robots are now equipped with sophisticated sensors and manipulators, enabling them to carry out complex manual operations. Similarly, self-driving vehicles augmented by cameras, GPS, and remote sensing technologies are increasingly capable of outperforming human drivers in areas such as environmental scanning, driving control, and safety.

Meanwhile, in addition to the matrix of cognitive vs manual and routine vs non-routine in categorizing tasks, Jaiswal et al. (2021) proposed an alternative typology dividing tasks into doing, thinking, and feeling tasks. They argue that "feeling" tasks, which are those involving empathy, interpersonal communication, and leadership, are the most difficult for AI to replicate, followed by "thinking" tasks that require analytical and intuitive intelligence within complex cognitive contexts. In contrast, AI is expected to take over "doing" tasks such as statistical operations in the near future.

2.4. The future workforce: the future-in-demand competencies

Following the job loss and job growth perspectives, much of the early literature adopts a reactive stance in identifying future human competencies. These studies focus on the "non-automatable" human competencies—that is, areas where AI falls short. This includes intuitive, empathetic, and social intelligence, as well as higher-order cognitive, analytical, and critical thinking abilities. These competencies are essential in communication, interpersonal interaction, and leadership, especially in complex, dynamic, and context-specific situations (Chuang, 2024; Ekuma, 2023; Jaiswal et al., 2021; Lund et al., 2019; Manyika et al., 2017). Moreover, skills that match up with technology capabilities are also anticipated to be indispensable across nearly all occupations (Babashahi et al., 2024; Lund et al., 2019). This includes digital literacy, computational proficiency, data analytics, cybersecurity, and expertise in leveraging AI and automation (Brasse et al., 2024; Granillo-Velasquez et al., 2025; Lund et al., 2019). Finally, scholars further emphasize skills that promote resilience to changes such as lifelong learning skills (Chuang, 2024; Lund et al., 2019). To summarize, previous studies highlight three categories of competencies that enable humans to co-evolve with technology: (1) Human Advantage Competencies (areas where humans outperform machines), (2) Tech-Complementary Competencies (knowledge, skills, and abilities that allow humans to effectively use and work alongside emerging technologies), and (3) Resilience Competencies (adaptability and lifelong learning to stay relevant in evolving labor market requirements). While this classification offers a useful framework, its underlying assumption that humans and technology are separate, competing entities may constrain our imagination. We may not be able to define the future of work in a full picture if we only view worker competencies in contrast to technological capabilities. This adversarial lens overlooks a more constructive possibility: *what can humans and AI achieve together?* The next section explores this collaboration paradigm, reimagining the future of work and the workforce.

2.5. New paradigms on the future of work: Human-AI collaboration

The emergence of AI has spurred a paradigm shift in how we conceptualize the future of work. Early discussions largely focused on the potential for AI to automate and replace human jobs, leading to widespread concerns about job loss and technological unemployment (Autor, 2015; Frey & Osborne, 2017). Over time, however, this narrative has evolved and undergone substantial revisions. Rather than focusing solely on job displacement, contemporary research increasingly positions AI as a transformative force that creates new categories of work and reshapes existing roles (Babashahi et al., 2024). The key question has thus shifted from whether jobs will disappear to how they will be redefined, what new skills will be required, and how workers can develop effective collaborative relationships with AI. This perspective represents a fundamental shift from the paradigm of human-AI competition to one of human-AI collaboration (Alexander Obaigbena et al., 2024; O'Neill et al., 2022). Furthermore, earlier "replacement" models to understand the future of work often fell into what might be described as the "leftover" assumption. That is, human work and competency were defined as the remaining domains beyond the reach of technological capability. However, this approach risked overlooking the dynamic interplay between human and AI and how technology might reshape the future of work beyond current occupational taxonomies (Babashahi et al., 2024; Parker & Grote, 2022). In order to gain a more nuanced understanding, we need to overcome such reductive assumptions and consider a wide array of factors in forecasting the future of work. Nevertheless, given the rapid and restless pace of technological change, predicting the future of work is inherently challenging. It requires advanced approaches that are capable of processing vast amounts of data effectively. Recent developments in machine learning (ML)/AI models excel in such prediction and analysis tasks (Babashahi et al., 2024). However, despite high predictive accuracy, many of these ML/AI models lack transparency, making it difficult to understand the rationale behind algorithmic outputs and decisions (Chang & Ke, 2024). Therefore, combining the requirements for both accuracy and explainability in predicting the future of work, this study leveraged the methodology of Explainable AI (XAI) to reconcile accurate prediction with an interpretable reasoning point in analyzing the future of work. The XAI approach will be described in more detail in the method section.

3. Explainable artificial intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a subfield of AI, machine learning, and statistics that aims to enhance the interpretability of complex models. As models become larger and more complex, improving interpretability becomes crucial for understanding the underlying mechanisms of their decision-making process and fostering trust in their outputs. XAI has evolved through a range of approaches, as outlined in thorough reviews of XAI methods (Longo et al., 2024; Meske et al., 2022). Existing XAI methods generally fall into two categories: those that build intrinsically interpretable models (ante-hoc), and those that use post-hoc approaches to analyze already-trained models or specific decisions (Longo et al., 2024; Retzlaff et al., 2024). Prominent XAI methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) belong to the second category. Both methods estimate the marginal contribution of variables to a specific prediction by restricting interpretation to individual data points. However, as Seo and Li (2024) pointed out, post-hoc methods can be misleading, as their interpretations may vary significantly even for similar data points. Furthermore, these methods fail to capture the overall structure of the data and models—an aspect particularly important in social research. On the other hand, among ante-hoc "white-box" models, decision tree (DT)-based methods are often used to detect non-linear patterns (Hatwell et al., 2020), owing to their relatively straightforward explainability. Nonetheless, linear models remain widely used in data-driven analyses in social science, as the primary focus is to validate the effects of explanatory variables and examine the underlying structure of a social phenomenon. However, a common concern is that linear models often exhibit poor predictive performance when the data involve non-linear and complex relationships, as linear functions are inherently limited in capturing such interactions.

As an ante-hoc method, the Mixture of Linear Models (MLM) approach (Seo et al., 2022) employed by this study approximates deep neural networks (DNNs) using a piecewise linear function by segmenting data into groups that share similar linear predictive structures. In this manner, MLM achieves predictive accuracy comparable to that of

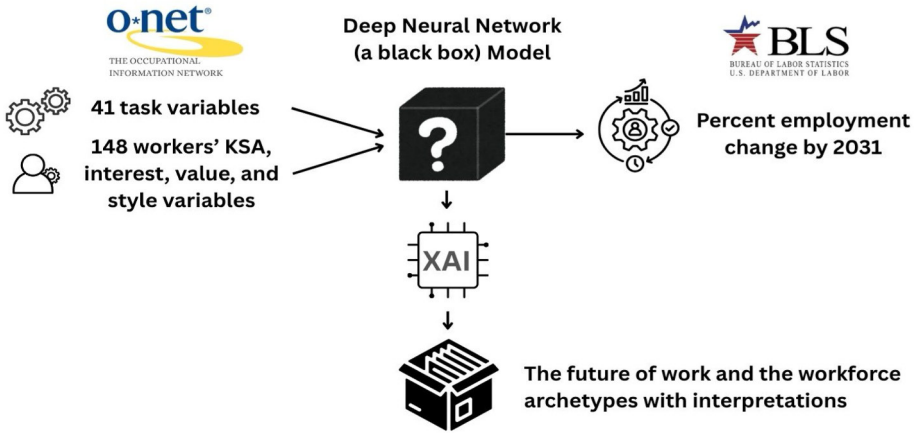


Fig. 1. The study design.

complex DNNs while offering interpretability similar to traditional linear models at the global level. This balance between predictive accuracy and explainability makes MLM particularly well-suited for this study, which aims to provide explainable insights into the future of work beyond mere prediction. The following section details the methodological framework of the study, including the design, data sources, variable construction, and implementation of the MLM approach for modeling and interpretation.

4. Methods

4.1. Design of the study

Fig. 1 illustrates the design of this study, which employs the XAI—specifically, Mixture of Linear Models (MLM) (Seo et al., 2022)—to learn the model that maps from O*NET occupational descriptors to the U.S. BLS occupational employment projections (EP) towards 2031. Both datasets are from the most authoritative data source in the United States and are cross walked through the Standard Occupational Classification (SOC) coding system, resulting in a total of 740 occupations. The goal of the model is to generate a function $f(X)$ relating O*NET variables to the future occupational demand Y, measured as the "Employment Percent Change in 2031" in the U.S. BLS-EP data. For the future of work, X includes 41 different work activities that specify different types of job behaviors and tasks performed in occupations. For the future workforce, X incorporates 151 variables in total, including 33 different types of knowledge, 35 skills, 52 abilities, 6 occupational interests, 9 work values, and 16 work styles, indicating worker characteristics and requirements. All data processing and analysis procedures were conducted using Python 3.9.

In this model, the index i denotes individual data points, each representing a distinct occupation. The learned model through MLM (Seo et al., 2022) is expected to associate the O*NET variables with the U.S. BLS occupational employment projections, in other words, to achieve $\widehat{Y}_i = f(X_i)$.

4.2. Data source and research variables

This study integrates data from two authoritative U.S. workforce databases: the Occupational Information Network (O*NET) and the Bureau of Labor Statistics (BLS) Employment Projections (EP) program. The O*NET database, developed and maintained by the U.S. Department of Labor, serves as the nation's primary source of occupational information, containing standardized descriptors for nearly 1000 occupations across the entire U.S. economy. The database employs a structured Content Model that classifies occupational characteristics into job-oriented descriptors (e.g., work activities, tasks, work contexts) and worker-oriented descriptors (e.g., knowledge, skills, abilities, work values, occupational interests, work styles). O*NET data are collected through a rigorous three-stage process involving statistically random sampling of businesses and workers, with data gathered via standardized questionnaires filled out by subject matter experts and job incumbents. This study utilizes the ONET 26.1, which is available at the O*NET website (https://www.onetcenter.org/dictionary/26.1).

The BLS-EP database provides employment projections including occupational openings, wages, education, and training for each occupation. It covers approximately 800 detailed occupations based on the SOC system. Unlike O*NET data, the BLS-EP is model-based rather than survey-based. It makes employment projections by applying economic theory and econometric techniques to process a wide range of economic data such as data from Current Population Survey (CPS), Current Employment Statistics (CES), the Quarterly Census of Employment and Wages (QCEW), Occupational Employment and Wage Statistics (OEWS), Price data from the Producer Price Index (PPI), the U.S. Census data, and other predominantly or publicly available private and government data sources. The projections are generated through a six-step integrated modeling process that begins with forecasts of labor force size and demographics, proceeds through predictions of overall economic growth and consumer demand, estimates industry-specific output, and ultimately projects employment levels for each occupation. The study utilizes the BLS-EP projection data towards 2031, which is available at the BLS website (https://www.bls.gov/emp/data.htm).

For the research variables, the independent variables of this study come from O*NET and consist of 192 specific occupational characteristics organized into seven categories. The first category, Work Activities (41 variables), captures what people actually do in a given occupation. For example, nurses would score high on "Assisting and Caring for Others" and "Making Decisions and Solving Problems," while software developers would score high on "Working with Computers" and "Thinking Creatively." The remaining six categories describe worker characteristics. Knowledge (33 variables) measures familiarity with different subject areas; for example, nurses require Medicine and Therapy knowledge, while developers need Computers and Electronics knowledge. Skills (35 variables) capture learned capabilities such as Active Listening (critical for nurses) or Programming (essential for developers). Abilities (52 variables) represent innate or developed capacities like Problem Sensitivity (recognizing when something is wrong) or Deductive Reasoning (applying general rules to specific problems). Occupational Interests (6 variables) reflect work environment preferences based on Holland (1985)'s RIASEC model: Realistic (hands-on work), Investigative (analytical work), Artistic (creative work), Social (helping people), Enterprising (persuading and leading), and Conventional (organized, systematic work). Work Values (9 variables) indicate what workers find satisfying, such as Achievement, Independence, or

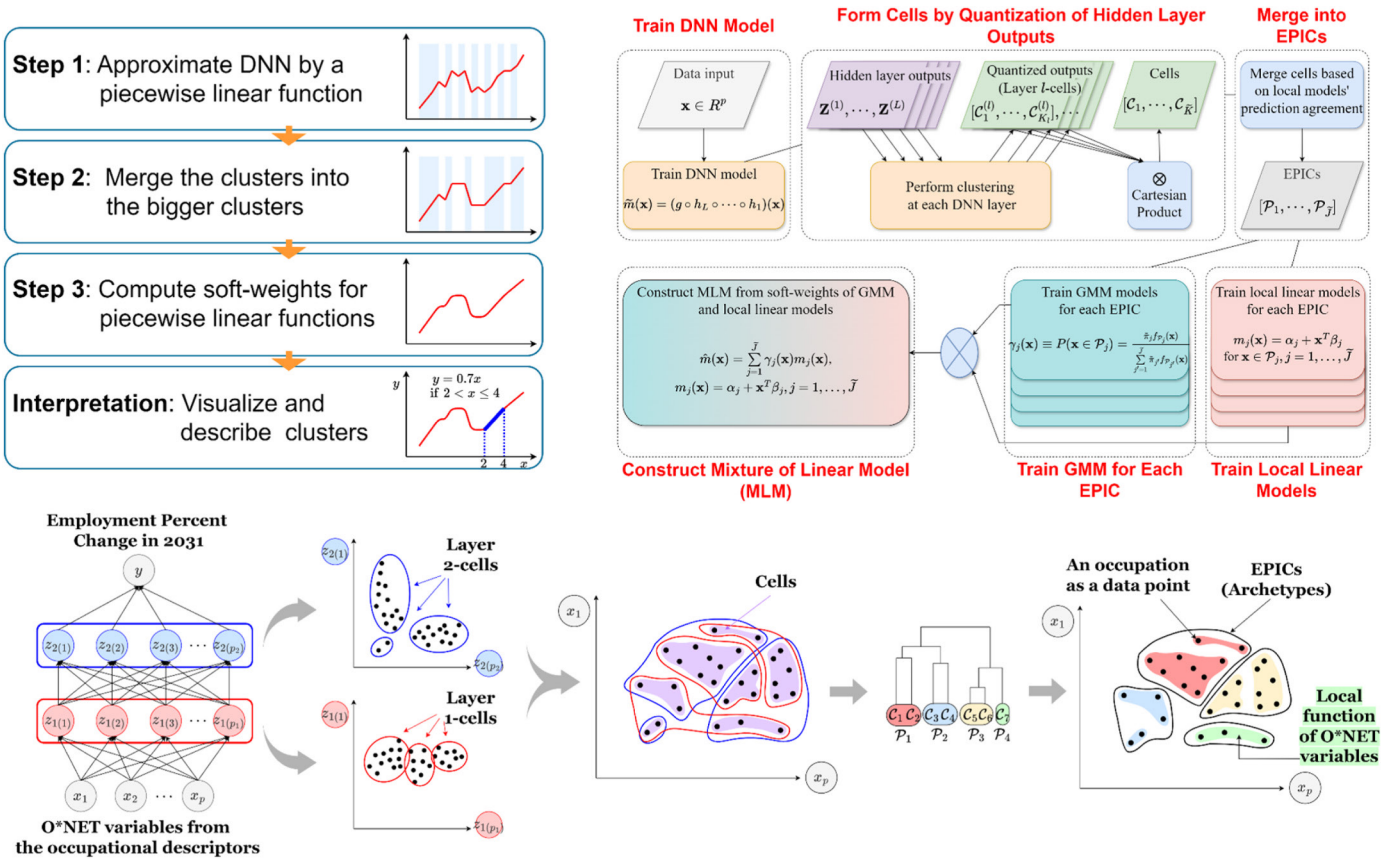


Fig. 2. The Mechanisms and Steps of MLM Modeling.
Note. MLM classifies each occupation into an archetype and conducts forecasting using the local function based on O*NET variables corresponding to each archetype. The model uncovers multiple archetypes that have different mechanisms when predicting the future employment change. Schematic plots from Seo et al. (2022).

supportive Working Conditions. Finally, Work Styles (16 variables) describe personality characteristics that affect job performance, such as Attention to Detail, Dependability, or Stress Tolerance. Each of these 192 variables is measured on an importance scale, indicating how important that particular characteristic is to successfully perform each occupation.

4.3. Mixture of Linear Models (MLM)

Fig. 2 illustrates how MLM helps to interpret the inner workings of a DNN model. The process begins by training a DNN model and then forming layer-level segments using clustering at each layer. These segments are then connected across layers to form network-level structures, referred to as *Cells*. Then, local linear models are fitted for each Cell, and Cells exhibiting similar predictive behaviors are merged into larger groups called *Explainable Prediction-induced Input Clusters* (EPICs). The local models within each EPIC are subsequently retrained, and the final interpretable model is constructed as a *Mixture of Linear Models (MLM)*. MLM partitions the sample space into segments such that the model achieves predictive accuracy comparable to that of a DNN, while allowing linear prediction and interpretation within each segment. This structure enables a trade-off between accuracy and interpretability, which is controlled by the number of segments: increasing the number of segments allows the model to capture more complex non-linear patterns, enhancing accuracy, whereas reducing the number of segments improves interpretability. We implement MLM in two versions: MLM-Cell, which includes all segments derived from the data, and MLM-EPIC, which further merges the segments to reduce their number, thereby offering greater interpretability than MLM-Cell.

In the context of this particular research, the EPICs with higher weights are interpreted as the major archetypes for the future of work and the workforce, and the local linear models can be used to interpret the unique composition of features within each archetype using original O*NET variables. By leveraging granular O*NET descriptors, the model gains analytical flexibility, allowing it to uncover latent archetypes that represent forms of future work not yet observed, based on all possible combinations of occupational characteristics. Finally, the top five in-demand workforce competencies are identified by calculating the weighted average scores of each knowledge, skill, ability, occupational interest, work value, and work style variable across all archetypes with the formula below.

$$\textit{Weighted Average Scores}_k = \sum_{j=1}^{n} \left(\frac{E_j}{\sum_{j=1}^{n} E_j} \times \beta_{jk} \right),$$

where j denotes the index of an archetype, k represents the index of an O*NET variable within the domains of knowledge, skills, abilities, occupational interests, work values, or work styles, E_j is the total projected employment in 2031 for archetype j, and β_{jk} is the coefficient for feature k in archetype j.

Overall, the use of MLM not only enhances the predictive accuracy of future job growth estimates but also provides interpretable insights into the key occupational characteristics underlined by each archetype.

5. Findings

The study first conducted an exploratory data analysis (EDA) and experiments using different machine learning models. First, the EDA

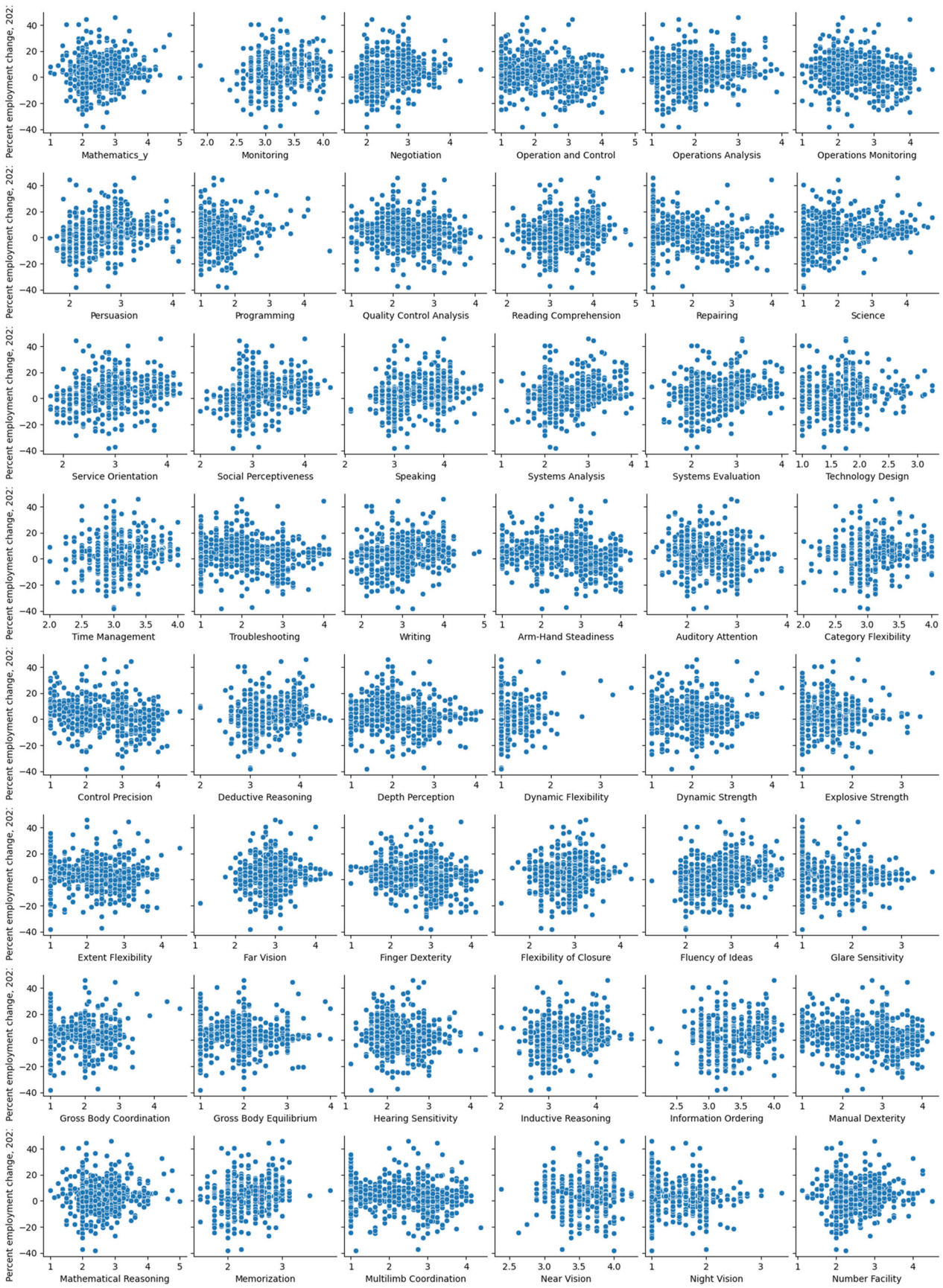


Fig. 3. Pairwise plot between employment projection and each independent variable.

Table 1
The predictive accuracy of machine learning models for the future of work and workforce.

Model	Non-Interpretable ML Models			Interpretable ML Models			
	RF	SVM	DNN	Linear Reg.	GAM	MLM-cell	MLM-EPIC
RMSE for the future of work	4.581	9.083	3.058	8.794	5.157	3.895	4.875
RMSE for the future of workforce	4.175	9.047	2.255	7.599	4.399	1.782	3.482

Note. MLM-cells consist of 628 and 671 linear segments estimated based on the DNN for the future of work and workforce, respectively. MLM-EPICs include 32 and 25 significant linear segments for the future of work and workforce, respectively.

results suggest that the relationships between employment projections for 2031 and the work- and worker-oriented variables are highly non-linear (Fig. 3). In other words, no single work activity, skill, or occupational characteristic exerts a uniform effect on employment growth across all occupations. This highlights the need for a more complex and context-sensitive model such as MLM to better capture the nuanced relationships between employment changes and O*NET variables related to work and worker characteristics.

Second, MLM models demonstrate their effectiveness by achieving significantly higher accuracy than other interpretable models, such as linear regression (Linear Reg.) and the Generalized Additive Model (GAM) (Hastie & Tibshirani, 1990), while achieving comparable performance to non-interpretable models, including Random Forest (RF) (Breiman, 2001), Support Vector Machine (SVM) (Cortes & Vapnik, 1995), and Deep Neural Networks (DNNs) (Table 1). An interesting result is that the MLM-cell for the model of the future of workers slightly exceeds DNN in predictive accuracy. This can be attributed to the MLM's architecture that regularizes the model complexity and reduces overfitting by limiting the number of segments. Furthermore, in terms of interpretability, MLM-EPICs identify 32 and 25 significant linear prediction functions for the future of work and the future of the workforce, respectively. This demonstrates that the MLM effectively balances the trade-off between predictive accuracy and interpretability. Consequently, the EDA and model experiments provide strong evidence that MLM is well-suited for addressing the study's research questions in understanding the future of work and the workforce.

5.1. The future of work

Based on size and average percentage employment growth, MLM identified 32 significant EPICs, which include a total of 712 occupations, explaining 96 % of the sampled 740 occupations. According to O*NET criteria, occupations with an employment increase of 8 % or more by 2031 are considered to be growing faster than average and have a "Bright Outlook." These occupations are expected to grow rapidly, offer a large number of job openings, or represent new and emerging fields. In our analysis, we identified ten EPICs with average employment growth of 8 % or more by 2031 and classified them as the fastest-growing work category. These ten EPICs encompass 128 occupations, representing 17 % of all US occupations and approximately 21.3 million employees in 2031. The research also explores declining archetypes, with six EPICs showing negative average employment growth rates from −2 to −10 % by 2031, which will impact over 20 million employments. Based on these results, the ten fastest-growing work archetypes along with the six fastest declining archetypes are summarized below.

5.1.1. Fastest-growing work archetypes by 2031 (RQ1)

The study identified ten rapidly growing archetypes including the *Creatives*, *Sages*, *Wizards*, *Supporters*, *Nurturers*, *Caregivers*, *Explorers*, *Hands-On*, *Data Scientists*, and *Frontliners,* with detailed descriptions summarized in Table 2. The *Creatives*, with an expected growth of 20 %, inspire and influence others through self-expression, aesthetics, and public performance. Their roles often center on communication, persuasion, and physical presence, involving activities such as selling, decision-making, and presenting ideas or products to diverse audiences. *Sage (Knowledge) Work*, expected to grow by 14 %, emphasizes creative thinking and information processing, responding to the demand for creativity and effective management of advancing information technologies. *Wizard (Applied Science) Work*, sharing the same growth projection, focuses on applying specialized knowledge and adaptability to new technologies, addressing complex challenges in a knowledge-based economy. Supporting *Work*, anticipating 13 % growth, highlights assistance and physical activities, responding to the growing emphasis on social and interpersonal aspects and the rising demand for animal care and recreation. *Nurturing Work*, with an expected 11 % growth, excels in human-centered roles that involve coaching, creative problem-solving, and interpersonal communication, relying on updated knowledge and empathetic strategies to guide and develop others. Its strengths lie in assisting, resolving conflicts, and thinking creatively, often in roles that require public interaction and team collaboration. *Caregiving Work*, projected to grow by 10 %, focuses on supporting vulnerable populations through compassionate care, medical assistance, and team-based procedures, combining emotional insight with practical support. These roles are characterized by strengths in caring, training, and advising others, with moderate involvement in physical activities and operational tasks. *Exploring Work*, expecting 9 % growth, thrives on discovery and innovation, engaging in activities that involve evaluating data, working with computers, and generating new ideas or experiences. These roles balance technical and interpersonal tasks, often contributing through scheduling, decision-making, and coordinating work across teams. *Hands-On Work*, with an anticipated 9 % growth, emphasizes practical problem solving through operating and maintaining machines, caring for others, and handling physical activities with precision and reliability. This role is grounded in technical control, equipment repair, and direct service, with a strong emphasis on action and support in physical and interpersonal environments. *Data Science Work*, projecting an 8 % growth, specializes in analyzing data and building functional solutions, often through extensive use of computers and quantitative estimation. While less focused on emotional or physical care, they excel in monitoring resources, innovating technical designs, and transforming data into impactful outcomes. Finally, *Frontline Work*, also anticipating 8 % growth, provides essential, people-centered services by directly assisting the public in areas like healthcare, food service, and administration. This type of work blends mid-level physical and planning tasks with interpersonal care, often requiring coordination, adaptability, and public-facing communication.

5.1.2. Fastest-declining work archetypes by 2031 (RQ2)

By 2031, several work archetypes are poised for rapid decline in employment, spanning the technical, operative, and administrative domains (see Table 3). The first to face a substantial decrease is the *Technicians: Mechanical Equipment and Machine* archetype, with an expected average employment change of −10 %. This role involves tasks such as repairing and maintaining mechanical equipment, performing administrative activities, and controlling machines. Key factors contributing to the downturn include a diminishing need for controlling machines and processes, performing administrative activities, and handling objects. Conversely, activities like monitoring resources, coaching, and assisting others are identified as potential catalysts for growth in this field. Following closely is the *Administrators* archetype, forecasting an average employment change of −5 %. Centered around administrative activities and computer work, this role is on the decline due to the increasing automation of administrative tasks and the prevalence of remote work. The *Operators: Technical Services or Retail*

Table 2
The fastest-growing work archetypes.

EPIC ID (Size)	Archetype	Descriptive Conditions	Work Activities that Are Most Associated with the Growth in Future Employment (Coefficients)	Average Percentage of Employment Growth by 2031	Total Employment in 2031 (Numbers in thousands)
23 (9)	**1. The Creatives:** Performing, influencing, communicating, and inspiring people through self-expression or creating and presenting aesthetics, visually appealing products to the world.	Performing for or Working Directly with the Public >3.72 (High) 4.125 < Performing General Physical Activities≤4.805 (High) Selling or Influencing Others >1.74 (Low to High) Communicating with Supervisors, Peers, or Subordinates≤4.78 (Low to High) Processing Information≤2.24 (Low to Mid) Inspecting Equipment, Structures, or Materials≤2.67 (Low to Mid) Judging the Qualities of Objects, Services, or People≤3.26 (Low to Mid) Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment≤1.31 (Low)	Performing General Physical Activities (8.43) Selling or Influencing Others (5.51) Making Decisions and Solving Problems (5.21) Developing Objectives and Strategies (5.19) Evaluating Information to Determine Compliance with Standards (5.01)	20	833 (0.5 %)
15 (14)	**2. The Sages:** Thinking creatively, analyzing, processing information through logical reasoning or scientific process to explore complex ideas or concepts in particular field of study, advance knowledge, and solve complex problems.	Thinking Creatively >3.69 (High) Communicating with Supervisors, Peers, or Subordinates >4.43 (High) Processing Information≤4.77 (Low to High) Controlling Machines and Processes≥1.605 (Low to High) 1.08 < Performing General Physical Activities≤1.33 (Low)	Getting Information (4.29) Monitoring Processes, Materials, or Surroundings (3.84) Operating Vehicles, Mechanized Devices, or Equipment (3.54) Evaluating Information to Determine Compliance with Standards (2.64) Controlling Machines and Processes (2.58)	14	182 (0.1 %)
30 (5)	**3. The Wizards:** Applying knowledge, methods, and creativity to create solutions to complex challenges, explore mysteries of the nature and science, design or innovate technical devices, or demonstrating innovative, cutting-edge insights in objects in organized, planned, prioritized manner.	Interpreting the Meaning of Information for Others >4.155 (High) 4.475 < Updating and Using Relevant Knowledge≤4.485 (High) Organizing, Planning, and Prioritizing Work >3.88 (High) Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment≤4.62 (Low to High) Thinking Creatively≤3.59 (Low to Mid) Judging the Qualities of Objects, Services, or People≤2.825 (Low to Mid) Performing General Physical Activities≤1.095 (Low)	Scheduling Work and Activities (14.76) Thinking Creatively (14.37) Repairing and Maintaining Electronic Equipment (12.69) Getting Information (11.49) Identifying Objects, Actions, and Events (9.51)	14	2987 (1.8 %)
24 (9)	**4. The Supporters:** Helping and caring for others through performing general physical activities or every-day, down-to-earth tasks such as provision of medical care or support, maintenance of equipment, household, and security.	Assisting and Caring for Others >4.165 (High) Performing General Physical Activities >4.295 (High) Repairing and Maintaining Mechanical Equipment >2.44 (Mid to High) Judging the Qualities of Objects, Services, or People≤3.35 (Mid to High) Developing Objectives and Strategies≤2.735 (Low to Mid) 2.455 < Making Decisions and Solving Problems≤2.56 (Low to Mid)	Communicating with People Outside the Organization (17.18) Performing General Physical Activities (11.86) Performing Administrative Activities (10.66) Communicating with Supervisors, Peers, or Subordinates (10.58) Coaching and Developing Others (10.14)	13	1910 (1.1 %)
17 (13)	**5. The Nurturers:** Nurturing and caring for others through direct human interaction such as communicating, developing, coaching and guidance with up-to-date knowledge as well as creative approaches and strategies.	Assisting and Caring for Others >4.495 (High) Resolving Conflicts and Negotiating with Others >4.125 (High) Thinking Creatively >4.045 (High) Interpreting the Meaning of Information for Others≤4.705 (Low to High) Communicating with Supervisors, Peers, or Subordinates≤4.7 (Low to	Performing for or Working Directly with the Public (7.74) Providing Consultation and Advice to Others (6.61) Performing General Physical Activities (5.09) Developing and Building Teams (4.58) Controlling Machines and Processes (3.64)	11	2432 (1.5 %)

(*continued on next page*)

Table 2 (*continued*)

EPIC ID (Size)	Archetype	Descriptive Conditions	Work Activities that Are Most Associated with the Growth in Future Employment (Coefficients)	Average Percentage of Employment Growth by 2031	Total Employment in 2031 (Numbers in thousands)
		High) Coaching and Developing others≤4.47 (Low to High) Judging the Qualities of Objects, Services, or People≤4.405 (Low to High) Updating and Using Relevant Knowledge≤4.085 (Low to High) Developing Objectives and Strategies≤2.83 (Low to Mid) Processing Information≤3.34 (Low to Mid)			
4 (40)	**6. The Caregivers:** Supporting, caring, or nurturing vulnerable populations with insights, compassion, empathy, and support by providing guidance, medical care, consultation and advice, or operating procedures with a team and mechanical equipment.	Assisting and Caring for Others >4.18 (High) Documenting/Recording Information >4.055 (High) Coaching and Developing Others >4.205 (High) Training and Teaching Others >3.535 (High) Developing and Building Teams >3.525 (Mid to High) Interpreting the Meaning of Information for Others≤3.385 (Low to High) Resolving Conflicts and Negotiating with Others≤4.045 (Low to High) Providing Consultation and Advice to Others≤4.345 (Low to High) Repairing and Maintaining Mechanical Equipment >1.78 (Low to High) Handling and Moving Objects >1.53 (Low to High) Staffing Organizational Units≤3.33 (Low to Mid) Operating Vehicles, Mechanized Devices, or Equipment <2.215 (Low)	Performing General Physical Activities (7.73) Training and Teaching others (7.37) Making Decisions and Solving Problems (7.23) Updating and Using Relevant Knowledge (6.46) Organizing, Planning, and Prioritizing Work (5.81)	10	8132 (4.9 %)
21 (12)	**7. The Explorers:** Exploring, adventuring, and creating new frontiers, knowledge, idea and solutions, and new experiences through working with computers, nurturing and organizing people, and evaluating quantifiable information.	3.605 < Estimating the Quantifiable Characteristics of Products, Events, or Information≤4.325 (High) Working with Computers≤4.965 (Low to High) Communicating with Supervisors, Peers, or Subordinates≤4.765 (Low to High) Training and Teaching Others≤4.56 (Low to High) Staffing Organizational Units >1.3 (Low to High)	Scheduling Work and Activities (6.62) Performing General Physical Activities (6.37) Making Decisions and Solving Problems (6.34) Identifying Objects, Actions, and Events (4.49) Coordinating the Work and Activities of Others (4.26)	9	1517 (0.9 %)
25 (7)	**8. The Hands-On:** Controlling, maintaining, repairing machines and processes and analyzing data or information to assist and care for others.	Assisting and Caring for Others >4.865 (High) Controlling Machines and Processes >4.41 (High) Repairing and Maintaining Electronic Equipment >3.04 (Mid to High) Analyzing Data or Information≤2.395 (Low to Mid)	Performing General Physical Activities (19.01) Communicating with People Outside the Organization (12.06) Repairing and Maintaining Mechanical Equipment (11.85) Communicating with Supervisors, Peers, or Subordinates (10.4) Assisting and Caring for Others (10.13)	9	455 (0.3 %)
13 (14)	**9. The Data Scientists:** Creating and innovating functional products or resource solutions based on the estimation of quantifiable data or information to assist and care for others, which is often largely involved working with computers.	4.595 < Working with Computers≤4.77 (High) Analyzing Data or Information >4.495 (High) Estimating the Quantifiable Characteristics of Products, Events, or Information≤4.0 (Low to High) 2.19 < Drafting, Laying Out, and	Establishing and Maintaining Interpersonal Relationships (10.31) Controlling Machines and Processes (9.43) Performing Administrative Activities (8.82) Monitoring Processes, Materials,	8	2012 (1.2 %)

(*continued on next page*)

Table 2 (*continued*)

EPIC ID (Size)	Archetype	Descriptive Conditions	Work Activities that Are Most Associated with the Growth in Future Employment (Coefficients)	Average Percentage of Employment Growth by 2031	Total Employment in 2031 (Numbers in thousands)
		Specifying Technical Devices, Parts, and Equipment >3.165 (Mid to High) Monitoring and Controlling Resources >2.65 (Mid to High) Repairing and Maintaining Mechanical Equipment≤1.515 (Low) Assisting and Caring for Others≤1.905 (Low)	or Surroundings (5.95) Getting Information (5.03)		
3 (54)	**10. The Frontliners:** Assisting and caring for others through performing for or working directly with the public such as serving for the provision of medical or personal care, food, or administrative services.	Assisting and Caring for Others >3.435 (Mid to High) Performing for or Working Directly with the Public >3.345 (Mid to High) Staffing Organizational Units >2.525 (Mid to High) Organizing, Planning, and Prioritizing Work≤3.605 (Low to High) Operating Vehicles, Mechanized Devices, or Equipment≤3.09 (Low to Mid) 1.87 < Repairing and Maintaining Electronic Equipment≤2.59 (Low to Mid) Scheduling Work and Activities≤3.275 (Low to Mid) Analyzing Data or Information≤3.49 (Low to Mid)	Performing General Physical Activities (6.95) Developing Objectives and Strategies (6.81) Organizing, Planning, and Prioritizing Work (5.68) Processing Information (5.34) Repairing and Maintaining Mechanical Equipment (5.07)	8	15511 (9.3 %)

Note. Because the O*NET importance scale adopted by this study ranged from 1 to 5, the study considered for each measure x, $1 \leq x < 2.3$ = Low; $2.3 \leq x < 3.6$ = Mid; $3.6 \leq x \leq 5$ = High.

archetype is next in line, anticipating a −3 % average employment change. The decline is attributed to automation and the transition to online services, affecting tasks such as controlling machines, administrative activities, and electronic equipment repair. Conversely, activities like inspecting equipment, assisting others, and monitoring resources are highlighted as crucial for fostering growth in this occupational category. The *Technicians: Craftsmanship or Process Control* archetype also expects a −2 % average employment change. This role, emphasizing monitoring processes and compliance, faces decline due to negative associations with information processing and creative thinking. On the other hand, developing teams, documenting information, and inspecting equipment are identified as vital components for potential growth in this particular field. Similarly, the *Operators: Precision and Process* archetype anticipates a −2 % average employment change. The decline is influenced by diminishing tasks such as operating machinery and obtaining information. Nonetheless, activities like inspecting equipment, assisting others, and building teams are recognized as positive contributors to future employment growth. Finally, the *Technicians: Process, Assembly, and Production* archetype projects a −2 % average employment change. Automation and advanced technologies in manufacturing processes contribute to the decline, with tasks related to information processing and machine control negatively impacting growth, whereas activities involving handling objects, inspecting equipment, and engaging in general physical tasks are associated with potential employment growth in this archetype.

5.2. The future workforce

MLM for the future workforce identified 25 significant EPICs, which include 702 occupations, explaining 95 % of the sampled 740 occupations. Through MLM, eight EPICs are projected to grow over 8 % in the next decade, representing major archetypes of the future workforce. The research also explores declining worker archetypes, with four EPICs showing negative average employment growth rates from −1 to −11 % by 2031. These EPICs are characterized by workers' knowledge, skills, abilities, interests, work values, and styles, as detailed below.

5.2.1. Fastest-growing worker archetypes by 2031 (RQ3)

Using MLM, the study identified the eight fastest-growing worker archetypes shaping the future workforce. These archetypes include the *Strategic Tech Workers*, the *Workers of Technology/Engineering Design*, the *Workers of Tech-Driven Leadership and Management*, the *Agile Healthcare Workers*, the *Tech Service Workers*, the *Workers of Social Justice and Change*, the *Business Analytics Workers*, and the *Empathetic Leaders,* with explanations formed by workers' knowledge, skills, abilities, interests, values, and styles. First on the list are *Strategic Tech Workers*, projected to see a substantial 23 % increase over the next decade, driven by their blend of technical expertise and strategic thinking, which is essential for navigating complex technological landscapes. Second, *Workers of Innovative Technology/Engineering Design* are anticipated to grow by 17 %, emphasizing their role in fostering innovation through creative ideas in technology and engineering. Third, *Workers of Tech-Driven Leadership and Management*, with a 13 % growth, showcase their proficiency in using technology to enhance business operations and decision-making. Fourth, *Agile Healthcare Workers* are expected to grow by 10 %, underlining their importance in providing medical care and ensuring healthcare systems are prepared for various challenges. Fifth, *Tech Service Workers*, projected to increase by 9 %, play a critical role in providing technology support and assistance across industries. Sixth, *Workers of Social Justice and Change*, set to grow by 9 %, are highlighted for their social and analytical skills, which are vital for addressing complex societal challenges. Seventh, *Business Analytics Workers*, with 9 % growth, are crucial for businesses that rely on data-driven decisions. Lastly, *Empathetic Leaders*, projected to grow by 8 %, are recognized for their

Table 3
The fastest-declining work archetypes.

EPIC ID (Size)	Archetype	Descriptive Conditions	Work Activities that Are Most Associated with the Changes in Future Employment (Coefficients)	Average Percentage of Employment Change by 2031	Total Employment in 2031 (Numbers in thousands)
28 (6)	**1. The Technicians: Mechanical Equipment and Machine** Working on or with mechanical equipment, technical devices or parts to control machines and processes or perform administrative activities	3.865 < Repairing and Maintaining Mechanical Equipment≤3.905 (High) Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment >2.715 (Mid to High) Performing Administrative Activities >2.655 (Mid to High) Controlling Machines and Processes≤4.645 (Low to High) Scheduling Work and Activities >1.725 (Low to High) 3.105 < Processing Information≤3.555 (Mid) Getting Information≤3.335 (Low to Mid) Thinking Creatively≤3.23 (Low to Mid)	Monitoring and Controlling Resources (17.14) Coaching and Developing Others (9.75) Assisting and Caring for Others (8.94) – Controlling Machines and Processes (−23.13) Performing Administrative Activities (−18.97) Handling and Moving Objects (−11.31)	−10	8783 (5.3 %)
20 (13)	**2. The Administrators** Working with computers to perform administrative tasks such as organizing, planning, prioritizing, and scheduling work and activities directly for the public.	Working with Computers >4.505 (High) Performing Administrative Activities >2.685 (Mid to High) Organizing, Planning, and Prioritizing Work≤3.645 (Low to High) Performing for or Working Directly with the Public≤2.815 (Low to Mid) Scheduling Work and Activities≤2.595 (Low to Mid) Judging the Qualities of Objects, Services, or People≤2.58 (Low to Mid) Interpreting the Meaning of Information for Others≤1.685 (Low)	Organizing, Planning, and Prioritizing Work (5.7) Performing for or Working Directly with the Public (5.19) Selling or Influencing Others (5.00) – Repairing and Maintaining Electronic Equipment (−7.11) Performing Administrative Activities (−6.14) Performing General Physical Activities (−5.81)	−5	4763 (2.9 %)
26 (7)	**3. The Operators: Technical Services or Retail** Directly performing services, sales, or physical tasks for or with the public	Performing for or Working Directly with the Public >4.71 (High) Identifying Objects, Actions, and Events >3.89 (High) Selling or Influencing Others≤4.71 (Low to High) Handling and Moving Objects≤4.585 (Low to High) Working with Computers >1.14 (Low to High) Documenting/Recording Information >1.425 (Low to High) Repairing and Maintaining Mechanical Equipment≤2.62 (Low to Mid) Communicating with People Outside the Organization≤2.31 (Low to Mid) Analyzing Data or Information≤2.115 (Low)	Inspecting Equipment, Structures, or Materials (14.92) Assisting and Caring for Others (11.72) Monitoring and Controlling Resources (8.41) – Controlling Machines and Processes (−9.99) Performing Administrative Activities (−9.77) Repairing and Maintaining Electronic Equipment (−6.3)	−3	245 (0.1 %)
0 (76)	**4. The Technicians: Craftmanship or Process Control** Monitoring processes, materials, and surroundings, performing general physical activities, working with computers, repairing and maintaining mechanical equipment, analyzing data and information, evaluating information to ensure compliance with standards, and drafting, laying out, and specifying technical devices, parts, and equipment to ensure optimal performance and efficiency	Monitoring Processes, Materials, or Surroundings >3.69 (High) Performing General Physical Activities >3.24 (Mid to High) 2.495 < Working with Computers≤3.695 (Mid to High) Repairing and Maintaining Mechanical Equipment >2.985 (Mid to High) Analyzing Data or Information >2.695 (Mid to High) Evaluating Information to Determine Compliance with Standards≤4.4 (Low to High)	Developing and Building Teams (9.44) Documenting/Recording Information (6.88) Inspecting Equipment, Structures, or Materials (4.45) – Processing Information (−7.01) Updating and Using Relevant Knowledge (−5.31) Thinking Creatively (−5.07)	−2	3148 (1.9 %)

(*continued on next page*)

Table 3 (*continued*)

EPIC ID (Size)	Archetype	Descriptive Conditions	Work Activities that Are Most Associated with the Changes in Future Employment (Coefficients)	Average Percentage of Employment Change by 2031	Total Employment in 2031 (Numbers in thousands)
		Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment >1.915 (Low to High)			
31 (4)	**5. The Operators: Precision and Process** Performing tasks according to specific processes or standards, typically in specialized environments and working collaboratively within a team	Monitoring Processes, Materials, or Surroundings >3.965 (High) Evaluating Information to Determine Compliance with Standards >3.68 (High) Developing and Building Teams >2.385 (Mid to High) Performing for or Working Directly with the Public≤1.35 (Low)	Inspecting Equipment, Structures, or Materials (29.22) Assisting and Caring for Others (20.97) Establishing and Maintaining Interpersonal Relationships (12.78) – Operating Vehicles, Mechanized Devices, or Equipment (−20.17) Getting Information (−13.6) Controlling Machines and Processes (−12.86)	−2	2557 (1.5 %)
8 (27)	**6. The Technicians: Process, Assembly, and Production** Controlling machines and processes, handling and moving objects, making decisions and solving problems, as well as effectively communicating, resolving conflicts, and negotiating with people to ensure smooth operations and resolve any issues that may arise	Controlling Machines and Processes >3.665 (High) Communicating with People Outside the Organization >3.38 (Mid to High) Resolving Conflicts and Negotiating with Others >3.285 (Mid to High) Handling and Moving Objects >2.515 (Mid to High) Making Decisions and Solving Problems≤4.405 (Low to High)	Handling and Moving Objects (10.07) Inspecting Equipment, Structures, or Materials (7.16) Performing General Physical Activities (5.84) – Getting Information (−8.69) Processing Information (−7.81) Controlling Machines and Processes (−6.76)	−2	591 (0.4 %)

Note. Because the O*NET importance scale adopted by this study ranged from 2 to 5, the study considered for each measure x, $1 \leq x < 2.3$ = Low; $2.3 \leq x < 3.6$ = Mid; $3.6 \leq x \leq 5$ = High.

ability to understand and relate to the emotions and perspectives of others, becoming increasingly important in diverse and inclusive workplaces. Each archetype's growth is underscored by its unique set of skills, knowledge, and work values in Table 4, reflecting the dynamic nature of the modern workforce and its alignment with emerging societal and technological trends.

5.2.2. Fastest-declining worker archetypes by 2031 (RQ4)

On the contrary, four archetypes were identified as the fastest-declining worker archetypes, including the *Administrative* Support *Workers: Mathematical Calculations*, the *Service and* Support *Technicians*, the *Mechanical and Production Technicians,* and the *Administrative* Support *Workers: Conventional Sales and Business* as summarized in Table 5. Despite their pessimistic employment outlook, the results also highlighted potential areas for each of these types of workers to reskill or upskill to enhance their employment opportunities. The first fastest-declining archetype, *Administrative* Support *Workers: Mathematical Calculations*, is projected to witness an average decrease of 11 % by 2031. This decline is attributed to factors such as automation of administrative tasks, outsourcing, and evolving job requirements. Strategies to mitigate this decline include enhancing knowledge in psychology and telecommunications, fostering skills like active learning and critical thinking, and adopting more innovative work styles. The second archetype, *Service and* Support *Technicians*, is anticipated to decline by 4 % due to automation of routine tasks, outsourcing, and the prevalence of self-service options. Opportunities for these workers lie in transitioning to medical or healthcare-related technical service and support, focusing on expertise in customer and personal service within these domains. The third archetype, *Mechanical and Production Technicians*, faces a 1 % decline, primarily influenced by technological advancements that automate routine tasks in production and processing. However, new job opportunities may emerge with specialized skills and knowledge in problem-solving, critical thinking, and technology design that address human needs. Lastly, *Administrative* Support *Workers: Conventional Sales and Business*, are expected to decrease by 1 %, driven by the automation of routine administrative tasks and the rise of remote work. Future success in this field may hinge on enhanced knowledge in psychology and customer relations, along with skills in social perceptiveness to provide humanized support. Despite the projected declines, potential avenues for growth and adaptability are highlighted within each archetype to navigate the evolving landscape of the workforce.

5.2.3. Most in-demand workforce competencies by 2031 (RQ5)

Finally, Table 6 presents the top five most in-demand competencies within each of the six major domains defined by the O*NET Content Model: knowledge, skills, abilities, occupational interests, work values, and work styles. The weighted average scores represent the weighted average coefficients that account for projected employment levels in 2031. In the knowledge domain, competencies like *Foreign Language* (4.00) and *Telecommunications* (3.08) highlight the growing need for global communication and digital infrastructure expertise, while *Customer and Personal Service* (2.87) and *Engineering and Technology* (2.77) reflect the rising importance of human-centered services and technical proficiency. The skills domain is dominated by analytical and adaptive skills, with *Active Learning* (6.36) and *Troubleshooting* (5.74) topping the list, suggesting that fast-growing occupations will require continuous learning and complex problem-solving. In abilities, *Dynamic Flexibility* (3.78), *Speed of Closure* (3.28), and *Fluency of Ideas* (3.23) point to a premium on cognitive agility and creativity in dynamic work environments. The occupational interests profile reveals a strong preference for *Realistic* (4.78) and *Artistic* (2.41) interests, hinting at the growth of practical, hands-on, and creative roles, possibly linked to skilled trades, design, and content creation. Within work values, *Achievement* (3.41) and *Recognition* (1.77) signal that high-demand

Table 4
The fastest-growing worker archetypes.

EPIC ID (Size)	Archetype	Descriptive Conditions	Worker-Oriented Variables that Are Most Associated with the Growth in Future Employment (Coefficients)	Average Percentage of Employment Growth by 2031	Total Employment in 2031 (Numbers in thousands)
23 (5)	**1. Strategic Tech Workers:** The professionals who possess both technical expertise and strategic thinking skills, allowing them to help organizations navigate complex technological challenges and opportunities.	Telecommunications≤4.572 (Low to High) Management of Material Resources≤3.312 (Low to Mid) Relationships≤3.247 (Low to Mid) Mathematics≤2.99 (Low to Mid)	**Knowledge** Mechanical (10.85) Engineering and Technology (7.1) Mathematics (6.82) **Skills** Operations Monitoring (8.29) Systems Evaluation (4.19) Management of Financial Resources (3.92) **Abilities** Rate Control (5.74) Dynamic Strength (5.71) Problem Sensitivity (5.49) **Interests** Investigative (6.52) Realistic (0.93) Enterprising (0.59) **Work Values** Achievement (5.49) Recognition (2.21) **Work Styles** Concern for Others (8.07) Achievement/Effort (5.72) Independence (2.5)	23	2464 (1 %)
19 (9)	**2. Workers of Technology/Engineering Design:** Professionals with new and innovative ideas in technology and engineering, while also emphasizing their expertise in designing and engineering solutions to best meet people's needs.	Concern for Others >4.615 (High) Persistence >2.825 (Mid to High) 3.29 < Dependability≤4.98 (Mid to High) Management of Material Resources≤3.185 (Low to Mid) Quality Control Analysis≤2.815 (Low to Mid)	**Knowledge** Engineering and Technology (9.15) Fine Arts (6.0) Foreign Language (5.68) **Skills** Coordination (3.83) Systems Evaluation (3.75) Troubleshooting (3.05) **Abilities** Time Sharing (5.85) Dynamic Strength (4.54) Originality (4.32) **Interests** Artistic (1.95) Realistic (1.75) Investigative (1.59) **Work Values** Independence (2.64) **Work Styles** Cooperation (2.61) Initiative (2.09) Adaptability/Flexibility (1.36)	17	4279 (3 %)
21 (10)	**3. Workers of Tech-Driven Leadership and Management:** Professionals who are skilled at using technology to streamline and improve business operations and decision-making. These workers may include managers, executives, consultants, and other professionals who specialize in using technology to enhance business performance.	Computers and Electronics >4.217 (High) Programming >3.285 (High) Social Orientation≤3.587 (Low to Mid) Engineering and Technology≤3.535 (Low to Mid) Mechanical≤1.701(Low)	**Knowledge** Personnel and Human Resources (12.99) Sociology and Anthropology (12.38) Administrative (11.3) **Skills** Equipment Selection (28.29) Complex Problem Solving (16.38) Management of Financial Resources (12.78) **Abilities** Oral Comprehension (18.39) Inductive Reasoning (14.83) Time Sharing (14.49) **Interests** Conventional (18.91) Realistic (5.32) Enterprising (1.64) **Work Values** Achievement (6.04) **Work Styles** Integrity (17.54) Leadership (13.48) Attention to Detail (11.41)	13	2066 (1 %)

(*continued on next page*)

Table 4 (*continued*)

EPIC ID (Size)	Archetype	Descriptive Conditions	Worker-Oriented Variables that Are Most Associated with the Growth in Future Employment (Coefficients)	Average Percentage of Employment Growth by 2031	Total Employment in 2031 (Numbers in thousands)
7 (30)	**4. Agile Healthcare Workers:** Professionals who provide medical or therapeutic care, monitoring patients, and make informed judgments and decisions based on complex data. They would likely excel in roles that require creativity, flexibility, ensuring that healthcare systems and organizations are prepared to respond effectively to a wide range of challenges, unexpected events and crises, from natural disasters to disease outbreaks.	Concern for Others >4.815 (High) Medicine and Dentistry >4.43 (High) Science≤4.44 (Low to High) Complex Problem Solving >3.0 (Mid to High) Multilimb Coordination >2.042 (Low to High)	**Knowledge** Chemistry (22.99) Therapy and Counseling (18.27) Mathematics (15.19) **Skills** Monitoring (16.94) Equipment Selection (13.48) Judgment and Decision Making (12.24) **Abilities** Fluency of Ideas (21.56) Number Facility (16.81) Dynamic Flexibility (12.18) **Interests** Investigative (6.76) Realistic (4.9) Artistic (2.71) **Work Values** Relationships (16.33) **Work Styles** Adaptability/Flexibility (14.16) Initiative (12.47) Attention to Detail (9.4)	10	5746 (3 %)
2 (49)	**5. Tech Service Workers:** Professionals who specialize in providing technology support and assistance to customers or clients in various industries. They typically have a strong understanding of technology and possess excellent problem-solving and communication skills. They also have knowledge of customer service principles and practices. Their skills include troubleshooting, system analysis, and communication, both verbal and written. They are able to quickly identify and solve technical issues for their clients or customers.	Attention to Detail >4.545 (High) Social Orientation >3.89 (High) Customer and Personal Service >3.535 (Mid to High) Troubleshooting >1.56 (Low to High) Arm-Hand Steadiness >2.317 (Low to High) Management of Material Resources >1.185 (Low to High) Sociology and Anthropology >1.62 (Low to High) Integrity≤4.095 (Low to High) Equipment Maintenance≤2.875 (Low to Mid) Judgment and Decision Making≤3.187 (Low to Mid) Innovation≤3.08 (Low to Mid)	**Knowledge** Telecommunications (23.82) Engineering and Technology (21.27) Foreign Language (19.84) **Skills** Equipment Maintenance (41.7) Operations Monitoring (23.58) Equipment Selection (22.06) **Abilities** Fluency of Ideas (25.25) Manual Dexterity (20.41) Information Ordering (19.75) **Interests** Realistic (27.93) Investigative (3.48) **Work Values** Achievement (21.88) Independence (2.38) **Work Styles** Leadership (25.32) Persistence (19.87) Integrity (17.27)	9	19318 (12 %)
11 (20)	**6. Workers of Social Justice and Change:** Professionals who are dedicated to promoting equality, fairness, and human rights in society. This type of workers is highly social and analytical. They are often motivated by a sense of compassion, empathy, and a desire to create a better world for all people. Their knowledge base includes communications and media, sociology and anthropology, and public safety and security, contributing to a deep understanding of social issues, political systems, and cultural dynamics. Their skills and abilities include troubleshooting, judgment and decision making, and active listening, dynamic flexibility, speed of closure, and rate control, allowing them to navigate complex social challenges and dynamics and work towards consensus and understanding. In terms of work style, this work archetype features leadership, social orientation, and analytical thinking.	Social >4.89 (High) Analytical Thinking≤4.105 (Low to High) Realistic≤3.0 (Low to Mid)	**Knowledge** Communications and Media (12.44) Sociology and Anthropology (9.52) Public Safety and Security (8.72) **Skills** Troubleshooting (17.05) Judgment and Decision Making (16.3) Active Listening (8.46) **Abilities** Dynamic Flexibility (26.91) Speed of Closure (18.02) Rate Control (16.14) **Interests** Artistic (5.49) Social (5.17) Enterprising (2.19) **Work Values** Independence (11.21) Achievement (10.84) **Work Styles** Leadership (11.87) Social Orientation (10.67) Analytical Thinking (8.7)	9	5425 (3 %)
6 (38)	**7. Business Analytics Workers:** Professionals who use analytical tools and techniques to analyze business information	Integrity≤4.655 (High) Achievement/Effort >4.225 (High)	**Knowledge** Economics and Accounting (11.32) Philosophy and Theology (11.19)	8	3682 (2 %)

(*continued on next page*)

Table 4 (*continued*)

EPIC ID (Size)	Archetype	Descriptive Conditions	Worker-Oriented Variables that Are Most Associated with the Growth in Future Employment (Coefficients)	Average Percentage of Employment Growth by 2031	Total Employment in 2031 (Numbers in thousands)
	and make data-driven decisions. high levels of integrity, achievement, analytical thinking, and learning strategies. They possess knowledge in the field of economics and accounting, philosophy and theology, and administration and management, which suggests an understanding of organizational financial operations, structures, and processes, as well as the knowledge to address questions related to ethics, morality, and spirituality. They also have skills in evaluating systems, which may involve analyzing data, identifying problems, and proposing solutions, monitoring, tracking and observing operations, processes or systems to ensure that they are working effectively.	Analytical Thinking >4.135 (High) Learning Strategies >3.815 (High) 2.5 < Science≤3.69 (Mid to High) History and Archeology >2.97 (Mid to High)	Administration and Management (9.55) **Skills** Systems Evaluation (10.85) Monitoring (8.56) Operations Monitoring (6.86) **Abilities** Deductive Reasoning (14.77) Dynamic Flexibility (9.01) Time Sharing (8.78) **Interests** Realistic (13.42) Artistic (6.29) Investigative (4.49) **Work Values** Independence (16.44) **Work Styles** Self-Control (6.61) Leadership (6.49) Adaptability/Flexibility (6.01)		
9 (22)	**8. Empathetic Leaders:** Professionals who possess a strong ability to understand and relate to the emotions, experiences, and perspectives of others. They have high levels of comprehension and active listening skills, indicating a strong ability to understand and communicate effectively. They are highly attuned to the needs and concerns of their team members and are able to create a supportive and inclusive work environment. They value communication, active listening, and collaboration, and are supportive and nurturing and able to create a positive work environment where team members feel heard, respected, and supported. Finally, the empathetic leaders are adaptable and flexible, able to navigate complex and challenging situations with grace and sensitivity.	Comprehension >4.328 (High) Active Listening >4.315 (High) Computers and Electronics >3.145 (Mid to High) Auditory Attention >1.345 (Low to High) Realistic≤3.835 (Low to High) Visualization≤2.56 (Low to Mid)	**Knowledge** Foreign Language (14.76) Public Safety and Security (10.43) Sales and Marketing (6.68) **Skills** Active Learning (7.75) Time Management (6.86) Complex Problem Solving (5.67) **Abilities** Far Vision (14.16) Dynamic Strength (11.05) Speech Clarity (10.58) **Interests** Realistic (7.39) Enterprising (2.58) Conventional (−1.28) **Work Values** Recognition (1.98) Support (0.8) **Work Styles** Cooperation (9.07) Leadership (7.28) Attention to Detail (6.14)	8	808 (0.5 %)

Note. Because the O*NET importance scale adopted by this study ranged from 1 to 5, the study considered for each measure x, $1 \leq x < 2.3$ = Low; $2.3 \leq x < 3.6$ = Mid; $3.6 \leq x \leq 5$ = High.

occupations increasingly attract individuals motivated by performance and visibility, potentially influenced by the rise of knowledge and creative economies. Finally, in the work styles category, *Leadership* (3.70), *Initiative* (3.37), and *Integrity* (2.43) indicate that future roles will emphasize self-direction, ethical conduct, and resilience. Collectively, these competencies suggest that future workforce readiness hinges not only on technical acumen but also on adaptability, creativity, and value-driven behaviors.

6. Discussion

The study delves into the future of work and the workforce by identifying the ten fastest-growing work archetypes (RQ1), six fastest-declining work archetypes (RQ2), eight fastest-growing worker archetypes (RQ3), four fastest-declining worker archetypes (RQ4), and the top five in-demand workforce competencies across knowledge, skills, abilities, occupational interests, work values, and styles (RQ5) projected through 2031. A matrix in Fig. 4 distinguishes between the fastest growing and declining work archetypes. The fastest declining archetypes, highlighted in red, are predominantly associated with tasks performed on objects through manual and physical processes. In contrast, the fastest growing archetypes, marked in blue, primarily center on producing outcomes from work activities for individuals, emphasizing a shift toward roles with a focus on people rather than technical processes.

Shifting focus to the future workforce, the study also scrutinizes the predictions of Frey and Osborne (2017), who assessed the likelihood of automation for various occupations based on only nine O*NET variables (circled in red in Fig. 5). However, our analysis as depicted in Fig. 5, reveals that the coefficients of these variables vary based on archetypes, challenging the notion of a universal set of skills and abilities uniformly influencing employment outcomes. Therefore, a contextualized approach, such as the MLM, is needed to better capture the complex relationships between occupational characteristics and workforce demand changes. Furthermore, the study addresses the evolving AI landscape, emphasizing a transition from traditional AI to Explainable AI (XAI), which enables transparent and interpretable processes for building trust and ensuring ethical, responsible AI use in decision-making involving people (Chang & Ke, 2024). By pioneering XAI applications in Human Resource Development (HRD) and Organizational Development (OD) research, this study advocates for greater transparency and accountability in the use of AI technologies for people analytics. Moreover, our use of XAI reveals new work and workforce archetypes, highlighting pathways not just for survival but for thriving within human-AI collaborative systems.

Finally, applying XAI on the BLS-EP projections for 2031, the study results indicate that 18 % of these occupations are rapidly declining,

Table 5
The fastest-declining worker archetypes.

EPIC ID (Size)	Archetype	Descriptive Conditions	Worker-Oriented Variables that Are Most Associated with the Growth in Future Employment (Coefficients)	Average Percentage of Employment Growth by 2031	Total Employment in 2031 (Numbers in thousands)
24 (5)	**1. Administrative Support Workers: Mathematical Calculations:** Workers with a focus on Mathematical Calculations requires a high level of mathematical ability, as well as knowledge in economics and accounting. They have lower abilities in written comprehension, operations monitoring, history and archeology, and physics. Their interests are primarily realistic and conventional, while they have little interest in artistic or enterprising pursuits.	Mathematics >3.925 (High) Economics and Accounting >3.4 (Mid to High) Written Comprehension≤3.56 (Low to Mid) Operations Monitoring≤1.28 (Low) History and Archeology≤1.01 (Low) Physics≤1.015 (Low)	**Knowledge** Psychology (16.7) Telecommunication (13.34) Administration and Management (10.4) **Skills** Operations Analysis (21.44) Science (15.44) Negotiation (14.72) **Abilities** Oral Comprehension (26.97) Category Flexibility (22.57) Reaction Time (14.54) **Interests** Realistic (24.56) Conventional (8.45) **Work Values** Independence (8.8) Recognition (5.05) Relationships (1.59) **Work Styles** Innovation (11.88) Independence (7.97) Adaptability/Flexibility (2.94)	−11	638 (0.4 %)
4 (46)	**2. Service and Support Technicians:** Workers who possess a high level of support and mechanical skills and ability in production and processing. They should also have knowledge of mechanical concepts but not necessarily high-level skills.	Support >4.135 (High) Coordination≤3.125 (Mid to High) Design >1.26 (Low to High) Telecommunications >1.095 (Low to High) Public Safety and Security≤2.495 (Low to Mid) Active Learning≤2.533 (Low to Mid) Writing≤3.0 (Low to Mid) Gross Body Coordination≤2.94 (Low to Mid) Response Orientation≤2.44 (Low to Mid) Analytical Thinking≤3.28 (Low to Mid) Learning Strategies≤2.06 (Low) Communications and Media≤1.845 (Low) Philosophy and Theology≤1.545 (Low)	**Knowledge** Medicine and Dentistry (38.89) Customer and Personal Service (33.91) Education and Training (14.75) **Skills** Active Learning (20.4) Troubleshooting (15.9) Science (12.6) **Abilities** Static Strength (37.63) Finger Dexterity (26.23) Perceptual Speed (23.16) **Interests** Artistic (5.83) Realistic (3.82) **Work Values** Recognition (38.78) **Work Styles** Achievement/Effort (27.92) Initiative (16.51) Leadership (14.97)	−4	4551 (3 %)
12 (21)	**3. Mechanical and Production Technicians:** Professionals who possess a high level of support and mechanical skills and ability in production and processing. They should also have knowledge of mechanical concepts but not necessarily high-level skills.	Support >4.443 (High) Production and Processing >3.105 (Mid to High) Mechanical≤4.805 (Low to High) Perceptual Speed≤4.0 (Low to High) Installation >1.845 (Low to High) Glare Sensitivity≤2.625 (Low to	**Knowledge** Psychology (18.35) Personnel and Human Resource (14.43) Chemistry (12.01) **Skills** Mathematics (25.94) Critical Thinking (17.99) Technology Design (15.71)	−1	1851 (1 %)

(*continued on next page*)

Table 5 (*continued*)

EPIC ID (Size)	Archetype	Descriptive Conditions	Worker-Oriented Variables that Are Most Associated with the Growth in Future Employment (Coefficients)	Average Percentage of Employment Growth by 2031	Total Employment in 2031 (Numbers in thousands)
		Mid) Reaction Time≤3.06 (Low to Mid) Telecommunications≤1.355 (Low)	**Abilities** Dynamic Strength (11.27) Problem Sensitivity (11.23) Extent Flexibility (9.53) **Interests** Enterprising (20.8) Conventional (6.55) **Work Values** Independence (18.75) Achievement (9.59) **Work Styles** Concern for Others (21.4) Dependability (16.19) Leadership (8.67)		
5 (41)	**4. Administrative Support Workers: Conventional Sales and Business:** Professionals who possess a high level of conventional interest and skills in sales and business. They may have good time management skills and value productivity as well as be socially perceptive and have some foreign language skills.	Conventional >4.61 (High) English Language >4.055 (High) Sales and Marketing >4.345 (High) Time Management >3.375 (Mid to High) Independence≤4.582 (Low to High) Social Perceptiveness≤3.875 (Low to High) Foreign Language≤3.84 (Low to High) Auditory Attention >2.06 (Low to High) Philosophy and Theology≤1.155 (Low)	**Knowledge** Psychology (10.3) Customer and Personnel Service (5.06) Personnel and Human Resources (4.37) **Skills** Operations Monitoring (10.07) Troubleshooting (6.28) Social Perceptiveness (6.07) **Abilities** Rate Control (5.17) Information Ordering (4.68) Flexibility of Closure (4.56) **Interests** Investigative (6.02) **Work Values** Relationships (3.75) Recognition (1.86) Achievement (1.68) **Work Styles** Concern for Others (2.95) Cooperation (2.12) Innovation (2.09)	−1	19754 (12 %)

Note. Because the O*NET importance scale adopted by this study ranged from 1 to 5, the study considered for each measure x, $1 \leq x < 2.3$ = Low; $2.3 \leq x < 3.6$ = Mid; $3.6 \leq x \leq 5$ = High.

Table 6
The most in-demand workforce competency.

Competency Domain	Competency	Weighted Average Scores
Knowledge	1. Foreign Language	4.00
	2. Telecommunications	3.08
	3. Customer and Personal Service	2.87
	4. Engineering and Technology	2.77
	5. Computers and Electronics	2.42
Skills	1. Active Learning	6.36
	2. Troubleshooting	5.74
	3. Systems Evaluation	4.28
	4. Operations Monitoring	3.21
	5. Time Management	3.11
Abilities	1. Dynamic Flexibility	3.78
	2. Speed of Closure	3.28
	3. Extent Flexibility	3.24
	4. Fluency of Ideas	3.23
	5. Originality	2.36
Occupational Interests	1. Realistic	4.78
	2. Artistic	2.41
	3. Social	0.82
	4. Investigative	0.37
	5. Enterprising	0.32
Work Values	1. Achievement	3.41
	2. Recognition	1.77
	3. Relationships	0.74
	4. Independence	0.59
	5. Support	0.43
Work Styles	1. Leadership	3.70
	2. Integrity	2.43
	3. Initiative	3.37
	4. Persistence	2.25
	5. Analytical Thinking	1.62

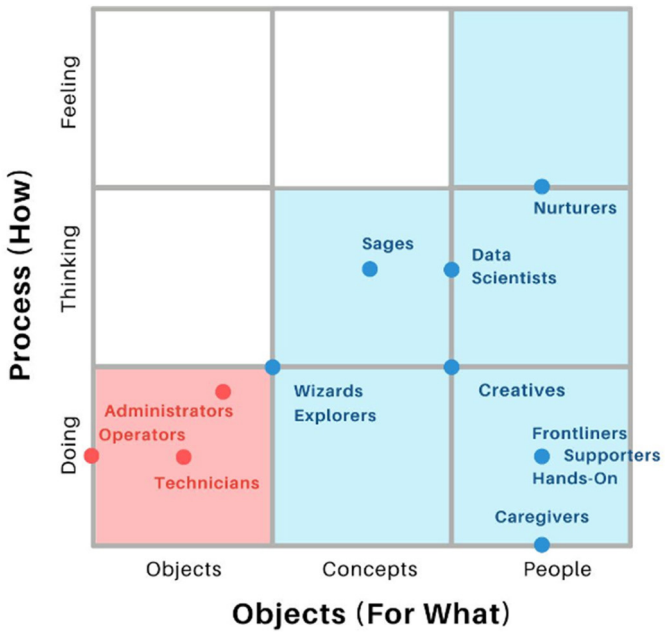


Fig. 4. Classification of work archetypes by objects and processes of activities: Fastest growing and declining.

while 24 % are experiencing rapid growth by 2031. The employment shift is reflected in workforce characteristics, with approximately 27 million workers at risk across 15 % of occupations and 44 million workers experiencing growth in 26 % of occupations. As illustrated in Fig. 6, these findings suggest that Frey and Osborne's (2017) earlier projection of a 47 % employment loss may be overestimated.

In summary, by leveraging XAI and contextualized modeling, this study moves beyond static predictions to uncover and explain the work and workforce archetypes, along with essential competencies for navigating the future of work. The findings offer a nuanced and evidence-based exploration of the evolving work and workforce landscape, challenging overly deterministic views of automation-driven displacement and instead highlighting emerging opportunities and areas for reskilling and upskilling. The following sections outline study limitations, implications for future research, and practical guidance for understanding how social forces shape technological decisions in workforce development, enabling decision-makers to thoughtfully engage with the economic, cultural, and political choices that will define technology's role in societal transformation.

6.1. Limitations

The study acknowledges several limitations particularly due to data availability. First, the variables used to describe the future of work and the workforce are based on the current scope and definition of tasks, work activities, knowledge, skills, abilities, work interests, values, and styles defined in the O*NET database. This may not be sufficient to capture new variables that may emerge in the future as technology and work environments continue to evolve. Second, the study relies on data from the U.S. government, which may limit the generalizability of the findings to other countries at different stages of development compared to the United States. Third, although O*NET data provides a standardized measure to investigate work, it lags behind job posting and recruitment platforms such as LinkedIn and other professional social media. Finally, the employment projection data used in the study covers only the period through 2031. Therefore, the findings can only provide insights into the future of work and the workforce within the time frame rather than a longer-term future. To address this limitation, the fields of study require ongoing analysis to provide timely and informed perspectives on the dynamic nature of work in the modern world.

6.2. Implications for research

The empirical findings presented in this study have the potential to inspire and guide future research and offer a roadmap for exploring the implications of the changing nature of work and the workforce. Building on this foundation, researchers can extend these recommendations to further advance the field and develop innovative solutions to the challenges facing the future of work. First, it challenges the highly influential future of work research, including Frey and Osborne (2017) and its followers, on the fundamental assumptions that the future of work is the "leftovers" from technology and automation and does not consider how the future of work will evolve and be created ahead. The study also challenges Frey and Osborne's (2017) methodology as oversimplified and encourages future research to employ more complicated and contextualized research methods such as interpretable non-linear XAI/MLM (Seo et al., 2022) to better capture the complex phenomena in the future of work. Second, the limitation of the data availability of this study also presents an opportunity for future research. As the study uses the BLS's employment projection through 2031 as the target variable, the results do not guarantee long-term future predictions but reflect the changes within the time frame. As the BLS continually announces its projections, researchers are recommended to replicate the present research efforts with the latest data. Such an iterative process will help the future of work research to be agile and adaptive to the rapidly changing technology and the world of work. Moreover, it is also advisable to incorporate dynamic data from job posting websites to capture the timely and evolving nature of work in the job market. Third, though the study only used data from the United States, we recommend researchers replicating the study design and methodology for indigenous or comparative research across different countries or regions and deepen our understanding of how different economic, cultural, and societal factors impact the future of work dynamics and develop strategies

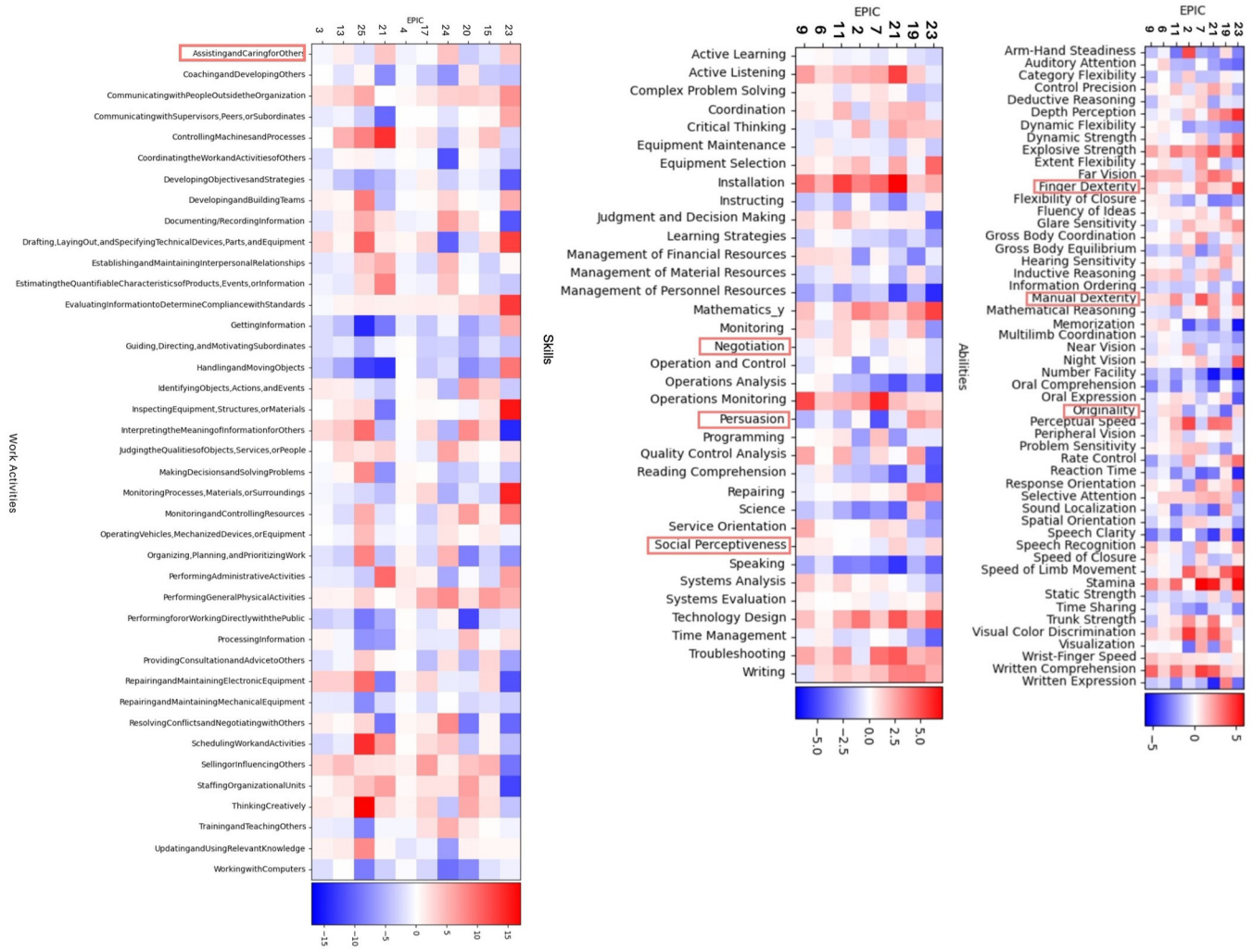


Fig. 5. The MLM regression coefficients of the fastest growing EPICs.

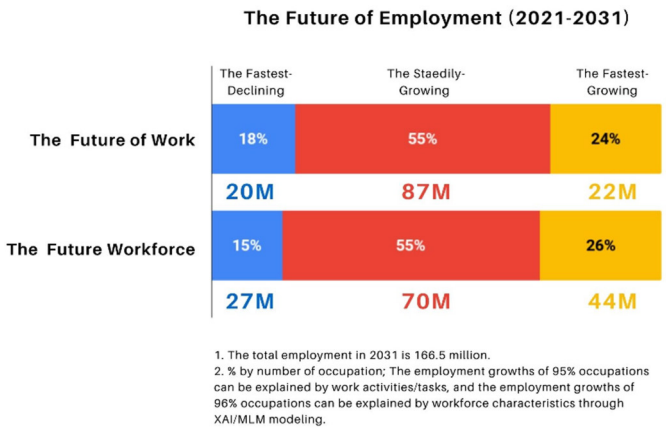


Fig. 6. The future of employment by 2031.

that are tailored to the unique needs of their workforce or to address issues related to diversity and inclusion with cross-country or cross-cultural perspectives.

6.3. *Implications for practices*

Finally, this study also has far-reaching practical implications across the entire spectrum of workforce development practices, which include career development (CD), talent development (TD), organization development (OD), and national HRD (NHRD). In CD, the findings encourage individuals to embrace future trends, cultivate hybrid skills that combine specialized and transferable competencies, and adopt agile, future-ready career plans. Leveraging the work/workforce archetype profiler is recommended for informed decision-making, enabling alignment with emerging archetypes and targeted skill development. The study underscores the importance of developing a broad skill spectrum to enhance adaptability and resilience. It also challenges deficit-based assumptions about the future of work by advocating parallel and flexible career strategies centered on meta-competence—continuous learning, skill diversification, and adaptability. Career professionals play a pivotal role in fostering this mindset and guiding individuals toward resilience and long-term success. In TD, the study highlights the need for targeted upskilling and reskilling initiatives aligned with emerging archetypes and in-demand competencies. TD professionals should tailor learning programs to match the specific needs of archetypes within their organizations, thereby enhancing individual capability and organizational agility. Furthermore, leadership development efforts should focus on two key emerging archetypes: *Tech-Driven Leadership and Management* and *Empathetic Leadership* and equip future leaders with the skills necessary to navigate complexity, foster innovation, and drive strategic transformation. In OD, professionals are encouraged to assess workforce structures, capabilities, and strategies in relation to both growing and declining archetypes. Future-oriented

workforce planning involves using archetype data to identify high-demand skills, align business strategies, and drive change initiatives. This includes redesigning roles, restructuring processes, fostering a culture of continuous learning, and integrating new technologies to remain competitive and adaptive. Finally, regarding NHRD and policy, the study offers valuable insights for designing forward-looking workforce development programs. Policymakers are urged to invest in education and training systems that support individuals transitioning from declining archetypes to new, sustainable career pathways. Importantly, identifying declining archetypes does not imply a value judgment about particular jobs or workers, but rather signals shifts in employment opportunities requiring strategic planning and support. In summary, the present study highlights the importance of focusing on specific work activities, tasks, and competencies associated with employment growth to build a resilient and adaptable workforce ecosystem that can thrive amid the rapidly evolving, technology-integrated future.

CRediT authorship contribution statement

Yu-Ling Chang: Writing – original draft, Validation, Resources, Formal analysis, Conceptualization, Writing – review & editing, Visualization, Software, Investigation, Data curation. **Beomseok Seo:** Writing – review & editing, Methodology, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. *Journal of Political Economy, 128*(6), 2188–2244. https://doi.org/10.1086/705716

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Babashahi, L., Barbosa, C. E., Lima, Y., Lyra, A., Salazar, H., Argôlo, M., Almeida, M. A. D., & Souza, J. M. D. (2024). AI in the workplace: A systematic review of skill transformation in the industry. *Administrative Sciences, 14*(6), 127. https://doi.org/10.3390/admsci14060127

Berg, M. (2017, December 11). The highest-paid YouTube stars 2017: Gamer DanTDM takes The Crown with $16.5 million. *Forbes*. https://www.forbes.com/sites/maddieberg/2017/12/07/the-highest-paid-youtube-stars-2017-gamer-dantdm-takes-the-crown-with-16-5-million/#2c9291441397.

Blinder, A. S. (2009). How many US jobs might be offshorable? *World Economics, 10*(2), 41–78. http://gesd.free.fr/blinder9.pdf.

Brasse, J., Förster, M., Hühn, P., Klier, J., Klier, M., & Moestue, L. (2024). Preparing for the future of work: A novel data-driven approach for the identification of future skills. *Journal of Business Economics, 94*(3), 467–500. https://doi.org/10.1007/s11573-023-01169-1

Brynjolfsson, E., & McAfee, A. (2011). Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy. *Brynjolfsson McAfee*.

Cass, O. (2018). *The once and future worker: A vision for the renewal of work in America. Encounter Books*.

Chang, Y. L., & Ke, J. (2024). Socially responsible artificial intelligence empowered people analytics: A novel framework towards sustainability. *Human Resource Development Review, 23*(1), 88–120. https://doi.org/10.1177/15344843231200930

Chuang, S. (2024). Indispensable skills for human employees in the age of robots and AI. *Eur. J. Train. Dev., 48*(1/2), 179–195. https://doi.org/10.1108/EJTD-06-2022-0062

Chui, M., George, K., Manyika, J., & Miremadi, M.. September 7). Human + machine: A new era of automation in manufacturing. McKinsey & Company. https://www.mckinsey.com/capabilities/operations/our-insights/human-plus-machine-a-new-era-of-automation-in-manufacturing.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Gao, P., Kaas, H.-W., Mohr, D., & Wee, D.. 1,). Disruptive trends that will transform the auto industry. McKinsey & Company. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/disruptive-trends-that-will-transform-the-auto-industry.

Goldin, C., & Katz, L. F. (2010). *The race between education and technology*. Harvard University Press.

Granillo-Velasquez, K. E., Musemeche, N. P., & Reyes, D. L. (2025). Mapping the modern workforce: An overview of workforce competency demands. *The Career Development Quarterly, 73*(1), 48–67. https://doi.org/10.1002/cdq.12367

Hatwell, J., Gaber, M. M., & Azad, R. M. A. (2020). *CHIRPS: Explain. Random For. Class. Artif. Intell. Rev., 53*, 5747–5788.

Holland, J. L. (1985). *Making vocational choices: A theory of vocational personalities and work environments*. Englewood Cliffs.

Jaiswal, A., Arun, C. J., & Varma, A. (2021). Rebooting employees: Upskilling for artificial intelligence in multinational corporations. *International Journal of Human Resource Management*. https://doi.org/10.1080/09585192.2021.1891114

Josten, C., & Lordan, G. (2020). *Robots at work: Automatable and non-automatable jobs* (pp. 1–24). Springer International Publishing.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., … Stumpf, S. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion, 106*, Article 102301.

Lund, S., Manyika, J., Segel, L. H., Dua, A., Hancock, B., Rutherford, S., & Macon, B. (2019). *The future of work in America: Peaople and place, today and tomorrow*. McKinsey Global Institute. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/future%20of%20organizations/the%20future%20of%20work%20in%20america%20people%20and%20places%20today%20and%20tomorrow/the-future-of-work-in-america-full-report.pdf.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., & Sanghvi, S. (2017). *What the future of work will mean for jobs, skills, and wages*. McKinsey Global Institute. https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53–63.

Paus, E. (2018). *Confronting dystopia: The new technological revolution and the future of work*. Cornell University Press.

Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Roettger, R., Mueller, H., & Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research, 86*, Article 101243.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Seo, B., & Li, J. (2024). Explainable machine learning by SEE-Net: Closing the gap between interpretable models and DNNs. *Scientific Reports, 14*(1), Article 26302.

Seo, B., Lin, L., & Li, J. (2022). Mixture of linear models co-supervised by deep neural networks. *Journal of Computational & Graphical Statistics, 31*(4), 1303–1317. https://doi.org/10.1080/10618600.2022.2107533

Störmer, E., Patscha, C., Prendergast, J., Daheim, C., Rhisiart, M., Glover, P., & Beck, H. (2014). *The future of work: Jobs and skills in 2030. UK commission for employment and skills*.

West, D. M. (2018). *The future of work: Robots, AI, and automation*. Brookings Institution Press.

Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning. *2* (3), 4.

여인권 교수

논문

- I.K. Yeo (2025). Comparison of correlation coefficients of dietary consumption and usual intake distributions by gender. The Korean Journal of Applied Statistics. 38, 469-483

- S. Kwon, W.S. Cho, and I.K. Yeo (2025). Forecasting peak demand for electricity using time series models and artificial intelligence models. The Korean Journal of Applied Statistics. 38, 823-833.

The Korean Journal of Applied Statistics
2025, Vol. 38, No. 4, 469–483

DOI: https://doi.org/10.5351/KJAS.2025.38.4.469
Print ISSN1225-066X / Online ISSN 2383-5818

Comparison of correlation coefficients of dietary consumption and usual intake distributions by gender

In-Kwon Yeo[1,a]

[a]Department of Statistics, Sookmyung Women's University

Abstract

Usual intake refers to the long-term average consumption of a particular food or nutrient over an extended period, serving as a crucial foundation for establishing nutrition and public health policies. Estimating usual intake requires long-term consumption data. Instead of directly collecting such data, conventional analytical methods use statistical models to generate simulated consumption data for analysis. The correlation of intake amount plays a crucial role in the process of generating simulated intake. To generate simulated intake, at least two days of dietary intake data are required, and these data are used to derive the correlation coefficient of intake amount. Previous studies have derived correlations by analyzing the entire dataset without dividing it based on demographic factors such as gender. In this study, we examine whether there are differences in the correlation coefficients obtained by dividing the data by gender. We also compare the usual intake distribution derived using the correlation coefficient from the entire dataset with the usual intake distribution derived using the correlation coefficients from gender-segregated data. An empirical analysis using the nutrient data from the 2019 Korea National Health and Nutrition Examination Survey (KNHANES) confirmed significant differences in both the correlation coefficients and the usual intake distributions.

Keywords: EMM, generalized estimating equation, generalized linear model, usual intake

1. 서론

일상섭취량(usual intake)은 어떤 식품이나 영양소를 오랜 기간 얼마나 섭취하는지를 평가하는 장기간의 평균 섭취량을 의미한다. 일상섭취량은 영양이나 보건 정책을 수립하는데 중요한 기초 자료로 사용되고 있다. 우리나라 일상섭취량 연구는 대부분 국민건강영양조사의 식품섭취조사 자료를 기반으로 이루어진다. 이 조사는 국민건강증진법 제16조에 근거하여 국민의 건강수준과 식품 및 영양 섭취 등의 실태를 파악하기 위해 전국적으로 이루어지는 조사이다. 다음 절에서 국민건강영양조사에 대해 자세히 알아본다.

장기간의 평균 섭취량을 계산하기 위해서는 표본으로 선택된 개체의 섭취 자료를 장기간 반복 조사해야 알 수 있다. 하지만 식품섭취조사가 이루어지고 있는 대부분의 나라에서는 한 개체에 대해 두 번 정도 조사가 이루어지고 있으며 현재 우리나라에서는 조사 대상자 중 일부에 대해 비연속 이틀간의 섭취량을, 나머지 대상자에 대해 하루 섭취량을 조사하고 있다. 제한된 빈도의 조사 자료로 일상섭취량을 유도하기 위해 대부분의 분석방법에서는 반복측정 자료를 분석할 때 많이 사용되는 혼합모형으로 개체 내와 간의 분산을 추정하여 상관계수를 유도한다. 개인의 섭취 패턴이 유지된다는 가정하에서 추정 모형으로부터 모의 섭취량을 추가로

[1]Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-Gu, Seoul 04310, Korea.
E-mail: inkwon@sookmyung.ac.kr

Published 31 August 2025 / journal homepage: http://kjas.or.kr
© 2025 The Korean Statistical Society. All rights reserved.

Table 1: Comparison of KNHANES

기수	기간	조사모집단	식품섭취조사 표본크기
제1기	1998	1995년 인구주택총조사(1995), 신축아파트목록(1997)	10400
제2기	2001	인구주택총조사(2000)	9968
제3기	2005	인구주택총조사(2000), 신축아파트목록(2001)	9047
제4기	2007~2009	인구주택총조사(2005)	(4099, 8641, 9397)
제5기	2010~2012	주민등록인구(2009), 아파트시세조사(2008)	(8027, 7709, 7213)
제6기	2013~2015	인구주택총조사(2010)	(7242, 6803, 6630)
제7기	2016~2018	인구주택총조사(2010), 주택공시가격 자료(2011~2014)	(7049, 7169, 7069)
제8기	2019~2021	인구주택총조사(2015)를 행정자료(2016)로 갱신	(7151, 5809, 5944)
제9기	2022~2024	인구주택총조사(2020)기반 등록센서스(2022)	(5832, - , -)

생성하여 일상섭취량을 추정한다. 이 과정에서 이틀간의 섭취량 자료의 상관관계는 모의 섭취량을 생성할 때 매우 중요한 역할을 한다.

대부분의 섭취량 분포는 정규분포 기반의 일반선형모형을 적용하기 어려울 정도로 양의 왜도를 가지며 어떤 식품은 조사 시점에 섭취하지 않은 비율이 높아 섭취량이 0인 자료를 과잉 포함하고 있다. 이를 모형화하기 위해 대부분의 분석법에서는 섭취 여부를 먼저 분석하고 섭취량이 모두 0보다 큰 개체의 자료만을 이용하여 섭취량 분석을 한다. 혼합모형(mixed effects model)을 사용하는 경우 섭취량이 0보다 큰 자료에 대해 Box-Cox 변환을 적용하여 등분산성 및 정규성과 같은 선형모형의 가정을 보다 만족하게 만든 후 모수를 추정한다. 이렇게 추정된 모형에서 모의 섭취량을 발생시키고 이를 역변환하여 최종 모의 섭취량을 유도한다. Yeo (2023)는 Box-Cox 변환을 적용한 혼합모형 대신 감마분포를 기반으로 한 일반화추정방정식(generalized estimating equation; GEE)을 적용하여 모의 섭취량을 유도하였는데 이 방법 또한 조사 시점의 모든 섭취량이 0보다 큰 자료에 대해서만 적용할 수 있다. 이에 대한 자세한 내용도 다음 절에서 알아본다.

앞에서도 언급한 것 같이 모의 섭취량을 생성하는 과정에서 상관계수가 매우 중요하다. 만약 성별이나 연령대와 같이 그룹별로 나눈 자료의 상관계수들이 유의하게 다르다면 이를 모의 섭취량을 생성할 때 반영해야 보다 정확한 일상섭취를 구할 수 있을 것이다. 이 논문에서는 주요 영양소에 대해 성별로 나누었을 때 상관계수에 유의한 차이가 있는지를 알아보고자 한다. 또한, 그룹별로 얻은 상관계수를 적용한 일상섭취 분포와 전체 자료의 상관계수를 적용한 일상섭취 분포를 비교해 본다.

2. 일상섭취

2.1. 국민건강영향조사

국민의 건강에 대한 보건정책을 마련하고 평가하기 위한 기초자료를 얻기 위해 1998년부터 국민건강영양조사가 수행되고 있다. 제1기부터 제3기까지 3년 주기로 조사를 수행하는 것으로 계획하였으나 국민건강증진종합계획(Health Plan 2010)의 지표생산을 위한 조사항목 재검토로 인해 제3기 조사는 2004년이 아닌 2005년에 수행되었다. 제4기부터는 매년 조사로 바뀌었으며 3년 주기로 조사환경을 고려하여 조사모집단을 변경하며 조사를 진행해 오고 있다. 자세한 조사모집단 변경 사항은 Table 1을 참고하기 바란다. 제4기를 제외한 기수에서는 조사구 선정 후 가구를 추출단위로 선택하는 2단계 층화집락추출법이 사용되었으며 제4기에서는 동읍면 선정 후 조사구와 가구를 선택하는 3단계 층화집락추출법이 사용되었다. Table 1에서 제4기부터의 식품섭취조사 표본크기는 3년동안 각 연도별로 조사에 참여한 표본크기이며 논문을 작성하는 시점에 제9기의 2023년과 2024년 자료를 찾을 수 없었다.

조사내용은 기수에 따라 조금씩 차이가 있는데 제8기의 경우 건강설문조사, 검진조사, 영양조사(식생활

조사, 식품안정성조사, 식품섭취조사), 가정 실내공기질 및 환경유해물질 생체지표 조사가 진행되었다. 식품섭취조사는 1세 이상을 대상으로 조사 1일 전 어떤 음식을 얼마나 섭취량 했는지를 24시간 회상법으로 조사한다. 2011년까지는 63개 식품에 대해, 2012년 이후부터는 112개 음식항목에 대해 조사가 이루어지고 있으며 2020년까지 영양조사는 모든 요일에 조사가 이루어진 반면 2021년 조사에서는 조사원의 근무조건 등의 이유로 주중에만 조사가 수행되고 있다. 이는 조사 참여 1일 전의 식사섭취 내용을 수집하기 때문에 2021년 조사에서는 금요일과 토요일의 섭취자료가 수집되지 않고 있다는 것을 의미한다.

국민건강영양조사에 참여한 가구원의 특성 중 시계열적으로 유의하게 변한 것은 연령으로 1998년과 2001년 조사에서는 30대가 17% 후반으로 가장 높은 비율을 차지했으며 70대 이상은 5%대로 가장 낮은 비율을 차지하고 60대가 8%대로 그 다음으로 낮은 비율을 차지했다. 이에 반해 2021년 조사에서는 70세 이상이 18.2%로 가장 높았고 60대가 16.2%로 두 번째로 높았으며 9세 이하가 7.8%로 가장 낮은 것으로 나타났다. 월평균 가구소득이 300만 원 초과인 비율이 제4기에는 26.5%이었으나 제5기와 제6기에는 40% 정도로 늘었고 제7기는 48.6%, 제8기는 50.3%로 늘어나고 있으나 제8기 동안의 가구중위소득을 고려할 때 300만원 초과 가구의 비율은 아직 낮은 수준으로 파악되었다.

조사와 관련된 보다 자세한 내용은 질병관리청에서 제공하는 원시자료 이용지침서를 참고하기 바란다. 이 지침서에 의하면 기수에 따라 조사모집단과 식품항목에 차이가 있고 식품섭취량 계산법이 다르기 때문에 기수별로 비교 분석하거나 추세 분석을 할 때 이런 내용을 고려할 필요가 있음을 강조하고 있다.

2.2. 일상섭취 분석모형

특정 시점에서의 모집단 섭취량은 해당 기간에 조사된 개별 섭취자료와 각 표본의 가중치를 이용하여 유도할 수 있다. 이에 반해 일상섭취량은 표본으로 선택된 개체를 장기간 반복 측정해야 알 수 있다. 하지만 식품섭취조사가 이루어지고 있는 대부분의 나라에서는 두 번 정도의 제한된 빈도로 조사가 이루어지며 우리나라의 경우 일주일 정도의 간격으로 이틀간 섭취량을 조사하고 있다. 그래서 일상섭취 분석에서는 수집된 섭취자료를 분석하고 개인의 섭취 패턴이 장기간 유지된다는 가정하에서 모의 섭취량을 추가로 생성하여 일상섭취량을 추정한다.

식이섭취를 분석하는 방법으로는 Iowa State University에서 개발한 ISU, 미국 국립암연구소(National Cancer Institute)에서 개발한 NCI, 네덜란드 RIVM (Dutch National Institute for Public Health and the Environment)에서 개발한 SPADE (Statistical Program to Assess Dietary Exposure), 독일 DIfE (Department of Epidemiology of the German Institute of Human Nutrition Potsdam-Rehbrücke)의 MSM (The Multiple Source Method)등이 있으며 관련 사이트는 References를 참고하기 바란다. Souverein 등 (2011)과 Laureano 등 (2016)은 이들 4개 모형으로 개체 내와 개체 간 분산에 대해 여러 시나리오를 가정하여 발생시킨 모의자료를 분석 비교하였다. 이들 결과에 의하면 표본이 적을수록 평균 편향과 분산이 커지고 섭취량의 분포가 정규분포에서 멀어질수록 분산이 커지는 것으로 분석되었다. 또한 영양 부족이나 과잉 섭취의 기준이 될 수 있는 10백분위수와 90백분위수 추정에서의 결과가 좋지 않은 것으로 나타났다. 이런 문제점이 있음에도 불구하고 북미와 우리나라 식이섭취 관련 연구에서는 NCI가 많이 활용되고 있다. 전 세계적으로 높은 인지도를 가진 NCI는 다음과 같은 4단계를 거쳐 일상섭취량을 추정한다.

단계 1. 식이섭취량의 정규성을 높이는 Box-Cox 변환 모수를 추정함.

단계 2. 혼합모형으로 평균과 개체 내와 개체 간의 분산을 추정함.

단계 3. 단계 2에서 추정된 평균을 평균으로, 개체 간 분산을 분산으로 설정한 정규분포로부터 k개의 모의 섭취량을 생성함.

단계 4. 개체 내 변동을 적용하고 Box-Cox의 역변환을 적용하여 식이섭취량을 유도함.

SPADE와 MSM도 세부 과정에서 약간의 차이만 있을 뿐 전체적으로 유사한 과정을 거친다. 단계 1에서 Box-Cox 변환을 적용하기 위해서는 섭취량이 0보다 커야 한다. 식품이나 일부 영양소의 경우 조사 기간에 섭취하지 않은, 즉 섭취량이 0인 경우도 있다. 이 문제를 해결하기 위해 실제 분석에서는 먼저 섭취여부에 대한 분석을 하고 섭취량이 0보다 큰 자료에 대해서 섭취량 분석을 한다. 그러므로 단계 3에서는 섭취여부 분석에서 유도된 섭취확률을 기반으로 섭취여부에 대한 난수를 생성하고 섭취하는 경우에 섭취량 난수를 생성한다. 개체별로 모형적합에 사용된 2개의 섭취자료와 k개의 모의 섭취량의 평균으로 일상섭취량을 추정한다. 모집단이나 특정 범주에 속하는 모집단의 섭취량 분포는 이 일상섭취량에 개체 가중치를 반영하여 유도한다.

본 논문에서는 기존 분석모형 대신 Yeo (2023)가 제안한 모형(ensemble method based on marginal-effect models; EMM)을 기반으로 분석하고자 한다. 이 모형의 초기 버전은 '2007년 영유아, 어린이 및 청소년의 식품섭취량 특별조사' 연구를 위해 한국보건산업진흥원의 의뢰를 받아 본 연구자에 의해 개발되었으며 관련 연구내용은 Kim 등 (2009)을 참고하기 바란다. 초기 버전의 분석모형은 이후 '위해평가를 위한 민감군 섭취량 조사(Kim 등, 2013)'와 '우리나라 적색육 및 가공육 섭취실태 및 섭취가이드 기초조사 (Lee 등, 2016)' 등에서 사용되었다. Lee 등 (2020)이 질병관리청의 정책연구용역사업을 수행하면서 붓스트랩 기법을 활용한 개선모형을 도출하고 Yeo (2023)에 의해 특허등록 되었다. Chi 등 (2023)은 다양한 섭취 시나리오 하에서 EMM과 기존 NCI의 결과를 비교하였는데 90 이상의 백분위수 추정에서 NCI보다 편향이 줄어든 것으로 나타났다. 이에 대한 이유를 아래에서 다시 설명한다.

본 논문에서 사용할 모형 및 방법론을 이해하기 위해 기존 연구에서 다루었던 모형을 좀 더 구체적으로 알아볼 필요가 있다.

Y_{ij}을 i번째 개체가 j번째 조사일에서 응답한 식품이나 영양소의 섭취량이고 $I_{ij} = I(Y_{ij} > 0)$는 섭취여부를 나타내는 확률변수라고 하자. 식품섭취조사 표본크기를 n이라고 하면 $i = 1, \ldots, n$이고 우리나라 식품섭취조사는 1주일 정도의 간격으로 이틀 조사하므로 $j = 1, 2$가 된다. 즉, i번째 개체의 관측값은 $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^T$, $\boldsymbol{I}_i = (I_{i1}, I_{i2})^T$로 표시할 수 있다. 섭취량에 영향을 주는 설명변수는 주로 범주형인 요인(factor)으로 동일 개체에 대해 j와 관계없이 동일한 값을 가진다. 즉, $\boldsymbol{x}_{ij} = \boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$이다. 이는 조사일 j와 관계없이 기댓값은 $E(I_{ij}) = \theta_i$와 $E(Y_{ij}) = \mu_i$로 가정해도 무리가 없다는 것을 의미한다.

기존 분석모형에서는 섭취여부에 대한 확률을 추정하기 위해 연결함수가 로짓(logit)인 일반화선형혼합모형(generalized linear mixed model; GLMM)을 적용하였다. 즉, 변량효과 $\boldsymbol{U}_i$가 주어졌을 때 I_{ij}의 조건부 확률분포는 베르누이 분포를 따르고 $E(I_{ij}|\boldsymbol{U}_i) = \theta_{ij} = \theta_i$는 다음의 관계식을 만족한다.

$$\text{logit}(\theta_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\alpha} + U_{ij},$$

여기서 변량효과 $\boldsymbol{U}_i$는 임의의 분포 $f(\boldsymbol{U}_i; \boldsymbol{\tau})$를 따르고 상관계수는 $\text{cor}(U_{i1}, U_{i2}) = \rho_I$라고 가정한다. 변환모수가 λ인 Box-Cox 변환 $g(y, \lambda)$는 양수 $y > 0$에 대해

$$g(y, \lambda) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0, \\ \log(y), & \lambda = 0 \end{cases}$$

이며 이 변환을 이용하여 섭취량 Y_{ij}는 다음과 같은 선형혼합모형을 가정하고 분석한다.

$$g\left(Y_{ij}, \lambda\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} + V_i + \varepsilon_{ij},$$

여기서 V_i는 변량효과로 $V_i \sim N(0, \sigma_V^2)$이고 개체 내 오차는 $\varepsilon_{ij} \sim \text{iid } N(0, \sigma_\varepsilon^2)$이라고 가정한다.

이들 모형의 모수를 추정하고 추정된 모형을 기반으로 개인별로 섭취여부와 섭취량 난수들을 각각 k개씩 생성한다. GLMM 분석 결과를 이용하여 i번째 개체에 대해 섭취여부 난수 $(\hat{I}_{i1}, \ldots, \hat{I}_{ik})$와 Box-Cox 변환 선형혼합모형 분석결과를 이용하여 섭취량 난수 $(w_{i1}, \ldots, w_{ik})$를 생성한다. Box-Cox 변환 척도 하에서의 섭취량을

다음과 같이 Box-Cox 역변환을 이용하여 관측 척도 하에서의 값으로 변환한다.

$$g(w_{ij},\hat{\lambda})^{-1} = \begin{cases} \left(\hat{\lambda}w_{ij}+1\right)^{1/\hat{\lambda}}, & \hat{\lambda}\neq 0 \\ \exp(w_{ij}), & \hat{\lambda}=0, \end{cases}$$

단계 4에서 얻어지는 최종 식이섭취량은 $y_{ij}(\hat{\lambda}) = \hat{I}_{ij}g(w_{ij},\hat{\lambda})^{-1}$로 표시할 수 있다.

위에서 언급한 것과 같이 선형회귀모형의 정규성이나 등분산성을 개선하기 위해 반응변수를 Box-Cox 변환하여 모형을 추정하고 예측값을 구한 후 역변환하여 원자료 척도에서의 예측값을 유도하였다. 이 경우 주의해야 할 점은 $\lambda \neq 1$인 Box-Cox 변환과 역변환은 볼록 또는 오목 함수이기 때문에 변환-역변환(transformation and back-transformation)을 통한 원자료 척도에서의 예측은 편향을 발생시킨다는 것이다. Box-Cox 변환 분석에서의 예측값을 $w(\lambda)$라고 하면 변환-역변환 과정에서 발생하는 편향을 줄인 예측값은 다음과 같이 유도된다.

$$\begin{aligned} y(\lambda) &\approx (\lambda w(\lambda)+1)^{1/\lambda}\left(1+\frac{(1-\lambda)\sigma^2(\lambda)}{2\,(\lambda w(\lambda)+1)^2}\right), \quad \lambda\neq 0, \\ y(0) &= e^{w(0)}\left(1+\frac{\sigma^2(0)}{2}\right), \quad \lambda = 0, \end{aligned}$$

여기서 $\sigma^2(\lambda)$는 Box-Cox 변환 자료의 분산을 의미한다. 기존 식이섭취 분석모형에서는 편향을 줄이기 위해서는 위와 같은 편향 조정이 필요하다.

2.3. EMM

기존 식이섭취모형에서 섭취여부는 일반화선형혼합모형, 섭취량은 Box-Cox 변환을 적용한 선형혼합모형으로 분석하였으나 EMM에서는 섭취여부와 섭취량을 모두 GEE로 분석한다. GEE에서는 이틀 동안의 섭취여부와 섭취량 간의 상관계수를 직접 추정하여 쉽게 활용할 수 있으며 섭취여부와 섭취량을 동일한 모형으로 분석할 수 있다는 장점이 있다. 섭취여부 분석에서는 베르누이 분포를, 섭취량 분석에서는 감마분포를 가정하는데 특별히 감마분포를 가정한 이유는 감마분포의 경우 상수 변동계수, 즉 평균과 표준편차가 비례관계를 가지는데 이는 평균 섭취량이 많은 그룹의 섭취량 변동이 적은 그룹의 변동보다 큰 식이섭취에서의 일반적 현상을 적절히 설명하기 때문이다. 분산상수화변환(variance stabilizing transformation) 관점에서 $\lambda = 0$인 Box-Cox 변환, 즉 로그변환으로 이러한 현상을 설명할 수 있지만 앞에서도 언급한 것과 같이 역변환 과정에서의 편향 문제가 발생한다. 그에 반해 로그연결함수는 모수에 대한 변환으로 역변환하여도 편향이 발생하지 않는다.

GEE 모형을 적용하기 위해서는 상관행렬을 지정해야 하는데 본 연구에서 분석할 국민건강영양조사 자료는 이틀 자료이기 때문에 독립이 아니면 상관행렬은 동일한 구조를 가진다. 분석에서는 교환가능(exchangeable) 상관행렬을 설정하였으며 식품이나 영양소의 섭취여부는 베르누이 분포와 로짓연결함수를, 이틀 모두 섭취한 자료에 대해 감마분포와 로그연결함수를 설정한 GEE로 분석한다. 즉, 섭취확률 θ_i와 평균섭취량 μ_i는 설명변수 $\boldsymbol{x}_i$에 대해 다음과 같은 관계식을 가진다.

$$g_P(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \boldsymbol{\alpha}^T\boldsymbol{x}_i, \quad g_A(\mu_i) = \log(\mu_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i.$$

GEE에서는 다음의 등식을 만족시키는 준가능도(quasi-likelihood) 추정법으로 모수 $\boldsymbol{\alpha}$와 $\boldsymbol{\beta}$를 추정한다.

$$\sum_i\left(\frac{\partial\boldsymbol{\theta}_i}{\partial\boldsymbol{\alpha}}\right)^T\mathbf{V}_P^{-1}(\boldsymbol{I}_i-\boldsymbol{\theta}_i)=\mathbf{0}, \quad \sum_i\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)^T\mathbf{V}_A^{-1}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)=\mathbf{0},$$

여기서 $\mathbf{V}_P$는 $\boldsymbol{I}_i$의 가공분산행렬(working covariance matrix), $\mathbf{V}_A$는 $\boldsymbol{Y}_i$의 가공분산행렬이며 등분산에 교환가능 가상관행렬를 곱한 것으로 이틀 자료분석에서는 다음과 같다.

$$\mathbf{V}_P = \sigma_P^2 \begin{bmatrix} 1 & \rho_P \\ \rho_P & 1 \end{bmatrix}, \quad \mathbf{V}_A = \sigma_A^2 \begin{bmatrix} 1 & \rho_A \\ \rho_A & 1 \end{bmatrix}.$$

섭취여부에 대한 분산은 $\sigma_P^2 = \theta_i(1-\theta_i)$, 섭취량에 대한 분산은 임의의 산포모수 $\phi > 0$에 대해 $\sigma_A^2 = \mu_i^2/\phi$를 가정한다. 이는 섭취여부는 베르누이 분포로, 섭취량은 감마분포로 모형화한 것을 의미하며 섭취량을 분석할 때에는 추가 모수인 ϕ에 대해서도 추정한다. EMM에서도 추정된 모형을 기반으로 개체별로 k개의 모의섭취여부 및 섭취량을 생성하여 일상섭취량을 유도한다. 가상관행렬에서의 ρ_P와 ρ_A, 감마분포의 산포모수 ϕ는 추정해야 할 모수이며 이 추정값은 k개의 모의 섭취자료를 생성할 때 매우 중요한 역할을 한다. 이들 모수에 대한 정보가 있다면 1일 조사자료에 대해서도 이 모수값을 이용하여 모의 섭취량을 생성하고 일상섭취량을 유도할 수 있다.

섭취여부들 간에 상관관계가 있을 때, 다음과 같은 확률분포를 따른다고 가정한다.

$$\begin{aligned} P(I_{ij}=1, I_{ik}=1) &= \theta_i^2 + \theta_i(1-\theta_i)\rho_P, \\ P(I_{ij}=0, I_{ik}=0) &= (1-\theta_i)^2 + \theta_i(1-\theta_i)\rho_P, \\ P(I_{ij}=0, I_{ik}=1) = P(I_{ij}=1, I_{ik}=0) &= \theta_i(1-\theta_i)(1-\rho_P). \end{aligned}$$

이 분포를 이용하면 조사 결과 I_{ij}의 값에 따른 조건부 확률은 다음과 같고 이를 이용하여 k개의 섭취여부 난수를 생성할 수 있다.

$$P(I_{ik}=1 \mid I_{ij}=0) = \theta_i(1-\rho_P), \quad P(I_{ik}=1 \mid I_{ij}=1) = \theta_i + (1-\theta_i)\rho_P,$$

여기서 θ_i와 ρ_p는 GEE 분석에서의 $\hat{\theta}_i = \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\alpha}})/(1+\exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\alpha}}))$와 $\hat{\rho}_P$로 추정한다. 이틀 조사의 경우 조사일은 $j = 1, 2$이므로 각 j에 대해 $k/2$개의 난수를 생성하고 기존 관측자료와 결합하여 $\tilde{I}_i = (I_{i1}, I_{i2}, \tilde{I}_{i1}, \ldots, \tilde{I}_{ik})^T$를 구한다.

다음 단계에서는 해당 식품 또는 영양소의 섭취량에 대한 난수를 생성하는데 위에서도 언급한 것 같이 섭취량은 감마분포를 가정했기 때문에 상관관계가 있는 감마난수를 생성해야 한다. 상관관계가 있는 감마난수는 Carpenter와 Diawara (2007)의 결과를 활용하여 생성할 수 있다. 이들의 결과를 활용하기 위해 먼저 형태모수 α^*와 척도모수 β_i^*가 다음과 같이 설정된 감마분포에서 난수 $\hat{Y}_{ik}$를 일단 생성한다.

$$\alpha^* = \frac{(1-\rho^*)\phi}{1+\rho^*}, \quad \beta_i^* = \frac{\mu_i(1+\rho^*)}{\phi}$$

모의 섭취량은 위에서 유도한 감마난수에 조사자료 Y_{ij}를 상관계수로 가중한 값을 합하여 유도한다. 상관계수 ρ_A가 음수일 때 두 합이 0보다 작을 수 있는 상황을 방지하기 위해 모의섭취량은 다음의 산식으로 유도한다.

$$\tilde{Y}_{ik} = \max(\hat{Y}_{ik} + \rho_* Y_{ij}^*, 0), \quad \rho^* = \text{sign}(\rho_A)\sqrt{|\rho_A|},$$

여기서 Y_{i1}, Y_{i2}가 모두 0보다 크면 $Y_{i1}^* = Y_{i1}, Y_{i2}^* = Y_{i2}$이지만, 하나만 크면 $\max(Y_{i1}, Y_{i2})$를 Y_{i1}^*, Y_{i2}^* 값으로 사용하고 둘 다 0이면 동일 설명변수 그룹 내에서 무작위로 선택한 섭취 개체의 값을 사용한다. 실제 모의 섭취량을 생성할 때, 모수 ρ_A, μ_i, ϕ는 GEE 분석에서의 $\hat{\rho}_A, \hat{\mu}_i = \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}), \hat{\phi}$로 추정한다.

섭취 조사자료와 모의 섭취량 자료를 결합하여 $\tilde{\boldsymbol{Y}}_i = (Y_{i1}, Y_{i2}, \tilde{Y}_{i1}, \ldots, \tilde{Y}_{ik})^T$로 표시하면 섭취여부와 섭취량을 결합한 $\tilde{I}_i\tilde{Y}_i = (Y_{i1}, Y_{i2}, \tilde{I}_{i1}\tilde{Y}_{i1}, \ldots, \tilde{I}_{ik}\tilde{Y}_{ik})^T$가 i번째 개체의 최종 모의 섭취량이 되고 최종 일상섭취량은 다음과 같다.

$$\bar{Y}_i = \frac{Y_{i1} + Y_{i2} + \sum_{j=1}^{k} \tilde{I}_{ij}\tilde{Y}_{ij}}{k+2}.$$

일상섭취분포는 섭취량이 0일 확률이 δ이고 섭취량이 0보다 크면 형태모수가 α_*, 척도모수가 β_8인 감마분포의 혼합분포(mixture distribution)를 가정한다.

$$f_Y(y) = \delta I(y = 0) + (1 - \delta) f_G(y; \alpha_*, \beta_*), \quad y \geq 0. \tag{2.1}$$

전체 모집단의 섭취분포를 유도하는 경우 전체 개체의 일상섭취자료 $\bar{Y}_1, \ldots, \bar{Y}_n$를 이용하여

$$\hat{\delta} = \frac{\sum_{i=1}^{n} I(\bar{Y}_i = 0)}{n}$$

으로 δ를 추정하고 일상섭취자료 중 0보다 큰 자료만을 이용하여 감마분포의 α_*와 β_*를 적률법 또는 최대가능도추정법으로 추정한다. 식품섭취조사의 경우 개별 개체에 대한 가중치가 있기 때문에 적률법을 이용하여 추정하는 것이 상대적으로 용이하다. 특정범주에 속하는 집단의 섭취분포를 구한다면 해당 개체의 일상섭취량 자료만을 이용하여 동일한 방법으로 분포를 구할 수 있다. 영양 및 보건 분야에서는 극단적인 섭취군이 어느 정도 되는지에 관심이 많으며 추정한 일상섭취분포를 이용하여 특정 백분위수를 추정하거나 얼마 이상 또는 이하로 섭취할 비율 등을 추정할 수 있다.

기존 분석모형에서는 한번의 모형 추정과 난수 생성단계를 거치지만 EMM에서는 다음과 같이 붓스트랩 방법을 적용한다.

단계 1. 개체별 가중치를 표집확률로 반영하여 n명의 이틀 섭취자료를 재표집함.

단계 2. 재표집 자료에 대해 GEE로 섭취여부와 섭취량 분석을 실시하여 상관계수, 회귀계수와 섭취량분석의 경우 추가로 산포모수를 추정함.

단계 3. 단계 2에서 추정값을 이용하여 각 재표집 자료에 대해 k 주간의 모의 섭취량을 발생시키고 n개의 일상섭취량을 유도함.

단계 4. n개의 일상섭취량 자료를 이용하여 식 (2.1)의 모수를 추정함.

단계 5. 위의 작업을 B번 반복 수행함.

위와 같은 작업을 통해 얻어진 모수 추정값들은 두 가지로 활용될 수 있다. 첫 번째는 B개의 $\hat{\delta}$, $\hat{\alpha}_*$, $\hat{\beta}_*$를 평균하여 식 (2.1)의 모수의 추정값으로 사용하는 것이다. 이를 통해 하나의 일상섭취분포를 유도하고 이 분포를 기반으로 어떤 식품이나 영양소를 얼마 이상 또는 이하로 섭취할 비율을 계산하거나 분위수를 추정하는 것이다. 두 번째는 단계 4에서 추정된 각 $\hat{\delta}$, $\hat{\alpha}_*$, $\hat{\beta}_*$를 기반으로 섭취분포를 구하고 이 섭취분포에서 섭취비율이나 분위수를 구한다. 이렇게 얻어진 B개의 섭취비율이나 분위수의 평균을 이용하여 해당 모수의 추정값을 구할 수 있으며 이와 더불어 표준오차 및 신뢰구간 등을 함께 구할 수 있다.

3. EMM 상관계수와 일상섭취량 비교

앞 절에서도 언급한 것 같이 모의섭취 여부와 모의 섭취량을 생성하는 과정에서 가장 중요한 역할을 하는 것이 상관계수 추정값인 $\hat{\rho}_P$와 $\hat{\rho}_A$이다. 모의자료를 발생시킬 때, 섭취여부의 조건부 확률이나 섭취량의 감마분포 모수에 영향을 주며 모의 섭취량과 관측자료를 결합할 때에도 이 상관계수가 반드시 필요하다. 이들 상관계수가 있으면 하루 조사 자료를 일반화선형모형(generalized linear model, GLM)으로 분석한 결과에 적용하여 모의섭취 여부와 섭취량을 생성할 수 있다. Lee 등 (2020)의 연구에서도 하루 자료를 이용하여 일상섭취를 유도하는 과정에서 이틀 조사 자료 분석에서 유도한 $\hat{\rho}_P$와 $\hat{\rho}_A$를 이용하여 모의섭취 여부와 모의 섭취량을 발생시켜 일상섭취량을 산출하였다.

기존 분석에서는 섭취자료를 인구생태학적으로 구분하지 않고 상관계수 추정값 $\hat{\rho}_P$와 $\hat{\rho}_A$를 유도하여 사용하였다. 이는 섭취자료를 인구생태학적으로 나누었을 때 각 그룹에서의 상관계수는 유의한 차이가 없다는

Table 2: Comparison of correlation coefficients by gender

Nutrient	Total		Male		Female		*t*-test		KS-test	
	Mean	SD	Mean	SD	Mean	SD	Stat	*p*-value	Stat	*p*-value
EN	0.311	0.024	0.294	0.041	0.317	0.037	−4.174	0.000	0.320	0.000
WATER	0.385	0.025	0.384	0.042	0.387	0.030	−0.579	0.563	0.120	0.468
PROT	0.300	0.024	0.282	0.036	0.310	0.030	−6.018	0.000	0.380	0.000
FAT	0.187	0.025	0.165	0.032	0.190	0.034	−5.207	0.000	0.350	0.000
CHOL	0.150	0.030	0.188	0.046	0.114	0.031	13.301	0.000	0.680	0.000
CHO	0.340	0.024	0.309	0.040	0.338	0.036	−5.520	0.000	0.340	0.000
TDF	0.375	0.026	0.357	0.036	0.365	0.031	−1.762	0.080	0.220	0.016
SUGAR	0.318	0.032	0.295	0.044	0.317	0.044	−3.580	0.000	0.240	0.006
CA	0.297	0.029	0.263	0.040	0.313	0.034	−9.580	0.000	0.490	0.000
PHOS	0.375	0.021	0.328	0.034	0.396	0.029	−15.270	0.000	0.730	0.000
FE	0.241	0.027	0.276	0.044	0.212	0.038	10.986	0.000	0.597	0.000
NA	0.234	0.028	0.239	0.035	0.202	0.031	7.954	0.000	0.430	0.000
K	0.392	0.028	0.358	0.036	0.393	0.036	−6.808	0.000	0.410	0.000
VA-RAE	0.150	0.029	0.209	0.069	0.134	0.038	9.660	0.000	0.640	0.000
CAROT	0.188	0.028	0.184	0.044	0.188	0.040	−0.688	0.492	0.090	0.813
RETIN	0.104	0.031	0.202	0.076	0.049	0.031	18.705	0.000	0.870	0.000
B1	0.223	0.029	0.220	0.051	0.201	0.033	3.076	0.002	0.250	0.004
B2	0.268	0.028	0.217	0.039	0.295	0.044	−13.205	0.000	0.720	0.000
NIAC	0.274	0.027	0.245	0.037	0.289	0.033	−8.724	0.000	0.480	0.000
FOLATE	0.313	0.025	0.334	0.039	0.287	0.033	9.202	0.000	0.560	0.000
VITC	0.183	0.024	0.180	0.032	0.178	0.031	0.421	0.674	0.070	0.967

가정하에서 분석한 것이다. 이 논문에서는 이러한 가정이 적절한지를 알아보기 위해 자료를 남녀별로 나누어 상관계수를 유도하고 그룹 간에 차이가 있는지를 알아본다. EMM에서는 B번 반복 수행하는 과정에서 B개의 $\hat{\rho}_P$와 $\hat{\rho}_A$를 그룹별로 유도할 수 있어 이 값을 비교 분석할 수 있다.

본 논문에서는 2019년 국민건강영양조사 자료 중 주요 영양소에 대해 상관계수를 비교한다. 이 조사자료에 의하면 하루 조사는 5,546명, 이틀 조사는 1,601명, 총 7,147명을 대상으로 조사가 이루어졌다. 이 중 3,202(하루 2,493, 이틀 709)명이 남자, 3,945(하루 3,053, 이틀 892)명이 여자인 것으로 나타났으며 식품영양학에서 주요하게 분석되고 있는 21개가 영양소에 대해 비교분석을 한다. Table 2에 영양소를 영문으로 표시하였는데 EN(에너지), WATER(수분), PROT(단백질), FAT(지방), CHOL(콜레스테롤), CHO(탄수화물), TDF(식이섬유), SUGAR(당), CA(칼슘), PHOS(인), FE(철), NA(나트륨), K(칼륨), VA-RAE(비타민 A), CAROT(카로틴), RETIN(레티놀), B1(티아민), B2(리보플라빈), NIAC(나이아신), FOLATE(엽산), VITC(비타민 C)를 의미한다. 영양소는 식품과 다르게 대부분 섭취하고 있었으며 이틀 자료 3,202개 중 RETIN은 80, CHOL은 56, CAROT는 5개, VA-RAE(와 VITC는 2개, WATER는 1개의 섭취량이 0이었고 나머지 영양소는 모두 0보다 큰 값이었다. 이는 GEE로 영양소 섭취여부를 분석했을 때 회귀모수와 상관계수를 추정하지 못할 수 있어 섭취여부에 대한 확률과 상관계수는 기술통계 형태로 추정하였다. 설명변수로 사용 가능한 인구생태학적으로 변수로는 sex(성), age(연령), region(17개 시도), town(시군)가 포함되어 있으며 영양 섭취의 정도는 연령대에 따라 차이가 있어 이를 생애주기별(0, 5, 9, 18, 29, 39, 49, 59, 69, 100)로 나누어 ageG라는 요인으로 처리하였다.

Table 2은 21개 영양소의 이틀 섭취자료에 대해 EMM으로 유도한 $B = 100$개 상관계수 $\hat{\rho}_A$를 비교한 결과이다. Total로 표시된 열은 남녀 구분하지 않고 전체 자료에 대해 선형예측식 'sex*ageG+region+town'으로

분석한 상관계수의 평균(Mean)과 표준편차(SD)이다. Male과 Female의 결과는 자료를 남녀별로 나누었기 때문에 선형예측식은 'ageG+region+town'으로 분석한 결과이다. 전체 자료 분석모형에서 'sex*ageG'는 sex와 ageG의 주효과와 상호작용항을 포함한 것을 의미하며 성별에 연령대의 섭취 패턴이 다를 수 있음을 설명하기 위한 것이고 남녀별로 나눈 모형에서 ageG의 효과가 다르게 나올 수 있는 부분과 연계된다. 상관계수의 비교분석에서는 단순히 평균이 같은지 검정하는 것와 더불어 전반적인 분포에도 차이가 있는지를 알아보기 위해 Kolmogorov-Smirmov-test (KS-test)를 추가하였다.

Table 2에 의하면 5% 유의수준에서 WATER, CAROT, VITC를 제외한 나머지 영양소 섭취량의 상관계수가 남녀별로 차이가 있는 것으로 나타났다. TDF의 경우 KS-검정에서만 유의한 차이가 없는 것으로 분석되었다. 반복수 B가 커지면 상관계수의 평균이나 분포가 조금이나도 차이가 있다면 같다는 가설을 기각할 확률이 높아지기 때문에 추가로 $B = 100$개 결과 중 무작위로 30개를 뽑아 비교해 보았다. 그 결과 TDF에 대한 t-검정 p-값이 0.060, KS-검정의 p-값이 0.035로 반대의 결과가 나왔고 CHO의 경우 KS-test의 p-값이 0.071로 차이가 없는 것으로 바뀌었을 뿐 나머지 영양소에는 모두 $B = 100$일 때와 같은 결론을 얻었다.

영양소 섭취량의 상관관계가 성별에 따라 차이가 있는 것으로 나타나 이를 반영하여 일상섭취분포를 유도한 것과 남녀를 구분하지 않고 전체 자료를 이용한 상관계수로 유도한 일상섭취분포 간에 차이가 있는지를 알아보았다. 이 비교에서는 관측수가 상대적으로 많은 하루 섭취자료를 이용했으며 하루 섭취자료에 EMM을 적용하기 위해 GEE 대신 GLM으로 회귀계수를 추정하였다. 이를 EMM의 단계와 비교하면 다음과 같이 수행한다.

단계 1. 개체별 가중치를 표집확률로 반영하여 n명의 하루 섭취자료를 재표집함.

단계 2. 재표집 자료에 대해 GLM으로 섭취여부와 섭취량 분석을 실시하여 회귀계수와 섭취량분석의 경우 추가로 산포모수를 추정함.

단계 3. 단계 2에서 추정값을 이용하여 각 재표집 자료에 대해 $k + 1$ 주간의 모의 섭취량을 발생시키고 n개의 일상섭취량을 유도함. 이때 상관계수는 이틀 자료를 EMM으로 분석하여 유도한 상관계수 집합에서 무작위로 선택함.

단계 4. n개의 일상섭취량 자료를 이용하여 식 (2.1)의 모수를 추정함.

단계 5. 위의 작업을 B번 반복 수행함.

각 반복에서 구한 일상섭취분포 (2.1)에서 평균, 5백분위수와 95백분위수를 B개 유도하고 이를 이용하여 비교하였다. 이때 각 평균과 100p백분위수는 다음과 같이 유도된다.

$$\hat{\mu} = (1 - \hat{\delta})\hat{\alpha}_*\hat{\beta}_*, \quad \hat{q}_p = \begin{cases} 0, & p \le \hat{\delta} \\ q(p, \hat{\delta}, \hat{\alpha}_*, \hat{\beta}_*), & p > \hat{\delta}, \end{cases} \tag{3.1}$$

여기서 $q(p, \hat{\delta}, \hat{\alpha}_*, \hat{\beta}_*)$는 형태모수가 $\hat{\alpha}_*$이고 척도모수가 $\hat{\beta}_*$인 감마분포에서의 $(p-\hat{\delta})/(1-\hat{\delta})$ 분위수를 의미한다. 이렇게 구한 평균과 백분위수의 평균을 Tables 3–6에 표시하였다.

표에서 M1로 표시된 부분은 위의 단계 1 과정에서 얻은 재표집 자료를 남녀별로 나누고 각 자료에 대해 다음 단계별 작업을 진행한 결과이다. 이때 상관계수는 위에서 유도한 남녀별 상관계수 집합에서 무작위로 선택하며 선형예측식이 'ageG+region+town'인 GLM으로 모수를 추정한다. Total에 해당되는 부분은 각 반복에서 유도된 남녀 모의 섭취량을 하나로 병합하여 도출한 결과이다. 이 방법을 분리-병합 방법이라고 부르겠다. M2로 표시된 부분은 단계 1 과정에서 남녀별로 나누지 않고 전체 자료에 대해 선형예측식이 'sex*ageG+region+town'인 GLM으로 모수를 추정한다. 이때 상관계수는 Table 2의 Total을 계산할 때 사용했던 상관계수 집합에서 무작위로 선택하여 사용한다. M2의 Male과 Female로 표시된 부분은 이 모의 섭취량

Table 3: Comparison of average usual intake(B = 100)

Nutrient	Total				Male				Female			
	M1	M2	t-test	W-test	M1	M2	t-test	W-test	M1	M2	t-test	W-test
EN	1915	1917	0.004	0.005	2243	2243	0.339	0.495	1588	1588	0.533	0.393
WATER	950	951	0.005	0.021	1027	1027	0.152	0.198	874	874	0.939	0.911
PROT	71.1	71.2	0.000	0.000	83.5	83.5	0.927	0.895	58.8	58.8	0.914	0.816
FAT	47.9	48.0	0.000	0.000	56.1	56.1	0.232	0.363	39.7	39.7	0.590	0.827
CHOL	267	267	0.004	0.002	307	307	0.005	0.003	227	228	0.001	0.005
CHO	275	276	0.000	0.000	310	310	0.988	0.960	241	241	0.165	0.216
TDF	22.8	22.8	0.003	0.002	24.6	24.6	0.875	0.919	21.0	21.0	0.068	0.080
SUGAR	60.3	60.3	0.001	0.002	63.8	63.8	0.089	0.133	56.8	56.8	0.248	0.272
CA	496	496	0.006	0.014	542	542	0.240	0.246	450	450	0.549	0.367
PHOS	1038	1039	0.000	0.000	1183	1183	0.300	0.330	893	893	0.538	0.605
FE	11.4	11.5	0.005	0.009	13.0	13.0	0.000	0.001	9.85	9.86	0.191	0.244
NA	3290	3294	0.000	0.000	3894	3894	0.601	0.748	2685	2686	0.017	0.013
K	2679	2682	0.000	0.000	2947	2948	0.312	0.308	2411	2412	0.193	0.182
VA-RAE	373	374	0.004	0.007	405	405	0.034	0.024	342	341	0.034	0.041
CAROT	2585	2586	0.026	0.083	2814	2813	0.226	0.318	2355	2355	0.469	0.581
RETIN	157	157	0.693	0.914	169	169	0.026	0.003	144	145	0.001	0.001
B1	1.308	1.309	0.001	0.002	1.523	1.523	0.957	0.873	1.092	1.092	0.364	0.402
B2	1.606	1.607	0.015	0.010	1.819	1.819	0.881	0.968	1.393	1.393	0.676	0.696
NIAC	13.1	13.1	0.000	0.000	15.2	15.2	0.220	0.300	10.9	10.9	0.456	0.152
FOLATE	291	291	0.031	0.012	314	314	0.493	0.428	268	268	0.136	0.161
VITC	62.0	62.0	0.614	0.785	62.9	62.9	0.003	0.003	61.1	61.2	0.000	0.000

자료를 단순히 남녀별로 나누어 얻은 결과이다. 이를 전체-분리 방법이라고 부르겠다. 동일한 자료에 대해 분리-병합 방법과 전체 분리 방법을 적용했기 때문에 두 결과를 비교하기 위해서는 쌍을 이룬 자료로 평가해야 한다. 표에서 t-test와 W-test로 표시된 것은 100개의 쌍을 이룬 자료에 대해 t-검정과 Wilcoxon 분호순위검정을 통해 계산된 p-값이다.

Table 3는 B = 100개의 섭취분포의 평균 $\hat{\mu}$들에 대한 M1과 M2의 비교 결과이다. 절대적인 값을 비교해 보면 M1과 M2 간 $\hat{\mu}$의 평균은 큰 차이가 없는 것으로 보이며 Total의 결과를 보면 RETIN과 VITC를 제외한 나머지 영양소에서 대부분의 영양소에서 통계적으로 유의한 차이가 있다고 분석되었다. 남녀별로 나누어 비교했을 때에는 CHOL, RETIN, VITC에서 유의한 차이가 있을 것으로 분석되었다. 재표집 반복수 B가 크면 클수록 약간의 차이에도 유의한 차이가 있다고 할 가능성이 높아지므로 B = 30로 만들어 추가 분석을 해 본 결과 Table 4에 표시된 것처럼 B = 100일 때와 B = 30일 때의 평균은 유사하며 M1과 M2의 비교에서 전체에 대한 PROT, FAT, CHO, PHOS, NA, K, B1, FOLATE에 유의한 차이가 있고 나머지 부분에서는 차이가 없는 것으로 분석되었다. 하지만 섭취량의 절대적인 차이는 크게 느껴지지 않는다.

Tables 3과 4에서 평균섭취량에 대해 알아 봤는데 영양학적 관점에서 특정 영양소나 식품을 너무 과소 또는 과잉 섭취하는 경우 문제가 발생할 수 있기 때문에 극단적인 섭취에 대한 추정이 더욱 중요할 수 있다. Tables 5와 6는 식 (3.1)에 표시된 B = 100개의 $\hat{q}_{0.05}$와 $\hat{q}_{0.95}$에 대한 평균과 비교 분석한 결과를 표시한 것이다.

Table 5를 보면 평균 비교과 다르게 5백분위수에서는 많은 부분에서 통계적으로 유의한 차이가 있는 것으로 분석되었으며 절대 섭취량에도 차이가 있는 것을 볼 수 있다. 또한, 남녀별로 나누어 M1과 M2를 비교했을 때 유의한 차이가 있는 영양소가 남녀별로 다른 것으로 나타났다. 특이한 점은 Table 2에서 남자의 상관계수가 전체 상관계수보다 유의하게 큰 영양소의 경우 M1의 값이 M2보다 작고, 상관계수가 작으면 M1의 값이 M2

Table 4: Comparison of average usual intake(B = 30)

Nutrient	Total				Male				Female			
	M1	M2	t-test	W-test	M1	M2	t-test	W-test	M1	M2	t-test	W-test
EN	1913	1913	0.768	0.761	2240	2240	0.485	0.490	1586	1587	0.069	0.052
WATER	950	950	0.110	0.318	1027	1027	0.136	0.119	873	873	0.890	0.952
PROT	71.1	71.2	0.002	0.002	83.3	83.3	0.296	0.221	58.9	58.8	0.062	0.050
FAT	47.8	47.9	0.033	0.045	56.0	56.0	0.319	0.490	39.7	39.7	0.352	0.428
CHOL	268	268	0.610	0.371	307	307	0.336	0.271	229	229	0.113	0.146
CHO	275	276	0.003	0.005	309	309	0.085	0.080	242	242	0.337	0.477
TDF	22.8	22.8	0.047	0.055	24.5	24.5	0.838	0.871	21.0	21.0	0.321	0.700
SUGAR	60.4	60.4	0.289	0.477	63.9	63.9	0.290	0.393	56.9	56.8	0.107	0.129
CA	495	495	0.426	0.715	540	540	0.945	0.792	451	451	0.684	0.584
PHOS	1038	1039	0.011	0.012	1184	1184	0.645	0.777	893	893	0.857	0.952
FE	11.4	11.5	0.476	0.371	13.0	13.0	0.027	0.040	9.9	9.8	0.785	0.685
NA	3284	3290	0.005	0.005	3895	3895	0.854	0.655	2673	2675	0.026	0.040
K	2676	2678	0.005	0.004	2941	2942	0.150	0.124	2411	2411	0.374	0.465
VA-RAE	375	375	0.115	0.184	406	406	0.134	0.198	343	343	0.289	0.253
CAROT	2580	2580	0.922	0.903	2810	2808	0.157	0.198	2349	2350	0.456	0.598
RETIN	157	157	0.657	0.516	170	169	0.063	0.084	145	145	0.235	0.328
B1	1.308	1.310	0.012	0.015	1.522	1.523	0.151	0.135	1.094	1.094	0.897	0.855
B2	1.608	1.608	0.467	0.404	1.821	1.821	0.851	0.452	1.394	1.395	0.753	0.935
NIAC	13.1	13.1	0.329	0.205	15.2	15.2	0.224	0.221	10.9	10.9	0.544	0.935
FOLATE	291	291	0.024	0.013	314	314	0.266	0.191	268	268	0.883	0.685
VITC	61.9	61.9	0.277	0.382	63.0	63.0	0.331	0.221	60.8	60.9	0.046	0.088

보다 큰 것을 볼 수 있다. 이는 여자의 경우에도 동일하다. 이와 반대로 95백분위수에 대한 결과인 Table 6에서는 남자의 상관계수가 전체 상관계수보다 유의하게 큰 영양소의 경우 M1의 값이 M2보다 크고, 상관계수가 작으면 M1의 값이 M2 보다 작을 것으로 나타났으며 여자의 경우에도 동일한 결과를 보였다. 이를 통해 상관계수가 섭취량에 영향을 주고 있으며 M1에서 유도된 남녀의 상관계수와 M2에서 유도된 전체 상관계수에 유의한 차이가 있는 경우 평균 섭취량보다는 과잉 또는 과소 영역에서의 M1 결과와 M2의 결과에 차이가 있는 것으로 분석되었다.

4. 결론

특정 영양소를 과잉 섭취하거나 결핍이 있는 경우 건강에 중대한 영향을 미칠 수 있어 일상섭취량을 정확히 추정하는 것은 영양이나 보건 정책 수립 과정에서 중요한 부분이다. 일상섭취 분석을 하기 위해서는 장기간의 섭취자료가 필요하지만, 국민건강영양조사 식품섭취조사에서는 하루 또는 비연속 이틀 자료가 수집되고 있다. 자료수집의 한계로 일상섭취분석에서는 이틀 자료의 상관관계를 이용하여 장기간의 모의 섭취량를 생성하고 이를 기반으로 일상섭취분포를 추정하고 있다. 일상섭취분포를 유도하는 방법 중 EMM 방법은 섭취자료에 대해 붓스트랩 방법을 적용하여 재표집 하는 과정을 거치는데 이 과정에서 여러 개의 상관계수를 유도할 수 있다. 이 논문에서는 EMM 분석과정에서 재표집된 섭취자료를 남녀별로 나누어 상관계수를 유도하고 남녀 간에 섭취량의 상관계수에 차이가 있는지를 알아보았다. 2019년 국민건강영양조사 영양소 자료에 대해 비교한 결과 많은 영양소에서 남녀 간 상관계수가 다른 것으로 분석되었다. 이렇게 다른 상관계수를 적용했을 때와 남녀별로 나누지 않고 유도한 상관계수를 적용했을 때 일상섭취분포를 비교하였으며 5백분위수와 95백분위수와 같이 과소 또는 과잉 섭취 영역에서 절대적인 추정값에 차이가 있었고 통계적으로도

Table 5: Comparison of the 5th-percentile usual intake($B = 100$)

Nutrient	Total				Male				Female			
	M1	M2	*t*-test	W-test	M1	M2	*t*-test	W-test	M1	M2	*t*-test	W-test
EN	1019	1012	0.000	0.000	1325	1306	0.000	0.000	976	979	0.057	0.082
WATER	400	398	0.000	0.000	442	442	0.926	0.941	378	372	0.000	0.000
PROT	33.2	32.9	0.000	0.000	43.4	42.5	0.000	0.000	31.0	31.3	0.000	0.000
FAT	18.8	18.6	0.000	0.000	24.3	23.8	0.000	0.000	17.6	17.7	0.836	0.727
CHOL	102	103	0.001	0.005	130	136	0.000	0.000	90.1	83.9	0.000	0.000
CHO	163	161	0.000	0.000	199	196	0.000	0.000	149	149	0.426	0.187
TDF	10.2	10.1	0.000	0.000	11.9	11.8	0.000	0.001	8.96	8.87	0.112	0.048
SUGAR	24.5	24.2	0.000	0.000	26.9	26.3	0.000	0.000	22.7	22.8	0.678	0.764
CA	255	253	0.000	0.000	304	292	0.000	0.000	223	230	0.000	0.000
PHOS	543	538	0.000	0.000	683	662	0.000	0.000	486	497	0.000	0.000
FE	5.59	5.66	0.000	0.000	6.73	6.98	0.000	0.000	5.35	5.17	0.000	0.000
NA	1539	1544	0.009	0.012	2025	2065	0.000	0.000	1491	1431	0.000	0.000
K	1269	1264	0.000	0.000	1592	1574	0.000	0.000	1046	1048	0.586	0.756
VA-RAE	135	147	0.000	0.000	138	165	0.000	0.000	148	140	0.000	0.000
CAROT	835	836	0.857	0.555	938	948	0.234	0.184	787	768	0.003	0.009
RETIN	20.7	30.9	0.000	0.000	15.6	34.8	0.000	0.000	42.5	30.0	0.000	0.000
B1	0.645	0.642	0.000	0.000	0.825	0.827	0.417	0.374	0.612	0.597	0.000	0.000
B2	0.789	0.776	0.000	0.000	0.996	0.947	0.000	0.000	0.685	0.708	0.000	0.000
NIAC	5.90	5.86	0.000	0.000	7.65	7.48	0.000	0.000	5.39	5.51	0.000	0.000
FOLATE	128	128	0.438	0.256	142	146	0.000	0.000	120	115	0.000	0.000
VITC	18.6	19.3	0.000	0.000	18.8	20.0	0.000	0.000	18.6	18.7	0.811	0.887

유의한 차이가 있는 것으로 나타났다. 자료를 남녀별로 나누어 유도한 상관계수에 차이가 있다는 것은 남녀별로 섭취 패턴에 차이가 있다는 것으로 이를 반영한 분석방법을 적용하는 것이 자료의 특성에 더 부합되는 것으로 판단된다. 향후 연구에서는 연령대와 같이 좀 더 세밀한 인구생태학적 기준으로 나누었을 때와 상대적으로 섭취빈도가 낮은 식품군을 분석할 때와 같이 분석자료가 적어 GEE이 분석이 되지 않는 문제를 해결하는 방법을 알아볼 것이다.

Table 6: Comparison of the 95th percentile usual intake($B = 100$)

Nutrient	Total				Male				Female			
	M1	M2	*t*-test	W-test	M1	M2	*t*-test	W-test	M1	M2	*t*-test	W-test
EN	3040	3055	0.000	0.000	3357	3386	0.000	0.000	2320	2317	0.078	0.109
WATER	1691	1695	0.000	0.000	1809	1808	0.752	0.816	1537	1546	0.000	0.000
PROT	121	121	0.000	0.000	134	136	0.000	0.000	93.8	93.4	0.000	0.000
FAT	87.7	88.3	0.000	0.000	98.6	100	0.000	0.000	68.9	68.9	0.683	0.569
CHOL	496	495	0.022	0.054	544	532	0.000	0.000	416	429	0.000	0.000
CHO	412	415	0.000	0.000	441	445	0.000	0.000	352	352	0.339	0.118
TDF	39.5	39.6	0.000	0.000	41.0	41.2	0.000	0.000	37.2	37.4	0.061	0.015
SUGAR	109	109	0.000	0.000	113	115	0.000	0.000	103	103	0.341	0.422
CA	801	805	0.000	0.000	836	854	0.000	0.000	741	729	0.000	0.000
PHOS	1663	1674	0.000	0.000	1794	1827	0.000	0.000	1401	1384	0.000	0.000
FE	19.0	18.9	0.000	0.000	21.0	20.6	0.000	0.000	15.5	15.8	0.000	0.000
NA	5575	5578	0.288	0.218	6260	6193	0.000	0.000	4165	4262	0.000	0.000
K	4515	4529	0.000	0.000	4642	4670	0.000	0.000	4240	4238	0.672	0.900
VA-RAE	709	685	0.000	0.000	790	735	0.000	0.000	602	618	0.000	0.000
CAROT	5120	5123	0.497	0.782	5514	5488	0.161	0.149	4610	4650	0.004	0.012
RETIN	407	367	0.000	0.000	481	391	0.000	0.000	298	337	0.000	0.000
B1	2.16	2.17	0.000	0.000	2.40	2.39	0.322	0.292	1.69	1.71	0.000	0.000
B2	2.66	2.68	0.000	0.000	2.84	2.92	0.000	0.000	2.30	2.27	0.000	0.000
NIAC	22.5	22.6	0.000	0.000	24.9	25.2	0.000	0.000	17.9	17.7	0.000	0.000
FOLATE	508	508	0.582	0.803	540	533	0.000	0.000	465	472	0.000	0.000
VITC	126	125	0.000	0.000	128	125	0.000	0.000	124	124	0.825	0.785

References

Carpenter M and Diawara N (2007). A multivariate gamma distribution and its characterizations, *American Journal of Mathematical and Management Sciences*, **27**, 499–507.

Chi SA, Lee H, Lee JE, Lee HS, Kim K, and Yeo IK (2023). An ensemble method based on marginal-effect models (EMM) for estimating usual food intake from single-day dietary data and internal/external two-day dietary data, *European Journal of Clinical Nutrition*, **77**, 325–334.

Kim CI, Jang YA, Moon JJ, Lee HS, Lee Y, Lee HJ, and Kim D (2008). *Dietary Intake Survey of infants, children and adolescents*, Korea Health Industry Development Institute, Available from: https://scienceon.kisti.re.kr/srch/selectPORSrchReport.do?cn=TRKO201000014836

Kim CI, Yon M, Lee HS, Kim D, Lee JY, Nam JW, and Lee Y (2013). *Dietary Intake Survey of Physiologically Vulnerable Population for Risk Assessment*, Korea Food & Drug Administration, Available from: https://doi.org/10.23000/TRKO201400011597

ISU. Available onelin: https://www.side.stat.iastate.edu/pc-side.php

Laureano GHC, Torman VBL, Crispim SP, Dekkers ALM, and Camey SA (2016). Comparison of the ISU, NCI, MSM, and SPADE methods for estimating usual intake: a simulation study of nutrients consumed daily, *Nutrients*, **8**, 1–11.

Lee HS, Yeo IK, Lee Y *et al.* (2016). *A Preliminary Study of red and processed meat consumption for development of food guide serving*, Ministry of Food and Drug Safety, Available from: https://doi.org/10.23000/TRKO201700017647

Lee JE, Yeo IK, and Kim DW (2020). *Dietary Assessment form 2-day 24 Hour Recalls and Estimation of Usual Intake in the Korea National Health and Nutrition Examination Survey(KNHANES)*, Korea Disease Control and Prevention Agency.

MSM. Available onelin: https://msm.dife.de/

NCI. Available onelin: https://epi.grants.cancer.gov/diet/usualintakes/method.html

SPADE. Available onelin: https://www.rivm.nl/en/spade

Souverein OW, Dekkers AL, Geelen A *et al.* (2011). Comparing four methods to estimate usual intake distributions, *European Journal of Clinical Nutrition*, **65**, S92-–S101.

Yeo IK (2023). Estimation method of usual intake based on resampling and electronic device thereof. Patent number: 10-2508934, Available from: https://doi.org/10.8080/1020200153184

Received March 06, 2025; Revised April 02, 2025; Accepted April 06, 2025

성별 식이 섭취의 상관계수와 일상섭취분포 비교

여인권[1,a]

[a]숙명여대 통계학과

요 약

일상섭취량은 어떤 식품이나 영양소를 오랜 기간 얼마나 섭취하는지를 평가하는 장기간의 평균 섭취량으로 영양이나 보건 정책을 수립하는데 중요한 기초 자료이다. 일상섭취량을 구하기 위해서는 장기간의 섭취량 자료가 필요한데 이를 직접 수집하는 대신 모형을 통해 장기간의 모의 섭취량을 생성하여 분석하고 있다. 모의 섭취량을 생성하기 위해서는 이틀 이상의 섭취자료가 있어야 하며 이 자료를 이용하여 섭취량의 상관계수를 유도한다. 상관계수는 모의 섭취량을 생성하는 과정에서 핵심적인 역할을 한다. 기존 연구에서는 자료를 성별과 같이 인구생태학적으로 나누지 않고 전체를 분석하여 상관관계를 유도하고 있다. 이 논문에서는 성별에 따라 자료를 나누어 구한 상관계수들 간에 차이가 있는지를 확인해 보고 전체 자료의 상관계수를 이용하여 유도한 일상섭취분포와 남녀별로 나눈 자료의 상관계수를 이용하여 유도한 일상섭취분포를 비교해 보았다. 2019년 국민건강영양조사의 영양소 자료를 실증분석하여 상관계수와 일상섭취분포에서 유의한 차이가 있는 것을 확인하였다.

주요용어: EMM, 일반화선형모형, 일반화추정방정식, 일상섭취

[1](04310) 서울특별시 용산구 청파로47길 100, 숙명여대 통계학과. E-mail: inkwon@sookmyung.ac.kr

The Korean Journal of Applied Statistics
2025, Vol. 38, No. 6, 823–833

DOI: https://doi.org/10.5351/KJAS.2025.38.6.823
Print ISSN1225-066X / Online ISSN 2383-5818

Forecasting peak demand for electricity using time series models and artificial intelligence models

Sukhui Kwon[a], Wan-Sup Cho[b], In-Kwon Yeo[1,c]

[a]Interdisciplinary Program in Big Data, Chungbuk National University;
[b]Management Information Systems, Chungbuk National University;
[c]Department of Statistics, Sookmyung Women's University

Abstract

Accurate power demand forecasting is essential for maximizing the efficiency of power grid operations. In this paper, daily peak electricity demands are predicted by using time-series models and artificial intelligence (machine learning and deep learning) models and their performances are compared. The analysis reveals that the XGBoost method outperforms others for one-step-ahead forecasts, while deep learning models, LSTM and GRU, exhibit superior performance for forecasts from two to seven steps ahead. For time-series models, predictions based on bias-corrected log-transformation and back-transformation methods yield better results than those using raw data.

Keywords: artificial intelligence models, forecasting maximum electricity demand, time-series models

1. 서론

현대 사회에서 전력은 가정의 일상생활부터 대규모 산업 활동에 이르기까지 필수적인 기반 자원으로, 국가 경제의 안정적 운영을 뒷받침하는 핵심 인프라이다. 전력의 특성상 대규모 저장이 어렵고 실시간 수요에 맞춰 공급되어야 하므로, 수요 예측의 정확성은 전력 운영 체계의 안정성과 효율성을 좌우하는 핵심 요소로 작용한다.

전력수요는 경제 성장, 계절 및 기후 변화, 기온, 인구 증가 등 다양한 변수에 의해 복합적으로 영향을 받으며, 특히 최근 들어 태양광 및 풍력 발전 등 변동성이 큰 신재생 에너지의 도입이 확대됨에 따라 전력수요 예측의 중요성은 더욱 부각되고 있다. 수요를 과소 예측할 경우 전력 공급 부족으로 인한 정전, 사회적 혼란, 경제적 손실이 발생할 수 있으며, 반대로 수요를 과대 예측할 경우 불필요한 발전으로 인한 자원 낭비와 높은 경제적 비용이 초래될 수 있다. 이러한 문제를 최소화하기 위해, 정확하고 신뢰할 수 있는 전력수요 예측 모델의 개발은 필수적이다.

Kwon 등 (2021)은 시계열모형인 계절지수평활법, 계절 ARIMA, AR-ARCH, AR-GARCH 회귀모형을 이용하여 최대 전력수요를 예측하는 방법을 연구하였다. 또한 추세를 갖는 시계열모형과 기온에 대한 회귀모형을 적절한 가중치로 예측 정확도를 높힌 가중평균모형도 연구하였다. 이 연구에서는 외부 변수(기온, 주말, 공휴일)를 고려한 AR-GARCH 회귀모형으로 예측하는 것이 가장 우수한 것으로 나타났다. 일별 최대 전력수요는 전력 공급 설계의 기준이 되는 주요 지표로, 이에 대한 정확한 예측은 전력 계통의 안정성 확보와

[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-Gu, Seoul 04310, Korea. E-mail: inkwon@sookmyung.ac.kr

Published 31 December 2025 / journal homepage: http://kjas.or.kr
© 2025 The Korean Statistical Society. All rights reserved.

효율적인 에너지 운영에 직결된다. 이에 따라 시계열모형 뿐만 아니라 인공지능 기반의 다양한 머신러닝 및 딥러닝모형을 전력수요 예측에 고려하면 더욱 신뢰성 높은 전력수요를 예측할 수 있으리라 기대된다.

본 연구는 2010년 1월부터 2024년 12월까지의 한국전력거래소 제공 일별 최대 전력수요 데이터를 다양한 전력수요 예측 기법으로 예측한 결과를 비교하는 데 목적이 있다. 비교 대상 모형으로 통계적 시계열 분석 모형인 자기회귀오차모형과 ARE-GARCH 모형, 기계학습 모형인 Random Forest, SVR, XGBoost와 딥러닝 모형 중 LSTM와 GRU을 고려하였으며 2절에서 이들 모형에 대해 간단히 서술하였다. 3절에서는 최대 일별 전력 사용량을 실증 분석을 위해 자료의 특징에 대해 알아보고 이에 부합한 설명변수 설정과 분석모형의 예측 결과를 비교하였다.

2. 전력수요 예측모형

Y_t는 t 시점에서의 최대전력수요이고 $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})^T$는 그 시점에서 Y_t에 영향을 줄 것으로 예상되는 설명변수라고 하자. 최대전력수요 자료는 시계열 자료이기 때문에 $\boldsymbol{x}_t$에는 이전의 최대전력수요가 포함되며 구체적 변수는 3절에서 설명한다. 이 논문에서 t 시점까지 관측된 자료를 이용하여 $t+h$ 시점의 최대전략수요를 예측하기 위해 다음과 같은 통계적 모형과 인공지능 모형을 적용해 본다. 모형 비교를 위해 $\boldsymbol{x}_t$는 동일한 변수로 설정하며 자기회귀오차모형과 ARE-GARCH 모형은 모수 추론 과정에 분포 가정이 들어가기 때문에 이 두 모형에 대해서는 로그변환을 추가로 적용하여 예측력을 비교하였다.

2.1. 자기회귀오차모형

자기회귀오차모형(Autoregressive Error Model; ARE)은 시계열 자료에 대해 회귀분석을 했을 때 오차항이 자기상관을 가지는 경우 적용되는 모형이다. Kwon 등 (2021)의 연구에 의하면 최대전력수요는 이전 최대전력수요와 기온, 주말 및 공휴일 여부에 영향을 받는 것으로 나타났으며 이들 변수를 설명변수로 설정한 회귀분석의 잔차들이 자기상관관계를 가지는 것으로 분석되었다. 이러한 자료의 특성을 모형화한 것이 ARE로 다음과 같은 관계식을 가진다.

$$
\begin{aligned}
Y_t &= \beta_0 + \boldsymbol{x}_t^T \boldsymbol{\beta} + \varepsilon_t, \\
\varepsilon_t &= \phi_1 \varepsilon_{t-1} + \cdots + \phi_k \varepsilon_{t-k} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2),
\end{aligned}
$$

여기서 k는 AR 차수이고 ϕ_i는 자기회귀 계수로 정상성을 만족하기 위해서는 $|\phi_i| < 1$의 조건을 만족해야 한다.

위 모형에서 최대전력수요 Y_t 대신 $W_t = \log(Y_t)$를 반응변수로 사용하는 경우, 일반적으로 모수를 추정한 후 추정모형 식에서 유도된 예측값 $\hat{W}_t$를 역변환하여 $\hat{Y}_t = \exp(\hat{W}_t)$으로 최대전력수요를 예측한다. 하지만 이 예측값은 과소예측하는 편향이 발생하는데 이 편향을 줄이기 위해 다음과 같은 방법으로 보정해 주어야 한다.

$$
\hat{Y}_t = \exp\left(\hat{W}_t + \frac{\hat{\sigma_W^2}}{2} \right), \tag{2.1}
$$

여기서 $\hat{\sigma_W^2}$는 로그척도 하에서의 분산추정량으로 추정회귀모형에서의 MSE가 사용된다.

2.2. ARE-GARCH 회귀모형

Engle (1982)은 시계열의 변동성 군집(volatility clustering) 특징을 모형화하기 위해 자기회귀조건부이분산(Autoregressive Conditional Heteroscedasticity; ARCH)모형을 제안하였다. 이후 Bollerslev (1986)은 ARCH 모형에 과거 변동성을 포함시켜 더 간결하지만 변동성이 지속적으로 군집화되는 금융 시계열 특성을 잘 설명하는 GARCH (Generalized ARCH) 모형을 소개하였다. Kwon 등 (2021)에서 비교한 모형 중 자기회귀오

차모형에 GARCH를 적용한 모형이 가장 설명력이 높은 것으로 분석 되었는데 본 논문에서도 아래와 같은 모형식의 ARE-GARCH(1,1) 모형을 적용한다.

$$
\begin{aligned}
Y_t &= \beta_0 + \boldsymbol{x}_t^T\boldsymbol{\beta} + \varepsilon_t, \\
\varepsilon_t &= \phi_1\varepsilon_{t-1} + \cdots + \phi_k\varepsilon_{t-k} + \epsilon_t, \quad \epsilon_t \sim N(0,\sigma^2), \\
\epsilon_t &= \sqrt{h_t}e_t, \quad e_t \sim N(0,1), \\
h_t &= \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \alpha_2 h_{t-1},
\end{aligned}
$$

여기서 h_t는 t 시점에서의 조건부 분산을 의미한다. ARE-GARCH 모형에 대해서도 로그변환을 적용하는 경우 원래 척도의 예측값은 (2.1)의 식을 통해 구한다.

2.3. Random forest 모형

Random Forest 모형은 Breiman (2001)에 의해 개발된 앙상블 기법으로 배깅(bagging) (Breiman, 1996) 접근 방식을 기반으로 한다. 이 방법은 분류 및 회귀문제에 주로 사용되며 붓스트랩(boostrap) 방법으로 K개의 의사결정나무를 유도한다. 각 의사결정나무의 노드에서 분지 할 때, 모든 설명변수(특성변수) p개를 사용하는 배깅과 달리 무작위로 선택한 $m(< p)$개의 일부 설명변수만 사용하여 속도를 개선시키면서 설명력이 큰 특정 변수의 영향력을 줄이고 의사결정나무 간 관계가 줄어들어 모델의 다양성과 예측 성능을 높인다. 일반적으로 $m = \sqrt{p}$를 사용한다 (Hastie 등, 2009; James 등, 2015). 이렇게 만들어진 K개의 의사결정나무에 대해 예측값을 얻고 이들의 평균으로 예측값을 구한다.

2.4. XGBoost 모형

Chen과 Guestrin (2016)에 의해 개발된 XGBoost (eXtreme Gradient Boosting) 모형은 약한 분류기를 순차적으로 개선해나감으로써 보다 강력한 분류기를 생성하는 트리모형에 경사도 부스팅(gradient boosting) 기법을 적용한 앙상블 알고리즘 중 하나이다. XGBoost는 병렬학습이 가능해 경사도 부스팅 대비 학습시간이 적게 걸리고, 과적합 제어 등 성능 향상에 필요한 기능들이 탑재되어있는 알고리즘이다. 경사도 부스팅과 유사하게 XGBoost 모형은 다음과 같이 표현된다 (Bentjac 등, 2019).

$$
L_{XGB} = \sum_{i=t}^{n} l(y_t, \hat{y}_t) + \sum_{k=1}^{K} \Omega(f_k),
$$

여기서 $l(y_t, \hat{y}_t)$는 예측값과 실제값의 차이를 나타내는 손실함수로 회귀문제에서는 $l(y_t, \hat{y}_t) = (y_t - \hat{y}_t)^2/2$가 주로 사용된다. $\Omega(f_k)$는 k번째 트리 f_k에 대한 정규화 항(regularization term)으로 모델의 복잡도와 과적합을 제어한다. 예측값 $\hat{y}_t$는 다음과 같이 K개 트리의 출력 합으로 계산된다.

$$
\hat{y}_t = \sum_{k=1}^{K} f_k(\boldsymbol{x}_t).
$$

2.5. SVR 모형

Support Vector Regression (SVR)은 SVM (Support Vector Machine)의 분류 개념을 연속형 반응 변수로 확장한 회귀 방법으로, Vapnik (1998)의 통계적 학습 이론에 기반한다. SVR은 오차가 특정 허용 범위 ε 이내일 경우 손실을 무시하는 ε-insensitive loss 함수를 사용함으로써, 이상값에 강인한 예측 모형을 구성할 수 있다.

SVR은 커널 함수를 통해 고차원 공간에서의 비선형 회귀 문제를 해결하며, 이때 적용되는 커널 함수는 Mercer 정리 (Cristianini와 Shawe-Taylor, 2000)에 따라 수학적 타당성이 보장된다. 일반적인 결정 함수는 다음과 같이 주어진다.

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b,$$

여기서 α_i, α_i^*는 라그랑주 승수이며, 예측 함수는 소수의 서포트 벡터에 의해서만 결정되므로 계산 효율성이 뛰어나다. SVR은 통계학적으로는 M-추정(M-estimation)의 성격을 가지며, 함수 추정의 편향-분산 균형을 조절할 수 있는 회귀모형으로 해석될 수 있다. Halil 등 (2019)은 기온, 습도, 요일 등 다양한 설명변수를 포함한 SVR 모형이 전통적 회귀모형 대비 우수한 단기 전력수요 예측 성능을 보인다고 연구하였다.

2.6. LSTM 모형

RNN 모형은 먼 시점의 과거 데이터에 대해 훈련의 성능이 떨어진다는 문제점이 있는데 이를 보완한 LSTM (Long Short-Term Memory) 모형이 Hochreiter와 Schmidhuber (1997)에 의해 제안되었다. LSTM 모형은 긴 의존 기간을 필요로 하는 학습에 적합하다. 일반적인 LSTM 모형은 셀, 망각 게이트(forget gate), 입력 게이트(input gate), 그리고 출력 게이트(output gate) 등으로 구성되며, 이 게이트들을 통하여 셀을 보호하고 제어한다. 망각 게이트는 f_t에 의해 삭제하는 내용을 제어하고, 입력 게이트는 i_t에 의해 g_t에서 셀 c_t에 더해지는 내용을 제어하고, 출력 게이트는 o_t에 의해 c_t 내용을 은닉층 h_t에 출력하는 것을 제어한다. 은닉층 h_t가 계산되는 과정을 식으로 표현하면 다음과 같다.

$$\begin{aligned}
f_t &= \sigma(w_{wf} x_t + w_{hf} h_{t-1} + b_f), \\
i_t &= \sigma(w_{\xi} x_t + w_{hi} h_{t-1} + b_i), \\
o_t &= \sigma(w_{xo} x_t + w_{ho} h_{t-1} + b_o), \\
g_t &= \tanh(w_{xg} x_t + w_{hg} h_{t-1} + b_g), \\
c_t &= (c_{t-1} \times f_t) + (i_t \times g_t), \\
h_t &= \tanh(c_t) \times o_t,
\end{aligned}$$

여기서 b_f, b_i, b_o, b_g는 편향(bias), 기호 '×'는 같은 차원의 두 행렬의 각 성분을 곱하는 연산, 그리고 σ는 시그모이드 함수를 나타낸다.

2.7. GRU 모형

Gated Recurrent Unit (GRU) 모형은 Cho 등 (2014)에 의해 제안되었으며, LSTM과 유사한 성능을 유지하면서도 LSTM과 대비하여 파라미터 수가 적고, 셀 상태를 별도로 두지 않아 구조가 단순하고, 계산 효율이 높고 메모리 사용량이 적은 특징이 있다. GRU 모형은 내부에 두 개의 게이트인 리셋 게이트(reset gate)와 업데이트 게이트(update gate)를 포함하며, LSTM과 달리 별도의 셀 상태(cell state)를 사용하지 않고 은닉 상태만을 통해 정보를 전달한다. 이러한 구조는 파라미터 수를 줄이면서도 장기 및 단기 정보를 효율적으로 조절할 수

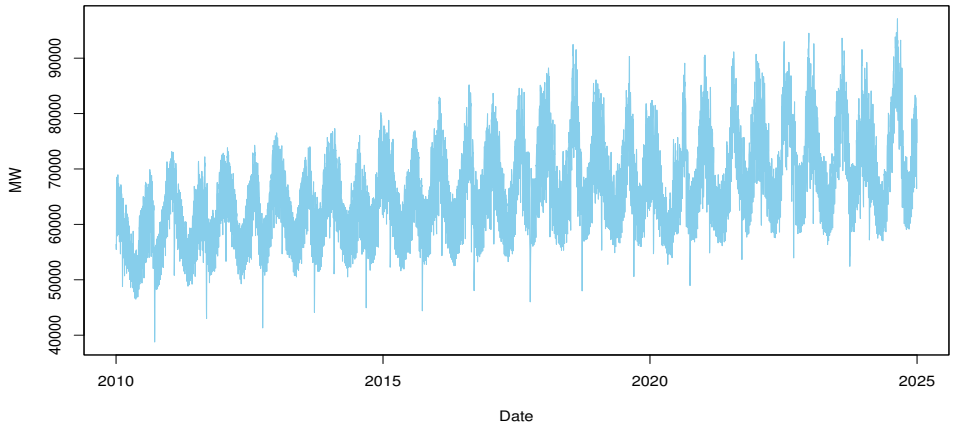


Figure 1: *Daily maximum electricity demand.*

있도록 한다. 연산 과정은 아래 수식과 같이 정의된다.

$$
\begin{aligned}
z_t &= \sigma(w_z x_t + w_{hf} h_{t-1} + b_z) && \text{(update gate)},\\
r_t &= \sigma(w_r x_t + w_{hi} h_{t-1} + b_r) && \text{(reset gate)},\\
\tilde{h}_t &= \tanh(w_h x_t + u_h (r_t \times h_{t-1}) + b_h) && \text{(candidate hidden state)},\\
h_t &= (1 - z_t) \times h_{t-1} + z_t \times h_t && \text{(new hidden state)},
\end{aligned}
$$

여기서 x_t는 시점 t의 입력, h_t는 은닉 상태(hidden state), σ는 시그모이드 함수, 기호 '$\times$'는 요소별 곱, 그리고 w, u, b는 학습되는 가중치와 편향을 나타낸다.

3. 데이터 및 자료 분석

3.1. 데이터 설명

본 논문에서 사용한 데이터는 한국전력거래소에서 운영하는 전력통계정보시스템(EPSIS)에서 다운로드한 최대전력수요 데이터로 측정단위는 메가와트(MW)이다. 한국전력거래소의 일별 최대전력은 가정용, 산업용, 공공용 등 모든 부문의 전력소비를 합산한 우리나라 전체 전력수요의 최대치로 정의된다. 이는 특정 부문을 별도로 분리하지 않고 계통 전체 부하를 기준으로 측정되며 하루(00:00~24:00) 동안 기록된 전력수요 가운데 가장 높은 값을 의미한다. 한국전력거래소의 일별 최대전력 자료는 우리나라 전력시장 및 계통운영의 대표적 수요 지표일 뿐만 아니라, 전력수요 예측 연구에서 가장 기초적이고 신뢰성 있는 데이터라 할 수 있다.

본 연구에서는 2010년 1월 1일부터 2024년 12월 31일까지 총 5,479일의 일별 최대전력 데이터를 수집하였다. 모형의 예측력을 비교하기 위해 2010년 1월 1일부터 2023년 12월 31일까지의 데이터를 시작으로 하나씩 추가하면서 모형을 적합하고 $h = 1, \ldots, 7$시차 이후의 예측값과 관측값의 비교를 통해 모형 성능을 평가하였다.

Figure 1은 14년 동안의 최대전력수요 추세를 표시한 시계열 그림으로 최대전력소비는 여름철 혹서기 냉방 수요와 겨울철 혹한기 난방 수요로 인한 계절적 변동성이 뚜렷한 계절성을 가지면서 변동폭이 조금씩 커지는 경향이 있는 것으로 보인다. 이는 통계적 모형으로 분석한다면 원래의 척도로 분석하는 것보다 로그 변환 후 분석하는 것이 더 적절할 수 있다는 것을 의미한다. 본 논문에서는 통계적 모형을 적용할 때 원래의

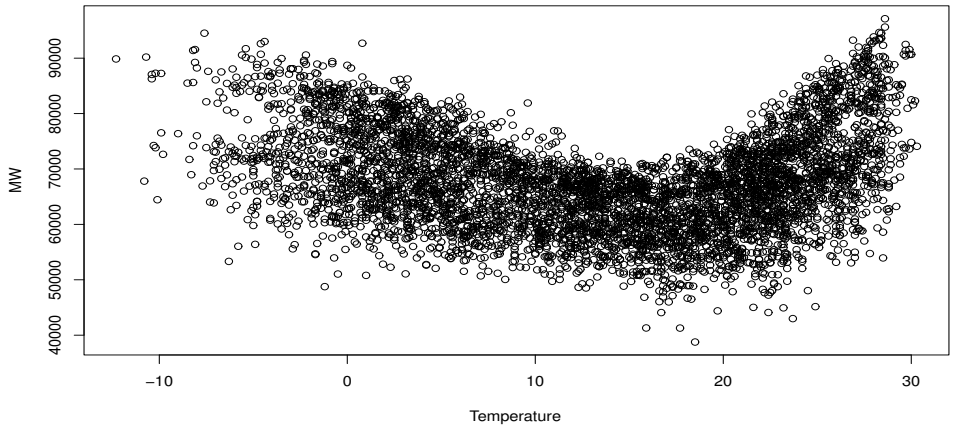


Figure 2: *Maximum electricity demand by temperature.*

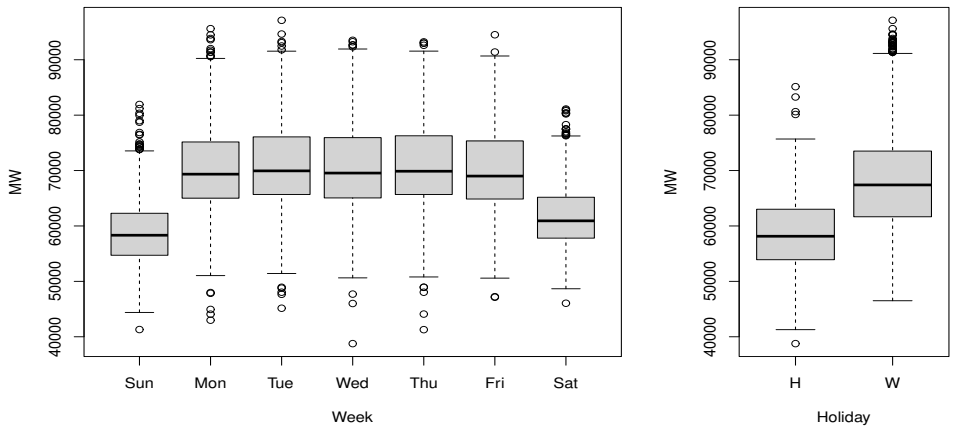


Figure 3: *Comparison of electricity demand by day of the week and by holiday status.*

척도 하에서 분석한 결과와 (2.1)과 같이 편향을 조정한 로그 변환-역변환의 결과를 비교해 본다. 기계학습 모형에는 특별히 분포 가정을 하지 않기 때문에 원래 척도 하에서 분석을 수행하였다. 연도가 증가함에 따라 수요도 증가하는 경향이 있지만 선형추세로 보기 어려운 구간도 있어 이전 수요량을 설명변수에 포함시킴으로써 해당 효과를 설명하고자 한다. 궁극적인 분석 목적이 t 시점까지의 관측자료를 이용하여 $t+h$ 시점의 수요를 예측하는 것이 때문에 모형 적합에서 Y_t에 대한 설명변수에 Y_{t-h}와 그 전의 관측자료가 포함되어야 하는데 모형의 단순화를 위해 본 논문에서는 Y_{t-h}만 포함시켰다.

Figure 2에 의하면 Figure 1에서의 계절성은 기온과 관련이 있은 것으로 보인는데 Kwon 등 (2021)의 연구에서는 15 °C를 기준으로 기온이 내려가거나 올라가면 전력수요가 늘어나는 경향이 있는 것으로 파악하고 |기온-15|을 설명변수에 포함시켰다. 본 연구에서는 Kwon 등 (2021)에서 사용한 |기온-15| 대신 기온의 제곱항을 모형에 포함시켜 자료에 의해 영향력의 형태가 만들어지도록 하였으며 두 방법을 비교한 결과 |기온-15|보다 기온과 기온의 제곱항을 추가했을 때 적합력과 예측력이 더 좋은 것으로 나타났다.

Figure 3은 요일과 공휴일 여부에 따른 최대전력수요의 분포를 비교한 상자그림이다. 요일별로 비교했

을 때 주말의 수요가 평일에 비해 유의하게 낮은 것을 볼 수 있으며 공휴일의 경우에도 수요가 낮은 것으로 나타났다. 월요일부터 금요일까지의 수요 분포가 비슷하고 토요일과 일요일의 수요 분포가 비슷하여 요일을 평일과 주말로 나누어 분석해 보았으나 요일별로 나누었을 때보다 적합력이나 예측력이 유의하게 떨어지는 것으로 분석되었다. 이런 결과를 바탕으로 최종 설명변수는 다음과 같이 설정하였다.

$$\boldsymbol{x}_t = (Y_{t-h}, T_t, T_t^2, W_{1t}, \ldots, W_{6t}, H_t)^T,$$

여기서 T_t는 t 시점에서의 기온, W_{1t}는 해당 시점이 월요일이면 1 아니면 0인 가변수이며 나머지 W 변수는 요일을 바꾸어 가면서 적용한 가변수이다. H_t는 t 시점이 요일과 관계 없이 공휴일이면 1 아니면 0인 가변수를 의미한다. 통계적 모형에서 수요자료를 로그변환한 경우 설명변수 Y_{t-h} 대신 로그 변환된 $\log(Y_{t-h})$를 사용한다. 시점 t에서 $t+h$ 시점의 예측을 위해서는 $\boldsymbol{x}_{t+h}$가 필요한데 일반적으로 $h(\leq 7)$차 후의 기온은 정확하게 예측될 수 있기 때문에 $\boldsymbol{x}_{t+h}$를 얻는데 특별한 문제가 없다.

3.2. 모형 적합 및 예측력 비교

본 논문에서는 7가지의 데이터 모형을 바탕으로 2010년부터 2023년까지의 일별 최대전력수요 자료를 시작으로 하나씩 자료를 추가하며 모형을 적합하며 $h = 1, \ldots, 7$차 후의 수요를 예측하여 비교하였다. 비교 대상 모형은 앞써 설명했던 ARE, ARE-GARCH, Random Forest (RF), XGBoost, SVR, LSTM, GRU이며 통계적 시계열분석 모형인 ARE와 ARE-GARCH의 경우 로그변환 분석도 추가하였다. 분석에 사용된 프로그램으로 ARE-GARCH는 R의 rugarch 패키지를, ARE는 Python의 statsmodels, Random Forest와 SVR은 scikit-learn, XGBoost는 xgboost, LSTM과 GRU는 tensorflow 패키지를 이용하였다.

ARE 모형을 활용한 사전 분석 결과 ARE 차수 k는 8이 적정한 것으로 분석되었으며 이를 기반으로 ARE와 ARE-GARCH에서 AR 차수는 $k = 8$로 지정하였다. Random Forest 분석을 위해 시계열 교차 검증 방식을 사용하여 트리의 개수는 300, 트리의 최대 깊이는 15, 최종 노드에 있어야 하는 최소 샘플 수는 10으로 하는 것이 적절한 것을 확인하였다. XGBoost 모형에서 손실함수는 reg:squarederro, 모형이 생성할 트리의 개수는 100, 각 트리의 최대 깊이는 6, 학습률은 0.05으로 지정하였다. SVR 모형에서는 StandardScaler를 사용하여 표준화한 자료를 LinearSVR를 통해 분석하였으며 최적화 조건의 민감도를 높이기 위해 오차(수렴기준)을 0.001로 지정하였다. LSTM 모형은 자료를 MinMaxScaler로 정규화한 후 LSTM 레이어의 뉴런 수를 50, 활성화함수를 relu, Dense 레이어를 1, 최적화 알고리즘은 Adam, 손실함수는 mse로 지정하고 분석하였다. GRU의 경우에도 LSTM과의 비교를 위해 동일한 세팅에서 분석이 이루어졌다.

모형의 예측 성능을 비교하기 위하여 다음과 같이 RMSE (Root Mean Square Error), MAE (Mean Absolute Error)와 MAPE (mean absolute percentage error)를 사용하였다.

$$\begin{aligned}
\text{RMSE} &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}, \\
\text{MAE} &= \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|, \\
\text{MAPE} &= \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|.
\end{aligned}$$

Table 1은 분석모형에 대해 시차 $h = 1, \ldots, 7$까지의 예측에 대한 RMSE, MAE, MAPE를 정리한 것이다. 이 표에서 각 기준과 시차별로 예측력이 가장 좋은 결과에 대해 밑줄을 그어 표시했다. 표에서 ARE와 GARCH 다음에 -O가 표시된 것는 원 자료에 대한 분석 결과이고 -L이 표시된 것은 편향조정을 거친 로그변환-역변환

Table 1: Comparison of prediction by time lag

Criterion	Model	Time lag(h)						
		1	2	3	4	5	6	7
RMSE	ARE-O	2976	3871	4293	4489	4320	4100	3949
	ARE-L	2871	3559	3866	4080	4067	3999	3938
	GARCH-O	2803	3519	3685	3872	3934	3898	3789
	GARCH-L	2705	3357	3579	3760	3812	3814	3750
	RF	3196	3834	4061	4192	4169	3907	3741
	SVR	2964	3748	3958	4130	4082	3912	3761
	XGBoost	**2574**	3316	3587	3659	3802	3844	3603
	LSTM	2840	2842	**2785**	2839	**2836**	**2781**	**2770**
	GRU	2881	**2798**	2815	**2747**	2859	**2781**	2829
MAE	ARE-O	2088	2814	3155	3370	3370	3143	3006
	ARE-L	2002	2559	2835	3034	3114	3056	2985
	GARCH-O	1976	2459	2629	2822	2971	2940	2849
	GARCH-L	1876	2322	2531	2731	2866	2890	2836
	RF	2300	2761	2985	3175	3206	2957	2778
	SVR	2095	2669	2850	3024	3130	2930	2794
	XGBoost	**1813**	2341	2600	2711	2905	2839	2628
	LSTM	2024	**1999**	**1989**	**1989**	**1994**	**1943**	**1983**
	GRU	2017	2006	2021	1998	2047	1950	2004
MAPE(%)	ARE-O	2.889	3.854	4.277	4.552	4.538	4.205	3.992
	ARE-L	2.774	3.490	3.823	4.074	4.173	4.071	3.941
	GARCH-O	2.720	3.328	3.513	3.770	3.973	3.927	3.791
	GARCH-L	2.583	3.142	3.384	3.645	3.826	3.847	3.755
	RF	3.025	3.607	3.901	4.167	4.222	3.905	3.657
	SVR	2.894	3.625	3.819	4.063	4.202	3.933	3.731
	XGBoost	**2.464**	3.142	3.464	3.621	3.896	3.791	3.486
	LSTM	2.789	**2.752**	**2.738**	**2.759**	**2.767**	**2.686**	**2.743**
	GRU	2.791	2.782	2.786	2.767	2.833	2.689	2.775

분석결과를 의미한다. 두 분석 모형 모두 원자료를 사용하는 것보다 편향조정한 로그변환-역변환한 결과가 더 예측력이 있는 것으로 나타났다.

분석 결과 시차에 따라 최적의 모형이 다른 것으로 나타났는데 1시차 후에 대한 예측에서는 모든 기준에서 XGBoost 모형이 가장 좋은 것으로 나타났다. GARCH-L도 1시차 후의 예측에서는 XGBoost를 제외한 다른 모형보다는 예측력이 높은 것으로 나타났다. 딥러닝 모형인 LSTM과 GRU의 경우 시차에 관계없이 비슷한 수준의 예측력을 가지는 반면 나머지 모형에서는 2시차 이후의 예측력이 1 시차의 예측력보다 현저히 떨어지는 것으로 분석되었다. MAE와 MAPE 기준에서 2시차 이후의 예측에서는 LSTM이 가장 좋은 예측력이 있는 것으로 나타났다. 하지만 초월모수의 세팅을 조금 변경한 경우에 GRU가 LSTM 보다 좋은 예측력이 가지는 시차도 있고 표에서 보는 것과 같이 LSTM의 결과와 GRU의 결과에 큰 차이가 없어 두 모형은 유사한 성능을 갖는 것으로 보아도 무리가 없을 것으로 생각된다. 시계열 교차 검증 방식으로 초월모수를 지정한 random forest와 ARE-O의 경우 다른 모형에 비해 전체적으로 예측력이 떨어지는 것으로 분석되었다. SVR의 경우에도 딥러닝 모형이나 GARCH-L과 XGBoost보다 낮은 예측력을 보였으나 보다 세밀한 커널이나 옵션 설정을 통해 개선될 여지가 있을 것으로 보인다.

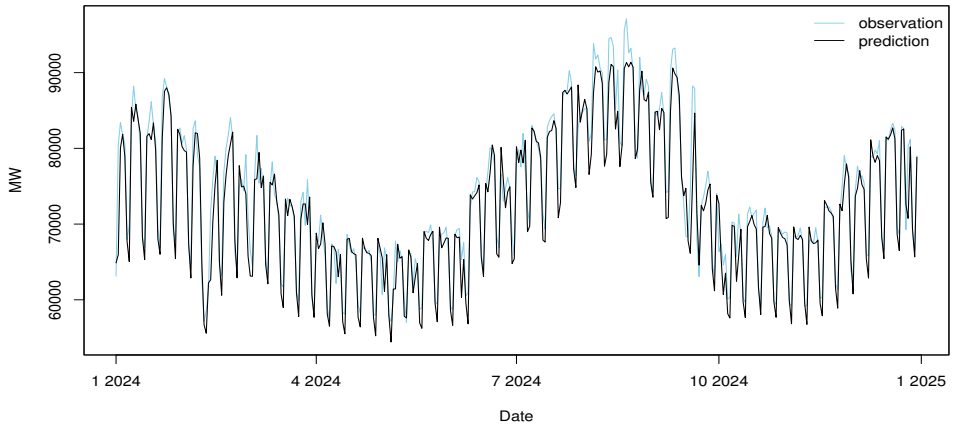


Figure 4: *XGBoost predictions.*

4. 결론

본 연구에서는 시계열모형과 기계학습 및 딥러닝 모형으로 최대전력수요를 예측하고 그 성능을 비교하고자 했다. 분석 결과에 의하면 1시차 후의 예측력은 XGBoost 모형이 가장 좋은 예측력을 가지고 있고 2시차에서 7시차 후에 대해서는 LSTM과 GRU와 같은 딥러닝 모형의 예측력이 높은 것으로 나타났다. 딥러닝 모형을 제외한 나머지 모형에서는 2시차 이후의 예측력이 1시차의 예측력보다 현저히 낮아지는 것으로 나타났다. 분포 가정에서 추론이 이루어지는 시계열 모형, ARE와 GARCH의 경우 원자료를 그대로 분석하는 것보다 로그변환 후 모형을 추정하고 예측값을 구한 후 편향을 보정한 역변환을 통해 예측 했을 때 예측력이 훨씬 좋은 것으로 나타났다. 본 논문에서는 기계학습 모형과 딥러닝 모형에 대해 로그변환-역변환 방법을 적용하지 않았으나 이를 적용하면 더 높은 예측력을 얻을 것으로 기대된다.

기계학습 모형과 딥러닝과 같은 인공지능 모형의 경우 초월모수의 선택에 결과가 영향을 받는 경향이 있는데 본 연구에서는 random forest를 제외한 모형에서 기본에 가까운 옵션 선택 하에서 분석이 이루어졌다. 이후 연구에서는 예측력을 높일 수 있는 옵션 선택에 대해 알아보고 하나의 모형으로 예측값을 구하는 것이 아니라 설명력이 높은 여러 모형의 결과를 종합하여 예측값을 구하는 방법에 대해 알아 볼 예정이다. Figure 4는 1시차 이후의 예측에서 가장 좋은 성능을 보였던 XGBoost 모형의 1시차 후 예측값과 관측값을 표시한 것으로 전체적으로 예측이 적절히 이루어 지고 있으나 8월 9월에서 과소 예측하는 것을 볼 수 있다. 이는 Figure 2에서 최대전력소비가 15 °C를 기준으로 기온이 낮아지면 완만하게 증가하는 반면 높아지면 좀더 급격하게 증가하는 형태를 가지는데 이를 기온의 2차 식으로 표현해서 상승 패턴을 완벽하게 설명하지 못해 발생한 것으로 예상된다. 향후 연구에서는 기온의 영향을 비대칭적으로 모형화 할 수 있는 방안에 대해 알아 볼 예정이다. 또한 최대전력소비는 통계적 관점에서 극한값에 속하는 자료이기 때문에 오차항의 분포로 정규분포나 t-분포와 같은 대칭적 분포를 가정하지 않고 극한값 분포와 같은 비대칭적 분포하에서 예측하는 방법들에 대해 알아보고자 한다.

References

Bentjac ML, Csallner C, and Sudholt D (2019). A theoretical analysis of the XGBoost algorithm, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2019)*, 1116–1123.

Bollerslev T (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, **31**, 307–327.

Breiman L (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Chen T and Guestrin C (2016). XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, and Bengio Y (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734.

Cristianini N and Shawe-Taylor J (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge.

Engle RF (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, **50**, 987–1007.

Halil K, Özgür Y, and Mehmet Ş (2019). Short-term electric load forecasting using support vector regression and multi-layer perceptron, *Energy*, **170**, 130-–147.

Hastie T, Tibshirani R, and Friedman J (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)*, Springer, Berlin.

Hochreiter S and Schmidhuber J (1997). Long short-term memory, *Neural Computation*, **9**, 1735–1780.

James G, Witten D, Hastie T, and Tibshirani R (2015). *An Introduction to Statistical Learning: with Applications in R (2nd ed.)*, Springer, Berlin.

Kwon S, Kim J, Sohn S, and Lee S (2021). Comparison of time series predictions for maximum electric power demand, *The Korean Journal of Applied Statistics*, **34**, 623–632.

Vapnik V (1998). *Statistical Learning Theory*, Wiley, New York City.

Received September 15, 2025; Revised September 23, 2025; Accepted September 29, 2025

시계열모형과 인공지능모형을 이용한 최대 전력수요 예측

권숙희[a], 조완섭[b], 여인권[1,c]

[a]충북대학교 빅데이터협동과정; [b]충북대학교 경영정보학과; [c]숙명여자대학교 통계학과

요 약

전력망 운영의 효율성 극대화를 위하여 정확한 전력수요 예측이 필수적이며, 이를 위해 본 연구에서는 시계열 모형과 인공지능(기계학습 및 딥러닝) 모형으로 일별 최대전력수요를 예측하고 성능을 비교해 본다. 분석결과 1시차 후의 예측은 XGBoost 방법이 가장 잘하는 것으로 나타났으며 2시차부터 7시차 후의 예측은 딥러닝 모형인 LSTM과 GRU가 잘하는 것으로 분석되었다. 시계열 모형의 경우 원자료를 분석하는 것보다 편향을 보정한 로그변환-역변환 방법으로 예측했을 때 보다 우수한 결과를 보이는 것으로 나타났다.

주요용어: 시계열 모형, 인공지능 모형, 최대전력수요 예측

[1]교신저자: (04310) 서울특별시 용산구 청파로 47길 100, 숙명여자대학교. E-mail: inkwon@sookmyung.ac.kr

RESEARCH INSTITUTE OF NATURAL SCIENCES

2025 자연과학연구소 연구결과백서

화학과

김효진 교수

논문

- Kim, H.†; Shin, D. Y.†; Choe, J. H.; Cheon, G.; Yun, H.; Lee, J.-H.*; Hong, C. S.*, (2026). Elucidating the Sigmoidal Adsorption Behavior of Xenon in Flexible Hofmann-Type MOFs through Experiments and Molecular Dynamics with Machine Learning Potentials. Small. 22, e09479. (†equal contribution)

Check for updates

Small

NANO · MICRO small
www.small-journal.com

RESEARCH ARTICLE

Elucidating the Sigmoidal Adsorption Behavior of Xenon in Flexible Hofmann-Type MOFs Through Experiments and Molecular Dynamics with Machine Learning Potentials

Hyojin Kim[1] | Dong Yun Shin[2] | Jong Hyeak Choe[3] | Gayoung Cheon[3] | Hongryeol Yun[3] | Jung-Hoon Lee[2,4] | Chang Seop Hong[3]

[1]Department of Chemistry, Sookmyung Women's University, Seoul, Republic of Korea | [2]Computational Science Research Center, Korea Institute of Science and Technology (KIST), Seoul, Republic of Korea | [3]Department of Chemistry, Korea University, Seoul, Republic of Korea | [4]KU-KIST Graduate School of Converging Science and Technology, Korea University, Seoul, Republic of Korea

Correspondence: Jung-Hoon Lee (jhlee84@kist.re.kr) | Chang Seop Hong (cshong@korea.ac.kr)

Received: 6 August 2025 | **Revised:** 12 December 2025 | **Accepted:** 16 December 2025

Keywords: kinetic analysis | machine learning potential | metal-organic frameworks | rotational dynamics | sigmoidal xenon adsorption

ABSTRACT

Sigmoidal adsorption in flexible metal-organic frameworks (MOFs) can be achieved through a dramatic phase transition from narrow to large pores; however, such a unique phenomenon, which is not induced by abrupt phase changes, has rarely been reported. We report a series of flexible Hofmann-type MOFs, **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe**, pillared with a dimensional (3D) piperazine linker. Owing to the rotational flexibility and 3D dimensionality of the piperazine linker, **CoNi-Pipe** and **CoPd-Pipe** exhibited Xe-induced gate opening and the corresponding sigmoidal adsorption isotherms. **CoNi-Pipe** revealed not only excellent Xe/Kr separation but also Xe recovery performance, attributed to the greatest linker rotation and flexibility, along with moderate Xe binding strength, as validated by the machine learning potential-based molecular dynamics simulations and van der Waals-corrected DFT calculations. Thus, this work provides an effective approach for gaining a deeper understanding of the structural dynamics in flexible frameworks and expanding the repertoire of computational studies.

1 | Introduction

Owing to their structural diversity, adjustable pore environment, and tunable functionality, metal–organic frameworks (MOFs) have become prominent in adsorptive applications [1–4]. Most rigid MOFs typically display a common Langmuir type I adsorption isotherm, and adsorption is primarily governed by thermodynamic effects [5, 6]. As an intriguing class of porous materials, flexible MOFs can show stimuli-responsive structural dynamics, offering fortuitous functionalities and unique type V or VI (S-shaped or multi-step) adsorption behaviors [6–14]. Recently, flexible MOFs have been particularly explored as an advantageous platform for gas separation of similar gas pairs and gas storage [12, 15–19]. The dynamic nature enables these MOFs to recognize a specific molecular stimulus while excluding others in a gas mixture, thereby enhancing gas separation efficiency. Moreover, the difference between the amount of gas adsorbed at the target pressure and the amount at the desired desorption pressure, commonly known as the working capacity, can be maximized to boost storage performance [19].

Structural dynamics generally involve either the rearrangement and/or rotation of molecular linkers within the frameworks, resulting from external stimuli such as temperature, pressure,

H.K. and D.Y.S. contributed equally to this work.

© 2025 Wiley-VCH GmbH

and guest molecules. The most commonly observed adsorption behavior in flexible MOFs is temperature-responsive gate opening, resulting in a sigmoidal S-shaped isotherm for certain gas molecules [13, 17, 20, 21]. This can typically be realized through a significant phase transition from the narrow phase to the large phase, associated with pore contraction and expansion, including significant variations in the unit cells and pore volumes [6, 7, 15, 17, 20, 22–26]. For example, $Mn(ina)_2$ and X-dia1-Ni showed temperature-responsive gas adsorption with a gate-opening phenomenon, which should be accompanied by an abrupt transformation from a guest-free and contracted pore structure into an expanded pore structure [17, 20].

However, a gate-opening phenomenon not induced by dramatic phase changes in the framework has been reported only in a few cases [24, 27–29]. Recently, ZUL-C6, which incorporates 3D alkane-bridged ligands, exhibited guest-responsive flexibility, thereby undergoing a unique single-crystal-to-single-crystal (SC-to-SC) transformation and achieving benchmark separation of hexane isomers [29]. Such unusual gate-opening characteristics can be triggered by flexible linker rotation, imparting local dynamics to the frameworks. However, it is not necessarily the case that local dynamics lead to sigmoidal sorption behaviors [16, 30]. The prevalent linker is primarily based on 2D, planar, and symmetrical structures, which can contribute to the high adsorption or separation capability of the frameworks, but it features common adsorption behaviors [31–35]. Replacing 3D linkers afforded temperature-dependent diffusion-controlled gas adsorption, leveraging the non-planar dimensionality and flexibility of the linker [30]. Although it is an exceptional phenomenon in gas adsorption, it does not cause sigmoidal adsorption, probably because of the regular pore channels based on the symmetrical linker. This study envisioned that incorporating nonplanar and non-centrosymmetric 3D linkers with flexibility and steric effect would generate sub-Å-controlled irregular pore channels that enable sigmoidal adsorption by providing distinct diffusion pathways for guest molecules that differ by sub-Ångström in size. (Figure 1) [28, 36–39].

Compared with rigid MOFs bearing static structures, elucidating the dynamic adsorption behavior of flexible MOFs is challenging. This is because molecular dynamics (MD) simulations using classical force fields, despite their ability to simulate large systems over extended time scales, fail to accurately capture complex electrostatic and van der Waals interactions between the molecules and the framework, as well as intermolecular interactions. Moreover, conventional force fields cannot accurately describe the flexibility-related characteristics of MOFs in MD simulations [40, 41]. Although density functional theory (DFT) calculations offer high accuracy, they are inherently limited in terms of the feasible simulation time and system size, making them unsuitable for kinetic analyses.

To overcome these challenges, machine learning potentials (MLPs) have emerged as promising alternatives, offering improved accuracy in capturing complex interactions while retaining the efficiency of classical MD simulations. This advancement enables large-scale and long-term simulations with DFT-level accuracy, thereby deepening our understanding of the fundamental properties of different materials and uncovering new physical and chemical insights [42–44]. Particularly, MLPs have been successfully applied to elucidate the mechanical and diffusion properties of flexible framework materials, showing their potential for accurately modeling structural dynamics and adsorption behavior [45–47].

This work reports a series of Hofmann-type ultramicroporous MOFs, termed **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe**, by employing non-centrosymmetric 3D piperazine as a pillar linker. To this end, gas adsorption properties of these MOFs were investigated through experiments, including isotherm and gas-loaded powder X-ray diffraction (PXRD) and MLP-based MD (MLP-MD) simulations. By virtue of the rotational flexibility and steric effect of the piperazine linkers and the ultramicroporosity of the frameworks, the frameworks featured structural dynamics in response to gas molecular stimuli, thereby regulating their gas adsorption behaviors [10]. Noble Xe and Kr are used as gas probes for investigating the structural flexibility and dynamics because of their atomic sizes, which are similar to those of the frameworks, spherical symmetry, and inertness. The rotational flexibility and different arrangements of the piperazine between adjacent 2D $\{Co[Ni(CN)_4]\}_n$ layers enable **CoNi-Pipe** and **CoPd-Pipe** to have two types of pore channels, referred to as small pores and large pores, endowing gate-opening behavior toward Xe with a sigmoidal sorption isotherm. Such sigmoidal Xe adsorption offers benefits in terms of energy reduction for not only Xe/Kr separation but also Xe storage. According to the MLP-MD simulations, **CoNi-Pipe** demonstrated the most pronounced fluctuations and rotation of the piperazine linkers, possessing the greatest structural flexibility and a dynamic pore environment. Hence, the high Xe mobility and suitable binding strength in the framework led to excellent Xe/Kr separation and Xe storage capability. Remarkably, **CoPd-Pipe** showed relatively restricted Xe mobility owing to its smaller pore channels, particularly showing limited diffusion in small pores. Therefore, **CoPd-Pipe** exhibited more distinct sigmoidal adsorption, thereby exhibiting a high working capacity and conspicuous Xe storage with low regeneration energy. Conversely, **NiNi-Pipe** did not exhibit such sigmoidal sorption toward Xe owing to its low Xe mobility and limited linker rotation. Thus, elucidating the structural dynamics and adsorption mechanisms of flexible systems is crucial for designing potential adsorbents for various industrial applications.

2 | Results and Discussion

2.1 | Synthesis and Characterization

3D Hofmann-type ultramicroporous MOFs, **CoNi-Pipe** and **NiNi-Pipe**, were constructed from 2D $\{Co[Ni(CN)_4]\}_n$ or $\{Ni[Ni(CN)_4]\}_n$ layers, respectively, which were further pillared with piperazine (Figure 2a). To investigate the effects of different open metal sites, **CoPd-Pipe**, a Hofmann-type ultramicroporous MOF based on $\{Co[Pd(CN)_4]\}_n$ layers (Figure 2b), was synthesized.

To understand the crystal structures of **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe**, structural optimization was first performed by rotating AA- (or BB-) and AB-type piperazine linkers using van der Waals (vdW)-corrected DFT calculations (Figures S1 and S2 and Table S1). Using the DFT-optimized structures, the final

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

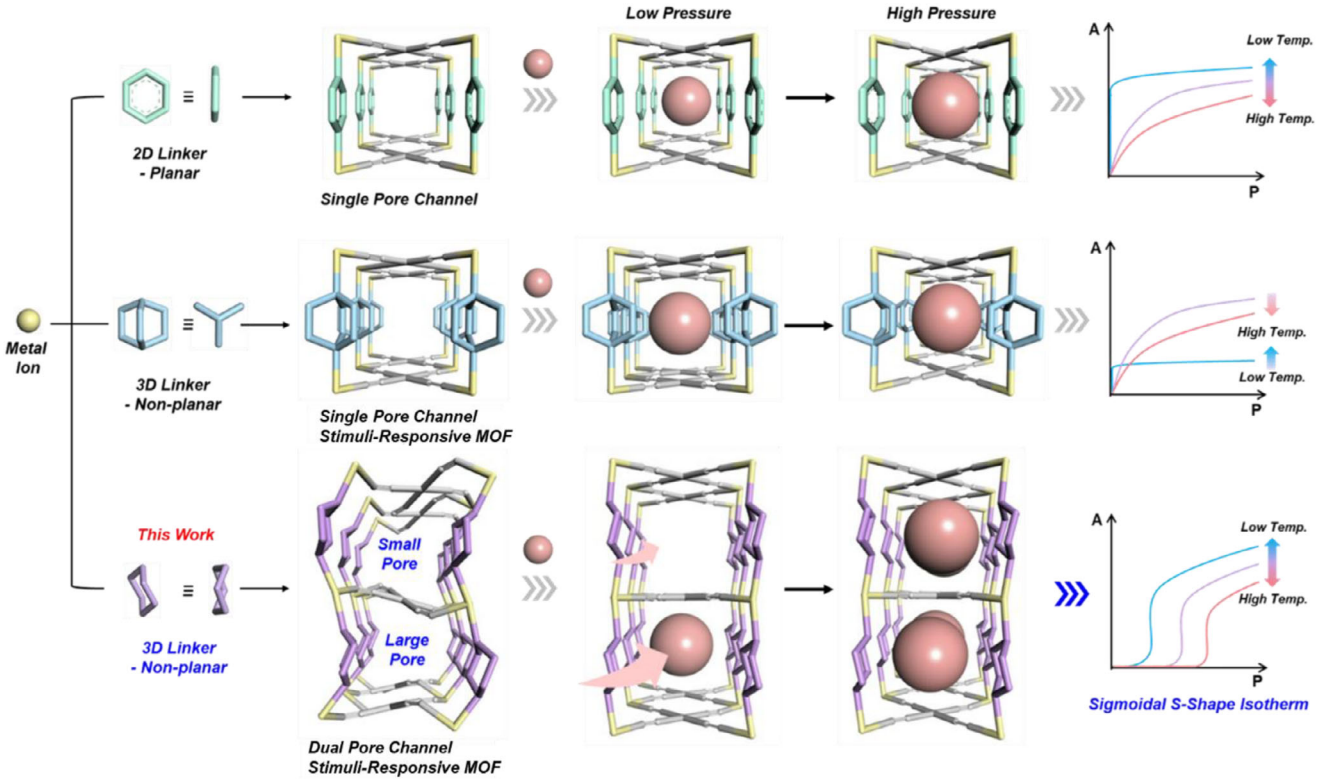


FIGURE 1 | Schematic of pore chemistry depending on the effect of the linker and the corresponding adsorption behaviors in flexible MOFs.

structures were successfully elucidated via Rietveld refinement of the synchrotron PXRD data (Figure 2c–e, Tables S2–S7).

The Rietveld refinement showed that the crystal structures of **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe** were characterized by the triclinic *P*1 space group (Tables S2, S4, and S6) [48–51]. As shown in Figure 2a, the cobalt(II) and nickel(II) cations are coordinated with the terminal N atoms of the $[Ni(CN)_4]^{2-}$ anion unit to form a 2D $\{Co[Ni(CN)_4]\}_n$ and 2D $\{Ni[Ni(CN)_4]\}_n$ layers, respectively. The cobalt(II) cations were also coordinated with the terminal N atoms of the $[Pd(CN)_4]^{2-}$ anion unit to form a 2D $\{Co[Pd(CN)_4]\}_n$ layer (Figure 2b). The respective 2D layers were further pillared with piperazine to generate the 3D frameworks. Interestingly, it was revealed that the piperazine linkers adopted different configurations in each framework: AB mode in **CoNi-Pipe**, BB mode in **NiNi-Pipe**, and AB mode in **CoPd-Pipe** (Figures S3–S5). Specifically, despite sharing the same AB mode, the 2D layers of the **CoPd-Pipe** showed significant structural distortion compared to those of **CoNi-Pipe**. The AB and BB stacking modes represent the alignment directions of piperazine linkers: in the AB stacking mode, piperazine linkers are oriented in alternating directions from layer to layer, whereas in the BB mode, they are aligned in the same direction, as shown in Figure S2 [52, 53].

In Figure 2a,b, the crystal structures of the MOFs clearly exhibit 1D pore channels formed by the piperazine linkers along the *a*- and *b*-axes. Additionally, oppositely adjacent open metal sites (OMSs) of Ni^{2+} or Pd^{2+} within the portal windows were observed, providing accessible binding sites for guest molecules. The OMS densities of the frameworks were calculated to be 9.737, 9.175, and 10.093 mmol cm^{-3} for **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe**, respectively. To the best of the authors' knowledge, these frameworks exhibit the highest OMS densities among all reported Hofmann-type MOFs, including CoNi-DAB (8.82 mmol cm^{-3}), ZJU-74-Ni (9.05 mmol cm^{-3}), and ZJU-74-Pd (8.50 mmol cm^{-3}), as well as other well-known MOFs with high OMS densities [30–35, 54, 55].

In addition to the 3D connectivity of the piperazine linker, the different arrangements led to non-uniform pore shapes in each framework (Figures S3b, S4b, and S5b). Notably, **CoNi-Pipe** and **CoPd-Pipe** featured dual-channel systems consisting of two irregular pore channels, corresponding to the A and B modes of the piperazine, thereby providing multiple binding pockets for guest molecules (Figures S3 and S5). The Ni···Ni distances between neighboring OMSs in **CoNi-Pipe** were 3.745 and 3.696 Å for the A and B modes, respectively, after excluding the vdW radius (Figure S3a). However, **NiNi-Pipe** exhibited a Ni···Ni distance of 4.232 Å, which is relatively longer than those of **CoNi-Pipe** due to its rhombus-shaped and tilted pore structure (Figure S4a). Similarly, in **CoPd-Pipe**, the Pd···Pd distances between neighboring OMSs were 3.49 and 3.314 Å for the A and B modes, respectively, after excluding the vdW radius (Figure S5a). The smaller Pd···Pd distances in **CoPd-Pipe** compared to those of **CoNi-Pipe** can be attributed to its distorted pore structure and the relatively larger vdW radius of Pd.

Furthermore, multiple hydrogen atoms from the piperazine linkers were exposed to these irregular pore channels, potentially serving as additional adsorption sites. The pore structures and the corresponding pore windows are shown in Figures S6–S8. Notably, **CoNi-Pipe** and **CoPd-Pipe** contained two distinct pore types, referred to as small pores and large pores, which

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

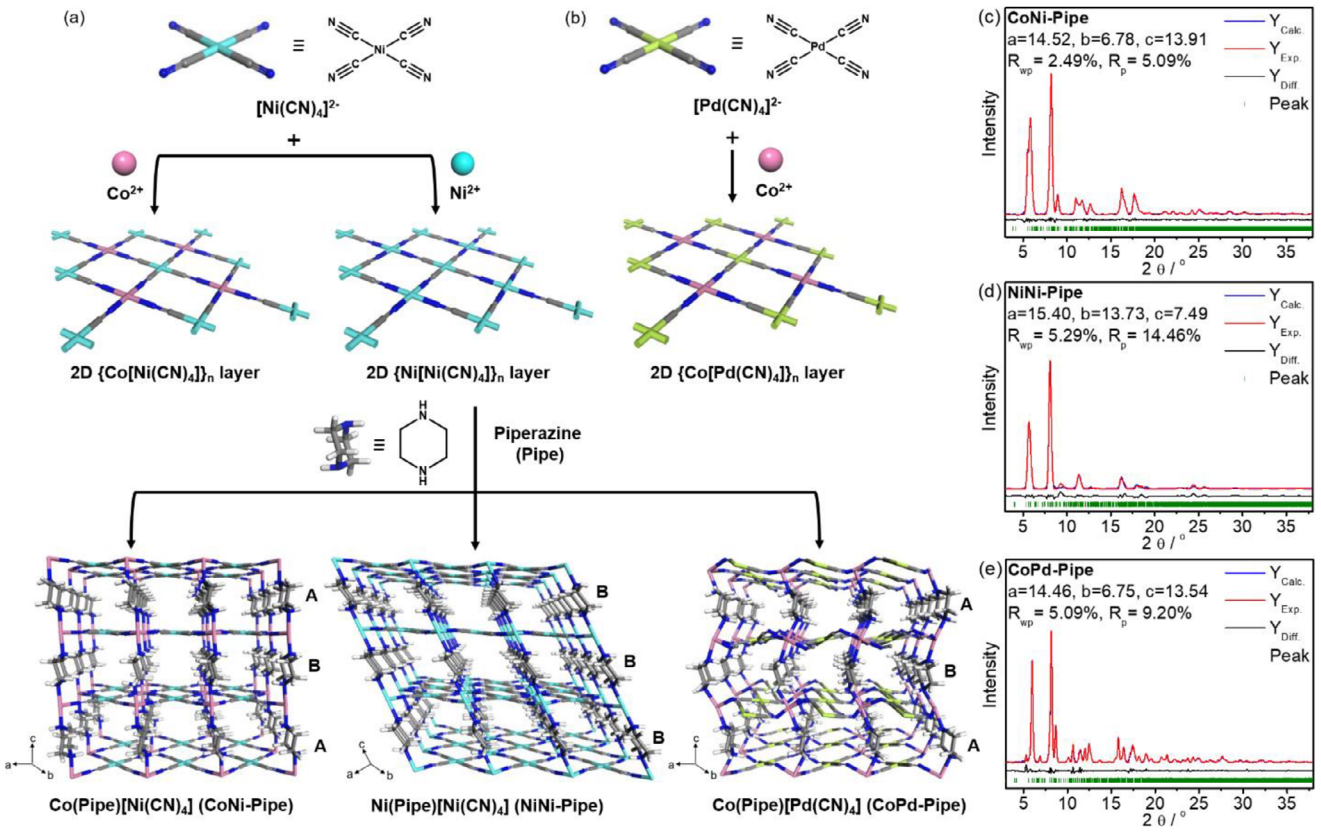


FIGURE 2 | Single-crystal structures of (a) **CoNi-Pipe**, **NiNi-Pipe**, and (b) **CoPd-Pipe**, pillared by piperazine from 2D $\{Co[Ni(CN)_4]\}_n$, $\{Ni[Ni(CN)_4]\}_n$, and $\{Co[Pd(CN)_4]\}_n$ layers, respectively. Color code: Co (pink), Ni (cyan), Pd (yellow-green), N (blue), C (gray), and H (white). PXRD pattern and the corresponding Rietveld refinement of (c) **CoNi-Pipe**, (d) **NiNi-Pipe**, and (e) **CoPd-Pipe**. Simulated data (blue line), experimental data (red line), the difference between the experimental PXRD and simulated patterns obtained by Rietveld refinements (black line), and the observed reflections (Bragg positions/peaks) (green ticks).

arise due to the A and B modes of piperazine [56]. The non-planar nature of piperazine led to the formation of pocket-like pores with varying window sizes along the *a* and *b* directions (Figures S6 and S8). In contrast, the uniform arrangement of piperazine in the **NiNi-Pipe** resulted in a single-pore structure (Figure S7).

As shown in Figures S6–S8, the sizes of the pocket-like pores are comparable to the kinetic diameter of Xe (4.05 Å), promoting effective Xe adsorption. However, the window sizes along the four surrounding directions were considerably smaller than the kinetic diameter of Xe, which could affect the Xe adsorption behavior by exploiting the intrinsic flexibility of piperazine. The guest-accessible void volumes were estimated to be 1364, 1448, and 1316 $Å^3$ for the **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe**, respectively [57].

The PXRD analysis confirmed that the bulk sample phases matched well with the simulated crystal structure patterns (Figure S9). Infrared (IR) spectroscopy further verified the presence of C≡N stretching peaks corresponding to the $\{Ni[Ni(CN)_4]\}_n$, $\{Co[Ni(CN)_4]\}_n$, and $\{Co[Pd(CN)_4]\}_n$ layers (Figure S10). Blue-shifts in the C≡N stretching vibrations of $K_2[Ni(CN)_4]\cdot H_2O$ and $K_2[Pd(CN)_4]\cdot H_2O$ were also observed, confirming the formation of C≡N-Co and C≡N-Ni coordination within the respective 2D layers [30, 58].

X-ray photoelectron spectroscopy (XPS) analysis was conducted to examine the atomic compositions and chemical environments of **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe** (Figures S11–S13). Thermogravimetric analysis (TGA) was performed to assess the thermal stability of the frameworks (Figure S14); the results showed that all structures remained intact up to 300°C. The initial weight loss was attributed to desorption of the piperazine linker. Scanning electron microscopy (SEM) images revealed the pebble-like spherical morphology of the MOFs, whereas energy dispersive spectrometry (EDS) elemental mapping confirmed the homogeneous distributions of C, N, Co, and Ni in the **CoNi-Pipe**; C, N, and Ni in the **NiNi-Pipe**; and C, N, Co, and Pd in the **CoPd-Pipe** (Figures S15–S17). Moreover, inductively coupled plasma atomic emission spectroscopy (ICP-AES) confirmed that the Co/Ni and Co/Pd ratios were approximately 1 for **CoNi-Pipe** and **CoPd-Pipe**, respectively (Table S8).

N_2 adsorption isotherms were obtained at 77 K to assess the permanent porosity of the MOFs (Figures S18–S20). The Brunauer–Emmett–Teller (BET) surface area determined from the N_2 isotherms was 535 $m^2\ g^{-1}$ for both **CoNi-Pipe** and **NiNi-Pipe**, and 270 $m^2\ g^{-1}$ for **CoPd-Pipe**. **CoPd-Pipe** showed a lower BET surface area and reduced N_2 uptake due to the relatively larger vdW radius of Pd compared with Ni [32]. As typically observed in the porosity analysis of ultramicroporous MOFs, CO_2 adsorption isotherms were measured at 195 K (Figures S21–S23)

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

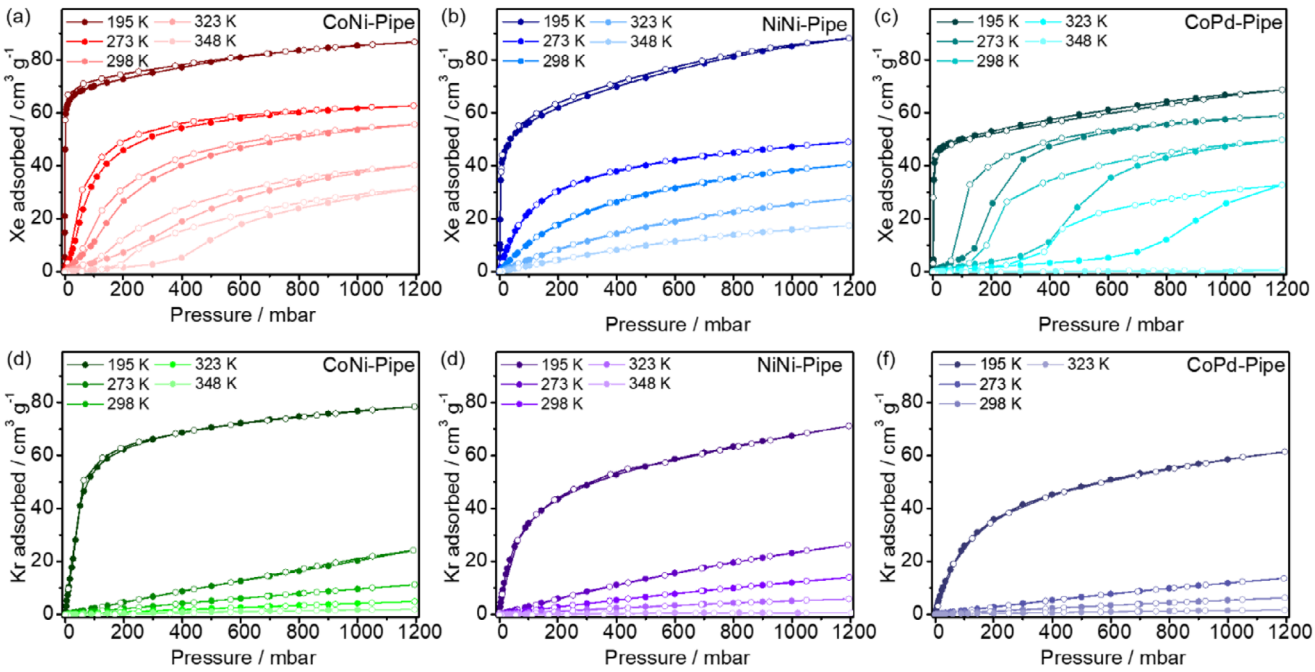


FIGURE 3 | Xe adsorption isotherms for (a) **CoNi-Pipe**, (b) **NiNi-Pipe**, and (c) **CoPd-Pipe** at 195, 273, 298, 323, and 348 K. Kr adsorption isotherms for (d) **CoNi-Pipe**, (e) **NiNi-Pipe**, and (f) **CoPd-Pipe** at the same temperatures. The filled and open symbols indicate adsorption and desorption, respectively.

[30, 59, 60]. All frameworks showed a type I adsorption isotherm, indicative of microporosity. The BET surface areas derived from the CO_2 isotherms were 250 m^2 g^{-1} for both **CoNi-Pipe** and **NiNi-Pipe**, and 233 m^2 g^{-1} for **CoPd-Pipe**. Additionally, pore size distributions calculated using the Horvath–Kawazoe method revealed a consistent pore width range of 3.8–4.3 Å across all frameworks.

2.2 | Xe and Kr Adsorption Performance

The conformational flexibility of the piperazine linkers, along with the appropriately sized pore apertures of **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe**, motivated the investigation of their potential for Xe/Kr separation [61, 62]. Single-component adsorption isotherms for Xe and Kr were measured at 195, 273, 298, 323, and 348 K (Figure 3). Notably, both **CoNi-Pipe** and **CoPd-Pipe** exhibited S-shaped Xe adsorption curves across all the mentioned temperatures, except at 348 K for **CoPd-Pipe** (Figure 3a,c). In contrast, such an S-shaped adsorption behavior was not observed for the **NiNi-Pipe** at any temperature (Figure 3b). Thus, it can be hypothesized that the presence of two distinct pocket-like pore channels in **CoNi-Pipe** and **CoPd-Pipe**, where the window sizes along the four surrounding directions differed, acted as a gating mechanism, enabling Xe to diffuse selectively into specific channels during the initial adsorption stages (Figure 2a; Figures S3, S5, S6, and S8). This selective diffusion likely caused the observed gate-opening behavior. In contrast, the **NiNi-Pipe** featured a single type of pore channel with a relatively larger pore size and more uniform pore windows (Figure 2a; Figures S4 and S7). Conversely, the Kr adsorption isotherms for **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe** exhibited linear adsorption behavior with low uptake, which can be attributed to the smaller atomic size of Kr and its lower diffusion barrier than that of Xe [30].

Among the three frameworks, **CoPd-Pipe** showed the most distinct S-shaped adsorption curves, which can likely be attributed to its more distorted pore structure characterized by narrower pore sizes and smaller pore windows than those of the **CoNi-Pipe** (Figures S5 and S8). Furthermore, the gate-opening pressures for **CoPd-Pipe** showed a significant shift depending on the temperature, occurring at 1 mbar at 195 K, 100 mbar at 273 K, 300 mbar at 298 K, and 600 mbar at 323 K, thereby resulting in a negligible Xe uptake at 348 K. In comparison, the gate-opening pressures for **CoNi-Pipe** were significantly lower, measured at 0.1 mbar at 195 K, 16 mbar at 273 K, 52 mbar at 298 K, 144 mbar at 323 K, and 300 mbar at 348 K (Figure 3a,c). Conversely, the adsorption capability of other flexible MOFs, which show S-shaped adsorption isotherms, was particularly affected by the temperature-dependent gate-opening behavior [17, 20]. Because the gate-opening behaviors of these MOFs are typically associated with a significant phase transition from closed to open pores, they tend to adsorb only when the pore is fully opened within a certain narrow temperature range; otherwise, adsorption does not occur. However, **CoNi-Pipe** and **CoPd-Pipe** followed the thermodynamic adsorption mechanism and exhibited excellent adsorption capabilities over a wide temperature range, even at low temperatures, such as 195 K. This could be attributed to the gate opening arising from the linker flexibility, not just to a dramatic phase change.

To investigate the relationship between gate-opening behavior, phase transition, and Xe mobility, the adsorption rates at different equilibrium pressures (50, 100, 200, 300, 400, 500, and 800 mbar) were calculated. The time-dependent Xe adsorption uptakes were normalized to the equilibrium uptake and fitted using the Avrami kinetic equation:

$$Q_t = Q_e[1 - \exp(-K_t{}^n)]$$

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Where Q_t represents the adsorption capacity at a given time, Q_e is the adsorption capacity at equilibrium, K is the rate constant, t is time, and n is the kinetic order of the model (Figures S24–S27 and Tables S9–S11) [63–65]. For **NiNi-Pipe**, which exhibits a conventional Langmuir-type adsorption, the rate constant K gradually increased with pressure. In contrast, **CoNi-Pipe** and **CoPd-Pipe** exhibited characteristic kinetic features associated with gate-opening transitions. For **CoNi-Pipe**, the rate constant K slightly decreased in the pressure range of 50–100 mbar, where gate opening starts to occur, followed by a pronounced increase above 200 mbar, corresponding to the fully open phase. Similarly, **CoPd-Pipe** showed a decrease in the rate constant within the gate-opening region (300–400 mbar), while the rate constant K increased again at higher pressures once the open phase was fully established. These observations indicate that during the intermediate state of the gate-opening transition (around 100 mbar for **CoNi-Pipe** and 400 mbar for **CoPd-Pipe**), Xe diffusion is temporarily hindered due to structural rearrangement of the framework, resulting in slower adsorption kinetics (Figure S27). The kinetic results thus support that the phase transition from the closed to the open phase governs Xe adsorption dynamics in these flexible MOFs.

To further understand the adsorption behavior, the isosteric heat of adsorption (Q_{st}) was calculated using the dual-site Langmuir–Freundlich model (Figures S28–S33 and Tables S12–S17). The Q_{st} values for **CoNi-Pipe** ranged from 30.9 to 32.7 kJ mol^{-1} for Xe and from 12.3 to 26.2 kJ mol^{-1} for Kr, while those for **NiNi-Pipe** ranged from 29.3 to 31.1 kJ mol^{-1} for Xe and from 10 to 22.5 kJ mol^{-1} for Kr (Figure S34a,b). The Q_{st} curves for both **CoNi-Pipe** and **NiNi-Pipe** exhibited nearly constant values across the entire adsorption range, indicating a strong affinity between Xe and the frameworks.

However, the Q_{st} values for the **CoPd-Pipe** were broader, ranging from 23.4 to 40.4 kJ mol^{-1} for Xe and from 19 to 34.7 kJ mol^{-1} for Kr (Figure S34c). Although the Q_{st} value for Xe at zero coverage was significantly higher than that for Kr, the Q_{st} values for Kr remained higher than those for Xe across the entire adsorption range. The unusual Q_{st} behavior of Xe on **CoPd-Pipe** arises from gate-opening in the flexible framework: initially, strong external/OMS sites are rapidly saturated, causing a drop in Q_{st}; near the gate-opening threshold, structural expansion allows Xe to access adsorption sites, leading to a transient rise, followed by a nearly constant Q_{st} at higher loadings once the open phase is established. In addition, this trend is consistent with the DFT-calculated Xe binding energies. Xe atoms are adsorbed at the open Pd sites with a binding strength of −26.31 kJ mol^{-1}, whereas configurations in which some Xe atoms occupy the weakly interacting internal channels exhibit the reduced binding energy (−11.32 kJ mol^{-1}) (Table S18). This clear hierarchy of binding strengths supports the experimentally observed Q_{st} behavior and indicates that the moderate rise followed by stabilization at higher loadings originates from framework expansion that facilitates Xe adsorption toward the open Pd sites. From the perspective of Xe storage, flexible adsorbents exhibiting S-shaped adsorption isotherms provide advantages in terms of temperature management [19, 66, 67]. Both **CoNi-Pipe** and **CoPd-Pipe** demonstrated relatively high Xe adsorption capacities at 1 bar compared to their respective Q_{st} values at the final coverage (Figure S35). Although pyrazine-based Hofmann MOFs, such as ZJU-74a-Ni and ZJU-74a-Pd, showed high Xe adsorption capacities, their inherently high Q_{st} values necessitated large temperature variations, thereby resulting in high regeneration energy requirements for a practical Xe storage system [32]. Additionally, while the Xe adsorption capacities of the DABCO-based CoNi-DAB and **CoNi-Pipe** were comparable, **CoNi-Pipe** showed a lower Q_{st} value, highlighting the advantages of S-shaped isotherms in reducing regeneration energy requirements [30].

The S-shaped isotherm and low Q_{st} value prompted the investigation of the potential applicability of the frameworks for Xe storage or Xe recovery [68–71]. To quantitatively assess the performance, working capacities (WCs) for pressure swing adsorption (PSA), termed as WC_{PSA}, were determined [6, 19, 72–74]. The WC_{PSA} at 298 K was 45.08, 42.20, and 27.30 cm^3 g^{-1} for the **CoPd-Pipe**, **CoNi-Pipe**, and **NiNi-Pipe** at 1–0.1 bar, respectively (Figures S36). Particularly, both **CoPd-Pipe** and **CoNi-Pipe** showed superior WC_{PSA} values compared to the reported benchmark ultramicroporous adsorbents, especially rigid frameworks with common Langmuir-type isotherms (Figure S36 and Table S19). Such impressive value for the WC_{PSA} in the **CoPd-Pipe** underlines the excellent performance in terms of regeneration energy reduction, thereby implying a promising Xe storage or recovery system.

2.3 | Understanding Xe and Kr Adsorption Behavior Using MLP-MD Simulations

To investigate the diffusion and adsorption behaviors of Xe and Kr in **CoNi-Pie**, **NiNi-Pipe**, and **CoPd-Pipe**, MLP-MD simulations were performed using a 4 × 4 × 4 supercell containing 1,664 atoms. Computational details are provided in Tables S20 and S21 and Figures S37–S39. Figure 4 shows the cumulative displacement and trajectories of Xe atoms in **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe** over 1 ns at 298 K. According to the MLP-MD-simulated Xe trajectories and cumulative displacements, Xe atoms show the highest mobility in **CoNi-Pipe,** followed by **CoPd-Pipe**, whereas Xe atoms barely move in **NiNi-Pipe** for 1 ns. Interestingly, in **CoNi-Pipe**, the Xe atoms can diffuse simultaneously along two distinct channels (Figure 4a). However, in **CoPd-Pipe**, Xe diffusion is more restricted, as atoms tend to migrate along only one of the two distinct channels, with a slight preference for the one with a larger diameter (Figure 4c). In contrast, in **NiNi-Pipe**, the Xe atoms in **NiNi-Pipe** showed reduced displacement compared to those in **CoNi-Pipe** and **CoPd-Pipe**, suggesting a relatively lower mobility (Figure 4b). In addition, the average diagonal distances between Co atoms forming the channel in CoNi-Pipe and CoPd-Pipe expanded from 10.46 and 10.42 Å before Xe diffusion to 10.69 and 10.64 Å during Xe diffusion, respectively (Figure S40). Thus, Xe atoms can effectively spread within the pore channels of both **CoNi-Pipe** and **CoPd-Pipe**.

This study identified that the different Xe adsorption properties of the three MOFs were closely related to the flexibility of their frameworks and that their Xe binding energies are comparable: −32.7 kJ mol^{-1} for **CoNi-Pipe**, −26.3 kJ mol^{-1} for **CoPd-Pipe**, and −30.0 kJ mol^{-1} for **NiNi-Pipe**, which are in good agreement with the experimental values (Figure 3a–c). However, the dynamics of their frameworks exhibited a significantly different

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

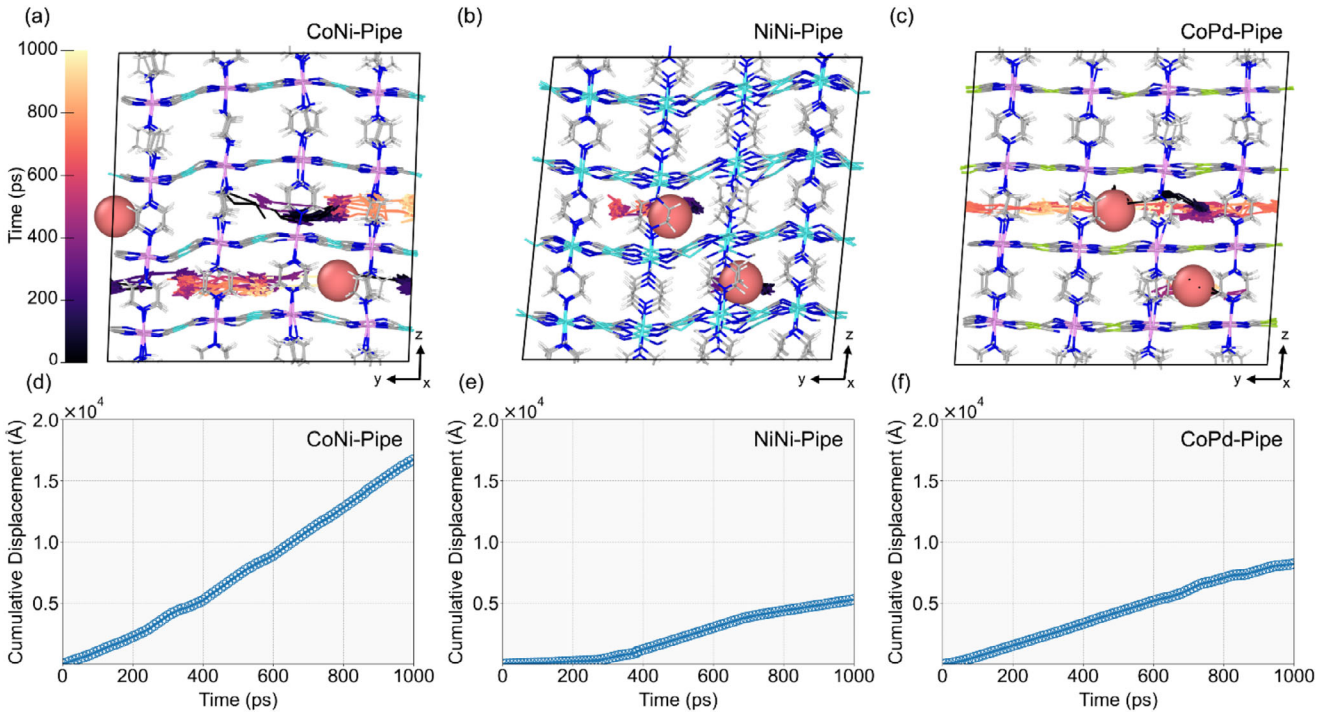


FIGURE 4 | MLP-MD-simulated trajectories and cumulative displacements of two Xe atoms in **CoNi-Pipe** (a and d), **NiNi-Pipe** (b and e), and **CoPd-Pipe** (c and f) at 298K. Color code: Xe (pink-red), Co (pink), Ni (cyan), Pd (yellow-green), N (blue), C (gray), and H (white).

pattern. As shown in Figure 2, all the MOFs comprise layered structures, in which the transition metal atoms are connected by piperazine linkers. By analyzing the cumulative displacements and trajectories of the C and H atoms in the piperazine linkers, the framework flexibility was assessed. In Figures S41–S43, it is evident that both **CoNi-Pipe** and **CoPd-Pipe** show more pronounced fluctuations and rotations of the piperazine linkers than **NiNi-Pipe**. This indicates that both **CoNi-Pipe** and **CoPd-Pipe** possessed greater structural flexibility, thereby creating a more dynamic pore environment that facilitated the adsorption and diffusion of Xe [20, 22]. The schematic diagram of structure changes upon Xe and Kr diffusion within flexible Hoffman-type MOFs is shown in Figure S44.

To investigate the structural factors contributing to the framework flexibility, the possible conformational changes (e.g., chair-to-boat form) of the piperazine linkers were examined. To this end, the N—C—C—N dihedral angles of the 64 piperazine linkers in a 4 × 4 × 4 supercell of all the MOFs were analyzed, as shown in Figure 5, because the dihedral angle is considered a reliable descriptor for capturing conformational changes in the linker [75, 76]. In Figure 5, the red color indicates significant changes in the dihedral angles, corresponding to a transformation into a boat-like conformation. In contrast, the blue color represents minimal changes, suggesting that the system maintains a chair-like conformation. Accordingly, snapshots at 600, 700, 800, 900, and 1000 ps were extracted from the MLP-MD simulations, and the N—C—C—N dihedral angles of the 64 piperazine linkers in the MOFs were evaluated. The piperazine linkers of all the MOFs predominantly adopted the chair conformation. As shown in Figures S45–S50, only minimal changes in the dihedral angles were observed, indicating that the framework flexibility was primarily governed by the rotational

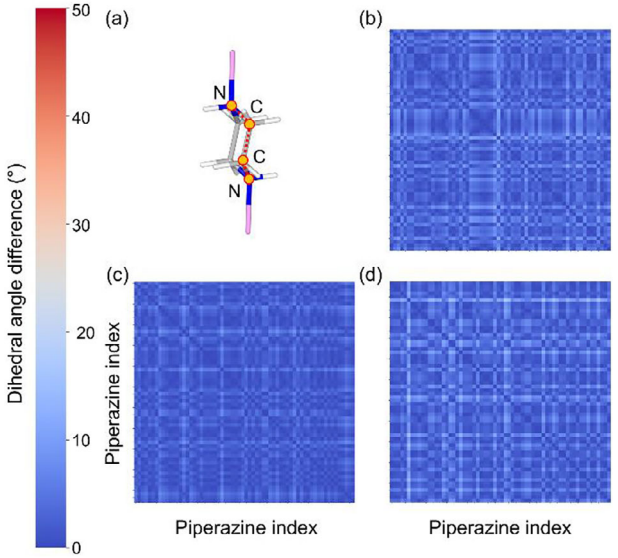


FIGURE 5 | Heatmaps comparing (a) N—C—C—N dihedral angles of the 64 piperazine linkers in three Xe-adsorbed (b) **CoNi-Pipe**, (c) **NiNi-Pipe**, and (d) **CoPd-Pipe** structures. The snapshots were extracted at 1,000 ps from MLP-MD simulations at 298 K.

motion of the piperazine linkers rather than by conformational transitions.

As previously mentioned, this study found that **CoNi-Pipe** provides a more favorable channel size for Xe diffusion than **CoPd-Pipe**. When two Xe atoms occupy each pore, **CoNi-Pipe** shows active Xe diffusion, whereas **CoPd-Pipe** shows more limited diffusion, particularly in slightly smaller pores (Figure 4a,c).

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To gain a deeper understanding of this behavior, the channel size was evaluated by measuring the distance between the two metal atoms (Ni-Ni or Pd-Pd) every 1 ps from 500 to 1,000 ps. Generally, the Ni-Ni distance in **CoNi-Pipe** (~4.1 Å) is larger than the Pd-Pd distance in **CoPd-Pipe** (~3.9 Å), leading to the relatively larger Xe mobility in **CoNi-Pipe** compared to that in **CoPd-Pipe** (Figure S51). Upon analyzing the probability distribution of the interatomic distances in the channel with Xe atoms, the two pore channels in **CoNi-Pipe** (blue curves in Figure S43a) exhibited similar behavior. Interestingly, **CoPd-Pipe** exhibited a distinct bimodal pore size distribution in the first small pore (middle panel of Figure S51a). In particular, the peak near 4.1 Å corresponded to cases in which Xe atoms were adsorbed at specific adsorption sites and became trapped, leading to prolonged residence times, and consequently, hindering Xe diffusion. For the second small pore, where Xe was absent, this peak disappeared (right panel of Figure S51a).

However, as the number of Xe atoms increased, Xe diffusion in **CoPd-Pipe** became more pronounced (Figure S52c). The probability distribution of the interatomic distances in **CoPd-Pipe** no longer showed the previously observed bimodal pore size distribution at higher Xe loadings (Figure S51b). As shown in Figure S53, the framework flexibility of **CoPd-Pipe** was enhanced by the addition of Xe atoms; specifically, the cumulative displacement of C and H atoms in **CoPd-Pipe** became comparable to that observed in **CoNi-Pipe** when three Xe atoms were present in the supercell. This implies that the framework flexibility of **CoPd-Pipe** can be significantly improved at high Xe loadings [77, 78]. This behavior is consistent with our experimental observations, as evidenced by the sigmoidal adsorption isotherm curves of the **CoPd-Pipe** (Figure 3c).

Owing to the smaller vdW radius of Kr compared to that of Xe, Kr exhibited significantly higher mobility in all MOFs, showing a larger cumulative displacement (Figures S54 and S55). Moreover, the Kr binding energies were generally weaker than the Xe binding energies, indicating that they could diffuse more freely through the pore channels (Table S22). Although a weaker binding strength does not always guarantee higher mobility, their combined effect resulted in a lower Kr uptake, which is consistent with the experimental results and single-point DFT calculations, thereby demonstrating a lower adsorption capacity for Kr than for Xe (Figure 3d–f; Figure S56). In addition, the strong adsorption of gas molecules generally plays a crucial role in enhancing the adsorption capacity within porous materials [79, 80]. In contrast, when the adsorption is weak, the confinement of gas molecules is reduced, leading to increased molecular mobility. Such high mobility shortens the residence time of the gas molecules, thereby resulting in an earlier breakthrough and a decrease in overall adsorption capacity [81, 82]. In addition, excessively strong host–guest interactions can induce structural deformations of the framework, such as pore contraction, which in turn reduces the adsorption capacity [83]. This implies that an optimal binding strength is essential for achieving enhanced uptake. Furthermore, these results align with previous studies reporting that Xe is selectively adsorbed over Kr in MOFs, such as MOF-74-based frameworks, UiO-66, and CALF-20 [84, 85].

2.4 | Xe- and Kr-Loaded PXRD Studies

To validate the Xe-induced gate opening and the corresponding subtle structural changes revealed by the MLP-MD simulations, synchrotron PXRD patterns for Xe-loaded MOFs were collected at 298 K (Figure 6; Figure S57). For comparison, the Kr-loaded PXRD data were also collected. Interestingly, the Xe-loaded PXRD patterns for **CoNi-Pipe** and **CoPd-Pipe** were different from those of the pristine patterns, whereas the Kr-loaded PXRD pattern for the **NiNi-Pipe** was retained (Figure S58). The observed changes in the Xe-loaded PXRD patterns for **CoNi-Pipe** and **CoPd-Pipe** corroborate that Xe induced gate opening and structural transformation, thereby resulting in a sigmoidal isotherm. Pawley refinements for the Xe-loaded **CoNi-Pipe** and **CoPd-Pipe** at 298 K were performed to validate the Xe-induced structural variations (Figure 6c,d). Remarkably, the unit cell volume of the Xe-loaded **CoPd-Pipe** increased significantly from 1316.14 $Å^3$ for the pristine MOF to 1495.71 $Å^3$. This indicates that despite the small pore channel size of **CoPd-Pipe**, Xe adsorption was facilitated by the flexibility of the framework, leading to the most pronounced structural transformation and substantial volume expansion (Figure 4; Figure S51). In contrast, the Xe-loaded **CoNi-Pipe** demonstrated a slight volume expansion from 1364.41 $Å^3$ for the pristine MOF to 1374.18 $Å^3$. This is consistent with the MLP-MD simulations, showing that **CoNi-Pipe** has a more favorable channel size for Xe diffusion and thereby showing a moderate structural transformation (Figure 4; Figure S51).

2.5 | Column Breakthrough Tests and Hydrolytic Stability

A Xe/Kr mixture at 20/80 (v/v), acquired as a by-product in air cryogenic distillation, typically undergoes an additional separation to produce high-purity Xe and Kr [86, 87]. To evaluate the dynamic Xe/Kr separation performance of the flexible frameworks, breakthrough experiments were conducted on a binary gas mixture of 4% Xe and 16% Kr balanced with He (Xe/Kr = 20:80, v/v; He should serve as a carrier gas for the experiments) at 273 and 298 K (Figures S59 and S60). Here, 4% Xe and 16% Kr correspond to the adsorbed amounts at 40 and 160 mbar, respectively, in the single-component Xe and Kr isotherms. Because the gate-opening pressures of **CoPd-Pipe** based on the Xe-adsorption isotherms were 100 mbar at 273 K and 300 mbar at 298 K, the difference between the amount of Xe adsorbed at 40 mbar and the amount of Kr adsorbed at 160 mbar was not significant. Therefore, breakthrough experiments were performed only for **CoNi-Pipe** and **NiNi-Pipe**. As expected, for **CoNi-Pipe**, Kr eluted within a short time of 8 min g^{-1} at 273 K and 6 min g^{-1} at 298 K, whereas Xe showed a much longer retention time of 43 min g^{-1} at 273 K and 17 min g^{-1} at 298 K (Figure S59). Similarly, for **NiNi-Pipe**, Kr first eluted within a short time of 8 min g^{-1} at 273 K and 6 min g^{-1} at 298 K, whereas Xe breakthrough did not occur until 58 min g^{-1} at 273 K and 23 min g^{-1} at 298 K (Figure S60). Notably, **NiNi-Pipe** showed a higher separation capability than **CoNi-Pipe,** probably because of the larger gap between the Xe and Kr binding energies in the **NiNi-Pipe** (Figure S34 and Table S22). Moreover, cycling breakthrough experiments were conducted at 298 K, showing

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

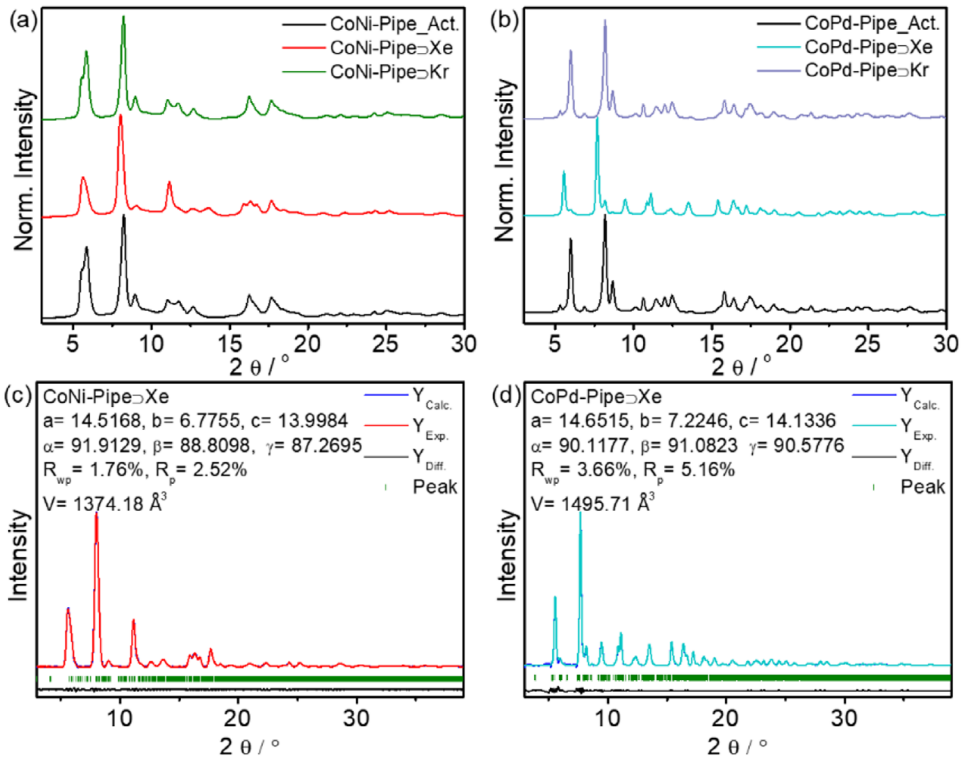


FIGURE 6 | Synchrotron PXRD patterns of (a) activated **CoNi-Pipe**, Xe-loaded **CoNi-Pipe**, and Kr-loaded **CoNi-Pipe**, (b) activated **CoPd-Pipe**, Xe-loaded **CoPd-Pipe**, and Kr-loaded **CoPd-Pipe** at 298 K. Pawley refinement results for synchrotron PXRD patterns of Xe-loaded (c) **CoNi-Pipe** and (d) **CoPd-Pipe**.

complete retention of the separation performance even after mild regeneration under He flow for 6 h at 298 K (Figures S59b and S60b).

Breakthrough experiments at 80% relative humidity (RH) were performed at 298 K to examine the separation capability in the presence of water vapor. Prior to the experiments, water vapor adsorption isotherms were measured at 298 K (Figure S61). For comparison, the isotherms for **CoNi-Pipe**, **NiNi-Pipe**, and **CoPd-Pipe** were analyzed. In all frameworks, the pore filling region was observed in the relative pressure range of 0.3–0.5. Moreover, the structural stability was investigated after exposure to 80% RH for 24 h and air for 1 month. Both **CoNi-Pipe** and **NiNi-Pipe** exhibited excellent hydrolytic and air stabilities, maintaining their structures after exposure to these conditions, as confirmed by the PXRD and IR data (Figures S62 and S63). The breakthrough curves at 80% RH are almost identical to those under dry conditions, displaying the same retention times for Xe and Kr as under dry conditions (Figures S59b, S60b, and S64). Consecutive breakthrough tests were conducted after mild regeneration under He flow for 6 h at 298 K (Figure S64). While **NiNi-Pipe** maintained its separation capability in the second cycle, even under a moist gas stream, **CoNi-Pipe** showed a slightly decreased Xe retention time in the second cycle. However, **CoNi-Pipe** displayed the same Xe retention times in the second and third cycles (Figure S64a). Therefore, it can be expected that the adsorbed water molecules from the first cycle were not fully desorbed under the He flow, which affected the Xe retention time in consecutive cycles. The PXRD analysis revealed that both **CoNi-Pipe** and **NiNi-Pipe** maintained their structural stability after humidity cycles, confirming that water vapor did not affect their structural robustness (Figures S65 and S66). Moreover, IR analysis confirmed that the O-H stretching band at approximately 3300 cm^{-1} of **CoNi-Pipe** was more distinct than that of **NiNi-Pipe** after humid cycles, as measured after the He flow, which supported the presence of unfully desorbed water molecules in the **CoNi-Pipe** system. According to dynamic breakthrough tests under dry and humid conditions, a stronger interaction between Xe and the frameworks was confirmed, implying the potential for practical Xe/Kr separation even under a moist gas stream.

3 | Conclusion

This study demonstrated a series of flexible Hofmann-type MOFs, CoNi-Pipe, NiNi-Pipe, and CoPd-Pipe, and their structural dynamics. The Xe-induced gate-opening behavior was elucidated using MLP-MD simulations. Notably, with a dual-pore channel system arising from the different modes of piperazine, CoNi-Pipe and CoPd-Pipe showed sigmoidal Xe adsorption isotherms. The MLP-MD-calculated Xe trajectories indicated different Xe mobilities and the corresponding diffusion in each framework, supporting the Xe-induced gate-opening and different gate-opening pressures. The MLP-MD-calculated C and H trajectories also validated the pronounced fluctuations and rotations of the piperazine linkers, supporting their structural flexibility and dynamic pore environments. Specifically, the distinct S-shaped isotherm of CoPd-Pipe enabled it to attain outstanding working capacity in the PSA system, providing the potential for practical Xe storage or recovery systems. Furthermore, breakthrough experiments for CoNi-Pipe and NiNi-Pipe revealed excellent Xe/Kr separation potentials. Notably, CoNi-Pipe featured not

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

only excellent Xe/Kr separation, but also Xe storage performance, leveraging the greatest flexibility together with optimal Xe binding strength, as verified by the MLP-MD simulations along with vdW-corrected DFT calculations. Thus, this study provides a pioneering perspective on elucidating the structural flexibility and Xe-induced sigmoidal adsorption behavior in flexible frameworks.

4 | Experimental Section

4.1 | Synthesis of Co(Piperazine)[$Ni(CN)_4$] (CoNi-Pipe)

Potassium tetracyanonickelate hydrate ($K_2[Ni(CN)_4]\cdot H_2O$, 0.1205 g, 0.5 mmol) and deionized H_2O (2 mL) were charged in a 20 mL vial and then sonicated for a few minutes. Subsequently, cobalt(II) nitrate hexahydrate ($Co(NO_3)_2\cdot 6H_2O$, 0.1455 g, 0.5 mmol) was added to the vial. The mixture was sonicated for 1 h to produce a mixture of 2D $Co[Ni(CN)_4]_n$. Piperazine (0.0431 g, 0.5 mmol) was dissolved in a mixture of deionized H_2O (5 mL) and methanol (5 mL). A clear solution of piperazine was then slowly added to a mixture of 2D $Co[Ni(CN)_4]_n$ and sonicated for 5 min. The final mixture was stirred at 40°C for 3 h, and the product was collected by centrifugation and washed thrice with acetone. The product was first solvent-exchanged with acetone several times to remove all guest solvents in the framework. Subsequently, the resultant product was dried overnight under vacuum. Elemental analysis (%) calculated for $Co(Piperazine)[Ni(CN)_4]\cdot 0.95H_2O$, $C_{10}H_{13.4}CoN_6NiO_{0.7}$: C 29.6, H 3.69, N 25.9; found: C 29.6, H 3.38, N 25.5.

4.2 | Synthesis of Ni(Piperazine)[$Ni(CN)_4$] (NiNi-Pipe)

NiNi-Pipe was synthesized according to the synthetic procedure similar to CoNi-Pipe except for the use of $Ni(NO_3)_2\cdot 6H_2O$ (0.1454 g, 0.5 mmol). Elemental analysis (%) calculated for $Ni(Piperazine)[Ni(CN)_4]\cdot 1.1H_2O\cdot 0.1C_3H_6O$, $C_{8.3}H_{12.8}N_6Ni_2O_{1.2}$: C 29.9, H 3.87, N 25.2; found: C 29.7, H 3.66, N 25.1.

4.3 | Synthesis of Co(Piperazine)[$Pd(CN)_4$] (CoPd-Pipe)

CoPd-Pipe was synthesized according to the synthetic procedure similar to CoNi-Pipe except for the use of $K_2[Pd(CN)_4]\cdot H_2O$; 0.144 g, 0.5 mmol. Elemental analysis (%) calculated for $Co(Piperazine)Pd(CN)_4]\cdot 0.95H_2O$, $C_8H_{11.9}CoN_6PdO_{0.95}$: C 22.6, H 3.22, N 22.6; found: C 22.3, H 2.84, N 22.3.

4.4 | Physical Measurements

The crystallographic data and details of the data collection for CoNi-Pipe, NiNi-Pipe, and CoPd-Pipe are listed in Tables S2–S7. Synchrotron PXRD data were collected at the Pohang Accelerator Laboratory (PAL) using a 2D SMC ADSC Quantum-210 detector with a silicon (111) double-crystal monochromator at 10 s exposure and 120 mm detector distance. The ADX program was applied for data collection, and the Fit2D3 program was applied to convert the 2D diffraction image into a 1D diffraction pattern. The structure of CoNi-Pipe, NiNi-Pipe, and CoPd-Pipe was determined using Rietveld refinement to adjust the conformation, molecular position, and orientation of the trial model in the Reflex module of Materials Studio (Accelrys. Materials Studio Getting Started, release 2019) [88]. Several parameters, including cell parameters, preferred orientation, and temperature factor, were considered to obtain the optimal structure for refinement. PXRD patterns were recorded using Cu Kα (λ = 1.5406 Å) on a Rigaku Ultima III diffractometer with a scan speed of 5°/min and step size of 0.02°. Infrared spectra were obtained with an ATR module using a Nicolet iS10 FT–IR spectrometer. TGA analyses were carried out in N_2 (99.999% purity) atmosphere (flow rate = 30 mL min^{-1}) at 30°C–700°C (heating rate = 5°C min^{-1}) using a TA Discovery. Scanning electron microscopy (SEM) images were obtained using SEM-energy dispersive spectrometry (EDS) elemental mapping images were obtained at the Semiconductor & Display Green Manufacturing Research Center at Korea University using a Quanta FEG 250. Elemental analyses of C, H, and N were performed at the Elemental Analysis Service Center of Sogang University. X-ray photoelectron spectroscopy (XPS) data were collected using a K-alpha Plus (Thermo Scientific, USA) at the Hanyang LINC + Analytical Equipment Center (Seoul).

Acknowledgements

This work was supported by the National Research Foundation of Korea (RS-2021-NR057842 and RS-2025-02218827), the KIST Institutional Program (2E33861), the KU-KIST Graduate Schoole of Converging Science and Technology (KU-KIST school program), and the National Center for Materials Research Data (NCMRD) through the National Research Foundation of Korea funded by the Ministry of Science and ICT (2021M3A7C2089739). Computational resources provided by KISTI Supercomputing Center (KSC-2024-CRE-0474). This research was supported by Sookmyung Women's University Research Grants (1-2503-2030). The X-ray crystallographic profile was obtained at the Pohang Accelerator Laboratory (PLS-II BL2D SMC).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

References

1. H. Kim and C. S. Hong, "MOF-74-type Frameworks: Tunable Pore Environment and Functionality Through Metal and Ligand Modification," *CrystEngComm* 23 (2021): 1377–1387, https://doi.org/10.1039/D0CE01870H.

2. J. R. Li, R. J. Kuppler, and H. C. Zhou, "Selective Gas Adsorption and Separation in Metal–Organic Frameworks," *Chemical Society Reviews* 38 (2009): 1477–1504, https://doi.org/10.1039/b802426j.

3. H. Kim, H. Y. Lee, D. W. Kang, et al., "Control of the Metal Composition in Bimetallic Mg/Zn(dobpdc) Constructed From a One-Dimensional Zn-Based Template," *Inorganic Chemistry* 58 (2019): 14107–14111, https://doi.org/10.1021/acs.inorgchem.9b02126.

4. H. Kim, J. H. Choe, H. L. Yun, et al., "High Ammonia Storage Capacity in LiCl Nanoparticle-Embedded Metal-Organic Framework Composites,"

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Chemical Engineering Journal 489 (2024): 151319, https://doi.org/10.1016/j.cej.2024.151319.

5. Q. Gao, J. Xu, D. Cao, Z. Chang, and X.-H. Bu, "A Rigid Nested Metal–Organic Framework Featuring a Thermoresponsive Gating Effect Dominated by Counterions," *Angewandte Chemie International Edition* 55 (2016): 15027–15030, https://doi.org/10.1002/anie.201608250.

6. J. H. Lee, S. Jeoung, Y. G. Chung, and H. R. Moon, "Elucidation of Flexible Metal-organic Frameworks: Research Progresses and Recent Developments," *Coordination Chemistry Reviews* 389 (2019): 161–188, https://doi.org/10.1016/j.ccr.2019.03.008.

7. S. Horike, S. Shimomura, and S. Kitagawa, "Soft Porous Crystals," *Nature Chemistry* 1 (2009): 695–704, https://doi.org/10.1038/nchem.444.

8. H. Sato, W. Kosaka, R. Matsuda, et al., "Self-Accelerating CO Sorption in a Soft Nanoporous Crystal," *Science* 343 (2014): 167–170, https://10.1126/science.124642.

9. N. Hosono, A. Terashima, S. Kusaka, R. Matsuda, and S. Kitagawa, "Highly Responsive Nature of Porous Coordination Polymer Surfaces Imaged by in Situ Atomic Force Microscopy," *Nature Chemistry* 11 (2019): 109–116, https://doi.org/10.1038/s41557-018-0170-0.

10. J. Dong, V. Wee, and D. Zhao, "Stimuli-Responsive Metal–Organic Frameworks Enabled by Intrinsic Molecular Motion," *Nature Materials* 21 (2022): 1334–1340, https://doi.org/10.1038/s41563-022-01317-y.

11. C. M. McGuirk, T. Runcevski, J. Oktawiec, A. Turkiewicz, M. K. Taylor, and J. R. Long, "Influence of Metal Substitution on the Pressure-Induced Phase Change in Flexible Zeolitic Imidazolate Frameworks," *Journal of the American Chemical Society* 140 (2018): 15924–15933, https://doi.org/10.1021/jacs.8b09631.

12. S. Hiraide, Y. Sakanaka, H. Kajiro, S. Kawaguchi, M. T. Miyahara, and H. Tanaka, "High-Throughput Gas Separation by Flexible Metal–Organic Frameworks With Fast Gating and Thermal Management Capabilities," *Nature Communications* 11 (2020): 3867, https://doi.org/10.1038/s41467-020-17625-3.

13. M. Bonneau, C. Lavenn, J. J. Zheng, et al., "Tunable Acetylene Sorption By Flexible Catenated Metal–Organic Frameworks," *Nature Chemistry* 14 (2022): 816–822, https://doi.org/10.1038/s41557-022-00928-x.

14. K. Koupepidou, A. Subanbekova, and M. J. Zaworotko, "Functional Flexible Adsorbents and Their Potential Utility," *Chemical Communications* 61 (2025): 3109–3126, https://doi.org/10.1039/D4CC05393A.

15. S. Krause, N. Hosono, and S. Kitagawa, "Chemistry of Soft Porous Crystals: Structural Dynamics and Gas Adsorption Properties," *Angewandte Chemie International Edition* 59 (2020): 15325–15341, https://doi.org/10.1002/anie.202004535.

16. Y. Huang, J. Wan, T. Pan, et al., "Delicate Softness in a Temperature-Responsive Porous Crystal for Accelerated Sieving of Propylene/Propane," *Journal of the American Chemical Society* 145 (2023): 24425–24432, https://doi.org/10.1021/jacs.3c10277.

17. S. M. Wang, M. Shivanna, S. T. Zheng, et al., "Ethane/Ethylene Separations in Flexible Diamondoid Coordination Networks via an Ethane-Induced Gate-Opening Mechanism," *Journal of the American Chemical Society* 146 (2024): 4153–4161, https://doi.org/10.1021/jacs.3c13117.

18. J. Qiao and D. Zhao, "Ethane-Gated MOFs Meet Industrial Ethylene Purification: Tiny Gates Open Big Possibilities," *Matter* 7 (2024): 2780–2783, https://doi.org/10.1016/j.matt.2024.06.046.

19. I. Senkovska, V. Bon, A. Mosberger, Y. Wang, and S. Kaskel, "Adsorption and Separation By Flexible MOFs," *Advanced Materials* (2025): 2414724, https://doi.org/10.1002/adma.202414724.

20. H. Wang, M. Warren, J. Jagiello, et al., "Crystallizing Atomic Xenon in a Flexible MOF to Probe and Understand Its Temperature-Dependent Breathing Behavior and Unusual Gas Adsorption Phenomenon," *Journal of the American Chemical Society* 142 (2020): 20088–20097, https://doi.org/10.1021/jacs.0c09475.

21. F.-X. Coudert, M. Jeffroy, A. H. Fuchs, A. Boutin, and C. Mellot-Draznieks, "Thermodynamics of Guest-Induced Structural Transitions in Hybrid Organic–Inorganic Frameworks," *Journal of the American Chemical Society* 130 (2008): 14294–14302, https://doi.org/10.1021/ja805129c.

22. I. Senkovska, V. Bon, L. Abylgazina, et al., "Understanding MOF Flexibility: An Analysis Focused on Pillared Layer MOFs As a Model System," *Angewandte Chemie* 62 (2023): 202218076, https://doi.org/10.1002/anie.202218076.

23. A. Schneemann, V. Bon, I. Schwedler, I. Senkovska, S. Kaskel, and R. A. Fischer, "Flexible Metal–Organic Frameworks," *Chemical Society Reviews* 43 (2014): 6062–6096, https://doi.org/10.1039/C4CS00101J.

24. S. Nandi, P. De Luna, R. Maity, et al., "Imparting Gas Selective and Pressure Dependent Porosity Into a Non-porous Solid via Coordination Flexibility," *Materials Horizons* 6 (2019): 1883–1891, https://doi.org/10.1039/C9MH00133F.

25. M. Agrawal, S. Bhattacharyya, Y. Huang, et al., "Liquid-Phase Multicomponent Adsorption and Separation of Xylene Mixtures by Flexible MIL-53 Adsorbents," *The Journal of Physical Chemistry C* 122 (2018): 386–397, https://doi.org/10.1021/acs.jpcc.7b09105.

26. S. Xiang, W. Zhou, J. M. Gallegos, Y. Liu, and B. Chen, "Exceptionally High Acetylene Uptake in a Microporous Metal–Organic Framework With Open Metal Sites," *Journal of the American Chemical Society* 131 (2009): 12415–12419, https://doi.org/10.1021/ja904782h.

27. Q. Wang, T. Ke, L. Yang, et al., "Separation of Xe from Kr with Record Selectivity and Productivity in Anion-Pillared Ultramicroporous Materials By Inverse Size-Sieving," *Angewandte Chemie* 59 (2020): 3423–3428, https://doi.org/10.1002/anie.201913245.

28. L. K. Macreadie, K. B. Idrees, C. S. Smoljan, and O. K. Farha, "Expanding Linker Dimensionality in Metal-Organic Frameworks for sub-Ångstrom Pore Control for Separation Applications," *Angewandte Chemie International Edition* 62 (2023): 202304094, https://doi.org/10.1002/anie.202304094.

29. K. Wen, J. Zhou, T. Ke, et al., "Metal–Organic Framework With Constrained Flexibility for Benchmark Separation of Hexane Isomers," *Angewandte Chemie International Edition* 64 (2025): 202500519, https://doi.org/10.1002/anie.202500519.

30. H. Kim, J. H. Choe, M. Kang, et al., "Switchable Xe/Kr Selectivity in a Hofmann-Type Metal–Organic Framework via Temperature-Responsive Rotational Dynamics," *Small* 19 (2023): 2301905, https://doi.org/10.1002/smll.202301905.

31. J. Pei, K. Shao, J.-X. Wang, et al., "A Chemically Stable Hofmann-Type Metal–Organic Framework With Sandwich-Like Binding Sites for Benchmark Acetylene Capture," *Advanced Materials* 32 (2020): 1908275, https://doi.org/10.1002/adma.201908275.

32. J. Pei, X.-W. Gu, C.-C. Liang, B. Chen, B. Li, and G. Qian, "Robust and Radiation-Resistant Hofmann-Type Metal–Organic Frameworks for Record Xenon/Krypton Separation," *Journal of the American Chemical Society* 144 (2022): 3200–3209, https://doi.org/10.1021/jacs.1c12873.

33. Y. Liu, J. Liu, H. Xiong, et al., "Negative Electrostatic Potentials in a Hofmann-Type Metal-Organic Framework for Efficient Acetylene Separation," *Nature Communications* 13 (2022): 5515, https://doi.org/10.1038/s41467-022-33271-3.

34. J. Ha, M. Jung, J. Park, H. Oh, and H. R. Moon, "Thermodynamic Separation of Hydrogen Isotopes Using Hofmann-Type Metal–Organic Frameworks With High-Density Open Metal Sites," *ACS Applied Materials & Interfaces* 14 (2022): 30946–30951, https://doi.org/10.1021/acsami.2c07829.

35. C. Lou, X. Li, B. E. Keshta, L. Wang, and Y. Zhang, "Hofmann-type Metal-organic Frameworks: A Review on Gas Separation and Purification," *Separation and Purification Technology* 374 (2025): 133691, https://doi.org/10.1016/j.seppur.2025.133691.

36. K. B. Idrees, Z. Chen, X. Zhang, et al., "Tailoring Pore Aperture and Structural Defects in Zirconium-Based Metal–Organic Frameworks for Krypton/Xenon Separation," *Chemistry of Materials* 32 (2020): 3776–3782, https://doi.org/10.1021/acs.chemmater.9b05048.

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

37. K. B. Idrees, Z. Li, H. Xie, et al., "Separation of Aromatic Hydrocarbons in Porous Materials," *Journal of the American Chemical Society* 144 (2022): 12212–12218, https://doi.org/10.1021/jacs.2c03114.

38. K. B. Idrees, K. O. Kirlikovali, C. Setter, et al., "Robust Carborane-Based Metal–Organic Frameworks for Hexane Separation," *Journal of the American Chemical Society* 145 (2023): 23433–23441, https://doi.org/10.1021/jacs.3c04641.

39. X. Zhu, T. Ke, P. Han, et al., "Pore Chemistry and Architecture Control in Anionic Functional Ultramicroporous Materials for Record Dense Packing of Xenon," *Journal of the American Chemical Society* 146 (2024): 24956–24965, https://doi.org/10.1021/jacs.4c06354.

40. E. Acuna-Yeomans, J. J. Gutierrez-Sevillano, S. Calero, and D. Dubbeldam, "Evaluation of ZIF-8 Flexible Force Fields for Structural and Mechanical Properties," *Microporous and Mesoporous Materials* 348 (2023): 112406, https://doi.org/10.1016/j.micromeso.2022.112406.

41. B. Zheng, G. X. Gu, C. D. Santos, R. Neumann Barros Ferreira, M. Steiner, and B. Luan, "Simulating CO_2 diffusivity in rigid and flexible Mg-MOF-74 with machine-learning force fields," *APL Machine Learning* 2 (2024): 026115, https://doi.org/10.1063/5.0190372.

42. J. Behler, "Perspective: Machine Learning Potentials for Atomistic Simulations," *The Journal of Chemical Physics* 145 (2016): 170901, https://doi.org/10.1063/1.4966192.

43. M. Kulichenko, B. Nebgen, N. Lubbers, et al., "Data Generation for Machine Learning Interatomic Potentials and Beyond," *Chemical Reviews* 124 (2024): 13681–13714, https://doi.org/10.1021/acs.chemrev.4c00572.

44. H. J. Lee, H. Kim, S. Ji, et al., "Lithium Localization by Anions in Argyrodite Solid Electrolytes From Machine-Learning-based Simulations," *Advanced Energy Materials* 14 (2024): 2402396, https://doi.org/10.1002/aenm.202402396.

45. M. Eckhoff and J. Behler, "From Molecular Fragments to the Bulk: Development of a Neural Network Potential for MOF-5," *Journal of Chemical Theory and Computation* 15 (2019): 3793–3809, https://doi.org/10.1021/acs.jctc.8b01288.

46. B. Zheng, F. L. Oliveira, R. Neumann Barros Ferreira, et al., "Quantum Informed Machine-Learning Potentials for Molecular Dynamics Simulations of CO_2 's Chemisorption and Diffusion in Mg-MOF-74," *ACS Nano* 17 (2023): 5579–5587, https://doi.org/10.1021/acsnano.2c11102.

47. A. Singh and Y. Li, "Reliable Machine Learning Potentials Based on Artificial Neural Network for Graphene," *Computation Materials Science* 227 (2023): 112272, https://doi.org/10.1016/j.commatsci.2023.112272.

48. M. Kurmoo, H. Kumagai, S. M. Hughes, and C. J. Kepert, "Reversible Guest Exchange and Ferrimagnetism (T_C = 60.5 K) in a Porous Cobalt(II)–Hydroxide Layer Structure Pillared With trans -1,4-Cyclohexanedicarboxylate," *Inorganic Chemistry* 42 (2003): 6709–6722, https://doi.org/10.1021/ic034787g.

49. J. Lannoeye, B. Van de Voorde, B. Bozbiyik, H. Reinsch, J. Denayer, and D. De Vos, "An Aliphatic Copper Metal-organic Framework as Versatile Shape Selective Adsorbent in Liquid Phase Separations," *Microporous and Mesoporous Materials* 226 (2016): 292–298, https://doi.org/10.1016/j.micromeso.2016.01.044.

50. H. Reinsch, J. Benecke, M. Etter, N. Heidenreich, and N. Stock, "Combined in- and Ex Situ Studies of Pyrazine Adsorption Into the Aliphatic MOF Al-CAU-13: Structures, Dynamics and Correlations," *Dalton Transactions* 46 (2017): 1397–1405, https://doi.org/10.1039/C6DT03998G.

51. Y.-Z. Zheng, M. Speldrich, P. Kögerler, and X.-M. Chen, "The role of π–π stacking in stabilizing a,a-trans-cyclohexane-1,4-dicarboxylate in a 2D Co(ii) network," *CrystEngComm* 12 (2010): 1057–1059, https://doi.org/10.1039/B919029E.

52. J. Yuan, M. Yang, B. Yang, et al., "Synthesis of pillar-layered metal–organic frameworks With variable backbones Through sequence control," *Nature Chemistry* 17 (2025): 421–428, https://doi.org/10.1038/s41557-024-01717-4.

53. F. ZareKarizi, M. Joharian, and A. Morsali, "Pillar-layered MOFs: Functionality, Interpenetration, Flexibility and Applications," *Journal of Materials Chemistry A* 6 (2018): 19288–19329, https://doi.org/10.1039/C8TA03306D.

54. N. L. Rosi, J. Kim, M. Eddaoudi, B. Chen, M. O'Keeffe, and O. M. Yaghi, "Rod Packings and Metal−Organic Frameworks Constructed From Rod-Shaped Secondary Building Units," *Journal of the American Chemical Society* 127 (2005): 1504–1518, https://doi.org/10.1021/ja045123o.

55. F. Luo, C. Yan, L. Dang, et al., "UTSA-74: A MOF-74 Isomer With Two Accessible Binding Sites per Metal Center for Highly Selective Gas Separation," *Journal of the American Chemical Society* 138 (2016): 5678–5684, https://doi.org/10.1021/jacs.6b02030.

56. Z. Zhang, Z. Deng, H. A. Evans, et al., "Exclusive Recognition of CO_2 From Hydrocarbons by Aluminum Formate With Hydrogen-Confined Pore Cavities," *Journal of the American Chemical Society* 145 (2023): 11643–11649, https://doi.org/10.1021/jacs.3c01705.

57. A. Spek, "Single-crystal Structure Validation With the Program PLATON," *Journal of Applied Crystallography* 36 (2003): 7–13, https://doi.org/10.1107/S0021889802022112.

58. K. R. Dunbar and R. A. Heintz, "Chemistry of transition metal cyanide compounds: Modern perspectives," *Progress in Inorganic Chemistry* 45 (1997): 283–392, https://doi.org/10.1002/9780470166468.ch4.

59. S. Sharma, S. Mukherjee, A. V. Desai, et al., "Efficient Capture of Trace Acetylene by an Ultramicroporous Metal–Organic Framework With Purine Binding Sites," *Chemistry of Materials* 33 (2021): 5800–5808, https://doi.org/10.1021/acs.chemmater.1c01723.

60. T. Islamoglu, K. B. Idrees, F. A. Son, et al., "Are you Using the Right Probe Molecules for Assessing the Textural Properties of Metal–Organic Frameworks?," *Journal of Materials Chemistry A* 10 (2022): 157–173, https://doi.org/10.1039/D1TA08021K.

61. C. Teixeira, N. Serradji, S. Amroune, et al., "Is the Conformational Flexibility of Piperazine Derivatives Important to Inhibit HIV-1 Replication?," *Journal of Molecular Graphics and Modelling* 44 (2013): 91–103, https://doi.org/10.1016/j.jmgm.2013.05.003.

62. A. Saha, S. Sengupta, A. Virmani, and A. Kumar, "Conformers of Piperazine on Air-water Interface Studied by VSFG Spectroscopy," *Journal of Chemical Sciences* 134 (2022): 98, https://doi.org/10.1007/s12039-022-02093-7.

63. J. D. Martell, P. J. Milner, R. L. Siegelman, and J. R. Long, "Kinetics of Cooperative CO_2 Adsorption in Diamine-Appended Variants of the Metal–Organic Framework Mg_2(dobpdc)," *Chemical Science* 11 (2020): 6457–6471, https://doi.org/10.1039/D0SC01087A.

64. J. H. Choe, H. Kim, M. Kang, et al., "Functionalization of Diamine-Appended MOF-Based Adsorbents by Ring Opening of Epoxide: Long-Term Stability and CO_2 Recyclability Under Humid Conditions," *Journal of the American Chemical Society* 144 (2022): 10309–10319, https://doi.org/10.1021/jacs.2c01488.

65. J. H. Choe, H. Kim, H. Yun, et al., "Boc Protection for Diamine-Appended MOF Adsorbents to Enhance CO_2 Recyclability Under Realistic Humid Conditions," *Journal of the American Chemical Society* 146 (2024): 646–659, https://doi.org/10.1021/jacs.3c10475.

66. H. Demir, M. Mobedi, and S. Ülkü, "Effects of Porosity on Heat and Mass Transfer in a Granular Adsorbent Bed," *International Communications in Heat and Mass Transfer* 36 (2009): 372–377, https://doi.org/10.1016/j.icheatmasstransfer.2009.01.008.

67. J. Wieme, S. Vandenbrande, A. Lamaire, V. Kapil, L. Vanduyfhuys, and V. Van Speybroeck, "Thermal Engineering of Metal–Organic Frameworks for Adsorption Applications: A Molecular Simulation Perspective," *ACS Applied Materials & Interfaces* 11 (2019): 38697–38707, https://doi.org/10.1021/acsami.9b12533.

68. S. K. Elsaidi, D. Ongari, M. H. Mohamed, et al., "Metal Organic Frameworks for Xenon Storage Applications," *ACS Materials Letters* 2 (2020): 233–238, https://doi.org/10.1021/acsmaterialslett.9b00468.

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

69. Y. Choi, K. M. Choi, and K. Park, "Four-Bed Vacuum Pressure Swing Adsorption for Xenon Recycling From Semiconductor Waste Gas," *Separation and Purification Technology* 359 (2025): 130477, https://doi.org/10.1016/j.seppur.2024.130477.

70. E. J. Karwacki, T. C. Golden, B. Ji, S. A. Motika, and T. S. Farris, "Xenon Recovery System," U.S. Patent 7285154, filed November 24, 2004, and issued October 23, 2007.

71. P. J. Ryan, O. K. Farha, L. J. Broadbelt, R. Q. Snurr, and Y.-S. Bae, "Metal-Organic Frameworks for Xe/Kr Separation," U.S. Patent 8518153, filed August 17, 2011, and issued August 27, 2013.

72. E. Perez-Botella, S. Valencia, and F. Rey, "Zeolites in Adsorption Processes: State of the Art and Future Prospects," *Chemical Reviews* 122 (2022): 17647–17695, https://doi.org/10.1021/acs.chemrev.2c00140.

73. T. M. McDonald, J. A. Mason, X. Kong, et al., "Cooperative Insertion of CO_2 in Diamine-Appended Metal-Organic Frameworks," *Nature* 519 (2015): 303–308, https://doi.org/10.1038/nature14327.

74. S. Xiong, Q. Liu, Q. Wang, et al., "A Flexible Zinc Tetrazolate Framework Exhibiting Breathing Behaviour on Xenon Adsorption and Selective Adsorption of Xenon Over Other Noble Gases," *Journal of Materials Chemistry A* 3 (2015): 10747–10752, https://doi.org/10.1039/C5TA00460H.

75. Y. Wang, L. Feng, K. Zhang, et al., "Uncovering Structural Opportunities for Zirconium Metal–Organic Frameworks via Linker Desymmetrization," *Advanced Science* 6 (2019): 1901855, https://doi.org/10.1002/advs.201901855.

76. S. Mahajan and M. Lahtinen, "3D Zinc–Organic Frameworks Based on Mixed Thiophene Dicarboxylate and 4-Amino-3,5-bis(4-pyridyl)-1,2,4-triazole Linkers: Syntheses, Structural Diversity, and Single-Crystal-to-Single-Crystal Transformations," *Crystal Growth & Design* 2024, 24, 747–762, https://doi.org/10.1021/acs.cgd.3c01172.

77. P. Zhao, H. Fang, S. Mukhopadhyay, et al., "Structural Dynamics of a Metal–Organic Framework Induced by CO_2 Migration in its Non-Uniform Porous Structure," *Nature Communications* 10 (2019): 999, https://doi.org/10.1038/s41467-019-08939-y.

78. A. Orellana-Silla, M. Meneses-Sanchez, R. Turo-Cortes, M. C. Munoz, C. Bartual-Murgui, and J. A. Real, "Symmetry Breaking and Cooperative Spin Crossover in a Hofmann-Type Coordination Polymer Based on Negatively Charged $\{FeII(\mu2\text{-}[MII(CN)_4])_2\}_n^{2n-}$ Layers (M^{II} = Pd, Pt)," *Inorganic Chemistry* 62 (2023): 12783–12792.

79. L. Huang, C. Xiong, Y.-L. Wang, X.-P. Fu, L. Chen, and Q.-Y. Liu, "Metal-organic Frameworks With Optimal Pore Cages for High Xe Adsorption and Efficient Xe/Kr Separation," *Separation and Purification Technology* 352 (2025): 128138, https://doi.org/10.1016/j.seppur.2024.128138.

80. R. A. Maia, B. Louis, W. Gao, and Q. Wang, "CO_2 Adsorption Mechanisms on MOFs: A Case Study of Open Metal Sites, Ultra-Microporosity and Flexible Framework," *Reaction Chemistry & Engineering* 6 (2021): 1118–1133, https://doi.org/10.1039/D1RE00090J.

81. S.-J. Lee, K. C. Kim, T.-U. Yoon, M.-B. Kim, and Y.-S. Bae, "Selective Dynamic Separation of Xe and Kr in Co-MOF-74 Through Strong Binding Strength Between Xe Atom and Unsaturated Co_{2+} Site," *Microporous and Mesoporous Materials* 236 (2016): 284–291, https://doi.org/10.1016/j.micromeso.2016.09.005.

82. Q. Liu, Y. Gong, B. Liu, S. Xiong, H.-M. Wen, and X. Wang, "Dense Packing of Xenon in an Ultra-Microporous Metal–Organic Framework for Benchmark Xenon Capture and Separation," *Chemical Engineering Journal* 453 (2023): 139849, https://doi.org/10.1016/j.cej.2022.139849.

83. S. Krause, V. Bon, I. Senkovska, et al., "A Pressure-amplifying Framework Material With Negative Gas Adsorption Transitions," *Nature* 532 (2016): 348–352, https://doi.org/10.1038/nature17430.

84. Y. Yang, C. Tu, L. Guo, L. Wang, F. Cheng, and F. Luo, "Metal-organic Frameworks for Xenon and Krypton Separation," *Cell Reports Physical Science* 4 (2023): 101694, https://doi.org/10.1016/j.xcrp.2023.101694.

85. J. Francis Kurisingal, D. W. Kim, and C. S. Hong, "Effective Approaches to Boost Xe/Kr Separation in Metal-Organic Frameworks: A Review," *Coordination Chemistry Reviews* 507 (2024): 215731, https://doi.org/10.1016/j.ccr.2024.215731.

86. M. Aneke and M. Wang, "Improving the Energy Efficiency of Cryogenic Air Separation Units (ASU) Through Compressor Waste Heat Recovery Using Direct Binary Heat Engine Cycle," *Computer Aided Chemical Engineering* 37 (2015): 2375–2380.

87. G. Yu, Y. Liu, X. Zou, N. Zhao, H. Rong, and G. Zhu, "A Nanosized Metal–Organic Framework With Small Pores for Kinetic Xenon Separation," *Journal of Materials Chemistry A* 6 (2018): 11797–11803, https://doi.org/10.1039/C8TA03509A.

88. Accelrys, Materials Studio, Release 2019, Dassault Systèmes BIOVIA, San Diego, CA, USA, 2019.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting File: smll72122-sup-0001-SuppMat.pdf

16136829, 2026, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202509479 by Sookmyung Womens University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

이영준 교수

논문

- V. Arun, E. Kim, H. Y. Kim, H.-J Park, W. Kim, Y. Lee, E. Kim (2025) Bisindolizine: A new AIEgen and its application in latent fingerprint imaging. Dyes and Pigments, 242, 112991.

Dyes and Pigments 242 (2025) 112991

Contents lists available at ScienceDirect

Dyes and Pigments

journal homepage: www.elsevier.com/locate/dyepig

Bisindolizine: A new AIEgen and its application in latent fingerprint imaging

V. Arun [a,1], Eunsu Kim [a,1], Hyeon Young Kim [a], Hyun-Ji Park [a,c], Wook Kim [a,c,d], Youngjun Lee [b], Eunha Kim [a,c,d,*]

[a] Department of Molecular Science and Technology, Ajou University, Suwon, 16499, South Korea
[b] Department of Chemistry, Sookmyung Women's University, Seoul, 04310, South Korea
[c] Advanced College of Bio-convergence Engineering, Ajou University, Suwon, 16499, South Korea
[d] Department of Otolaryngology, Ajou University School of Medicine, Suwon, 16499, South Korea

ARTICLE INFO

Keywords:
Bisindolizine
Aggregation induced emission
Latent fingerprint
Molecular symmetry

ABSTRACT

The concept of molecular symmetry was rationally applied to transform a monomeric asymmetric indolizine moiety into a dimeric symmetric bisindolizine, resulting a new fluorescent molecular framework having appealing aggregation induced emission (AIE) properties and a bathochromic shift in emission wavelength. The synthetic protocol utilized an operationally straightforward Pd-catalyzed dehydrogenative coupling reaction. Among the synthesized derivatives, the most efficient AIE compound was identified and employed for latent fingerprint (LFP) imaging. The LFP visualization protocol involves the application of the AIEgen solution to the fingerprint surface, followed by washing with water to reveal detailed fingerprint pattern.

1. Introduction

Traditional fluorophores typically undergo fluorescence quenching upon aggregation. In contrast, a distinct class of fluorophores exhibits enhanced emission in the aggregated or solid state—a phenomenon known as 'aggregation induced emission' (AIE) [1–4]. These fluorophores are collectively referred to as AIE luminogens (AIEgens). In our recent study, we developed an indolizine-based chemical framework with tunable emission wavelengths, optimized for the synthesis of diverse AIEgens [5,6]. This advancement enabled selective imaging of mitochondria and endoplasmic reticulum, highlighting the framework's potential as a versatile molecular palette for cellular bioimaging [7]. Given the advantages of longer-wavelength emission for biological and other visualization applications [8], we aimed to extend the emission range of the indolizine-based framework beyond the visible spectrum to develop improved AIE molecules for broader bio applications.

Over the past decade, considerable research efforts have focused on tuning the emission wavelength (λ_{em}) of fluorescent molecules [9]. Recent studies have highlighted that modulating the molecular symmetry of fluorophores offers an innovative strategy to induce a bathochromic shift in λ_{em} [10]. Based on the literature proceedings [11–13], to create symmetrical fluorophore the molecule should contain two identical electron donating group (EDG) and two identical electron withdrawing group (EWG) in a symmetrical fashion. Moreover, the strength of the EDG and EWG is playing one of the major key roles to induce the bathochromic shift in symmetrical fluorophore. Despite notable progress, tuning the molecular symmetry of fluorophores remains a relatively new approach (Fig. 1). For example, Erhong et al. developed a series of BODIPY dimers (**1a**) that exhibited significantly red-shifted absorption (λ_{abs}) and λ_{em}, along with large Stokes shifts [10]. Similarly, bisbenzo [*f*]isoindolylidenes (**1b**), reported by Hongli and co-workers, showed a red-shifted λ_{em} in the range of 545–640 nm, compared to their asymmetric monomeric counterpart [14]. Another example includes bisbenzofulvenes (**1c**), which demonstrated favorable optoelectronic properties [15]. Additionally, Pu and Chen synthesized symmetrical bipyrazolo [1,5-a]pyridine (**1d**) and explored their AIE behavior [16]. More recently, a single benzene-core fluorophore system (**1e**), bearing symmetrically positioned donor and acceptor groups, displayed a notably long λ_{em} (618 nm) [17]. Collectively, these findings underscore the promise of molecular symmetry as a powerful molecular design strategy for developing new fluorophores with extended λ_{abs} and λ_{em}.

* Corresponding author. Department of Molecular Science and Technology, Ajou University, Suwon, 16499, South Korea.
E-mail address: ehkim01@ajou.ac.kr (E. Kim).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.dyepig.2025.112991
Received 26 March 2025; Received in revised form 20 June 2025; Accepted 24 June 2025
Available online 26 June 2025
0143-7208/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

In this study, we successfully modulated the molecular symmetry of indolizine monomers **4a-c** to synthesize a new series of AIEgens, bisindolizines, **5a-c**, which exhibited bathochromically shifted λ_{em}. The synthesis was achieved using an operationally simple Pd-catalyzed dehydrogenative coupling reaction. Among the resulting compounds, bisindolizine **5a** demonstrated the most favorable AIE properties and was selected for application in latent fingerprint (LFP) imaging A key feature of the LFP detection protocol is its solution-phase approach: the **5a** solution is applied to the fingerprint surface, and a simple water wash subsequently reveals the hidden fingerprint pattern.

2. Results and discussion

2.1. DFT study

Prior to the chemical synthesis of dimeric indolizines, we performed DFT calculations on indolizin-3-yl (phenyl)methanone—a representative monomeric building block of the bisindolizine scaffold—to investigate its electronic distributions at the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) levels using B3LYP/6-31G (d,p) method (Fig. 2). The results revealed that the C_9 position in the indolizine moiety contributed most significantly to the HOMO (Fig. 2, left), indicating that introducing an electron-rich π-conjugated system at this position could elevate the HOMO energy level and reduce the HOMO-LUMO energy gap. At the LUMO level, a pronounced electron density shift toward the pendant phenyl group was observed (Fig. 2, right), suggesting that *para*-functionalization of the phenyl ring could further modulate the excited-state properties of indolizine systems. Based on these insights, we aimed to design a bisindolizine π-conjugation system at the C_9 position to create a bisindolizine scaffold, while introducing diverse functional groups onto the pendant phenyl moiety to enhance electronic versatility.

2.2. Chemical synthesis

Aligned with our initial vision, we recognized the importance of establishing an efficient synthetic protocol for this study. Inspired by recent advancements in metal-catalyzed dehydrogenative strategies [18], [-19] we noted that such approaches eliminate the need for pre-functionalization of starting materials, enabling the synthesis of target molecules in fewer steps with minimal byproducts—offering a clear advantage over conventional coupling reactions [20]. Accordingly, our envisioned symmetrically molecular scaffold was successfully constructed using a metal-catalyzed process [21]. As shown in Scheme 1a, the synthesis of bisindolizine scaffolds proceed via a three-step protocol. First, the commercially available picoline **1** and various 4-substituted 2-bromoacetophenones (**2a-c**) were reacted in acetone at 80 °C to afford the corresponding pyridinium salts (**3a-c**). Using a previously reported method [22], these intermediates were then converted to indolizin-3-yl (aryl)methanone (**4a-c**) via an intermolecular condensation reaction in the presence of *N*,*N*′-dimethylformamide, dimethyl sulfate, and triethylamine. This step yielded indolizine monomers (**4a-c**) in 55 %–65 % (Scheme 1b). Finally, a metal-catalyzed oxidative dimerization of **4a-c** using $Pd(OAc)_2$ and K_2CO_3 in DMF at 100 °C for 12 h afforded desired symmetrical bisindolizine products [1,1′-biindolizine]-3,3′-diylbis (arylmethanone) **5a-c** in moderate yields (20–30 %) (Scheme 1c). As a result, we successfully synthesized bisindolizine compounds with neutral (**5a**), electron-donating (**5b**), or electron-withdrawing (**5c**) groups attached to the dimeric phenyl moieties. Detailed experimental procedures and characterization data for all compounds are provided in the supporting information.

Based on the literature proceedings [23], we proposed that the reaction initiates with an electrophilic palladation of **4** by $Pd(OAc)_2$, forming intermediate **A**. A second molecule of **4** then undergoes a similar palladation and reacts with intermediate **A** to generate species **B** (Scheme 1d). Subsequent reductive elimination from **B** affords the desired bisindolizine product **5** via C–C bond formation, accompanied by the release of Pd (0). Finally, aerial oxidation and the acetic acid byproduct facilitate the regeneration of the active Pd(II) catalyst.

2.3. Photophysical properties

We next compared the photophysical properties of indolizine monomers (**4a-c**) and their corresponding bisindolizine dimers (**5a-c**) to examine the impact of dimerization on the indolizine scaffold (Table 1, Fig. 3). Initial observations indicated that dimerization did not markedly alter the absorption properties. Although modest bathochromic shift of approximately 20 nm in the absorption maxima was observed, along with a decrease in molar absorptivity. In contrast, the AIE behavior exhibited pronounced differences between indolizine monomeric and dimeric species. Notably, the emission maxima of the dimers (**5a-c**) displayed a substantial bathochromic shift—averaging around 70 nm—relative to the monomers monomers (**4a-c**), leading to significantly larger Stokes shifts (Table 1, Fig. 3). These findings suggests that the dimeric bisindolizines undergo more extensive electron redistribution upon photoexcitation than their monomeric counterparts. Considering

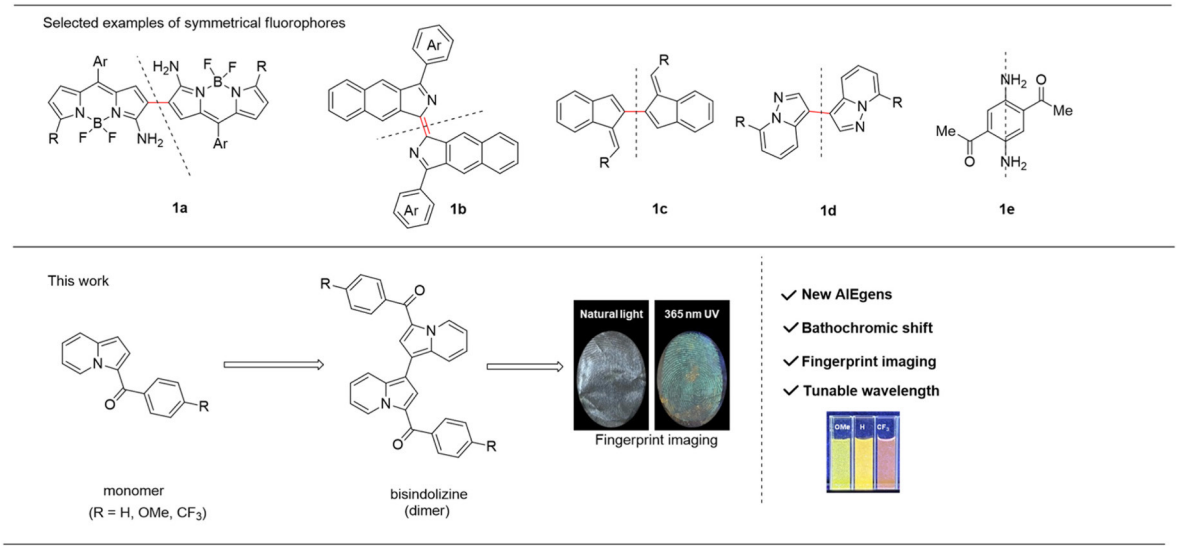


Fig. 1. Chemical structures of selected symmetrical fluorophores (up) and a schematic representation summarizing the current work (down).

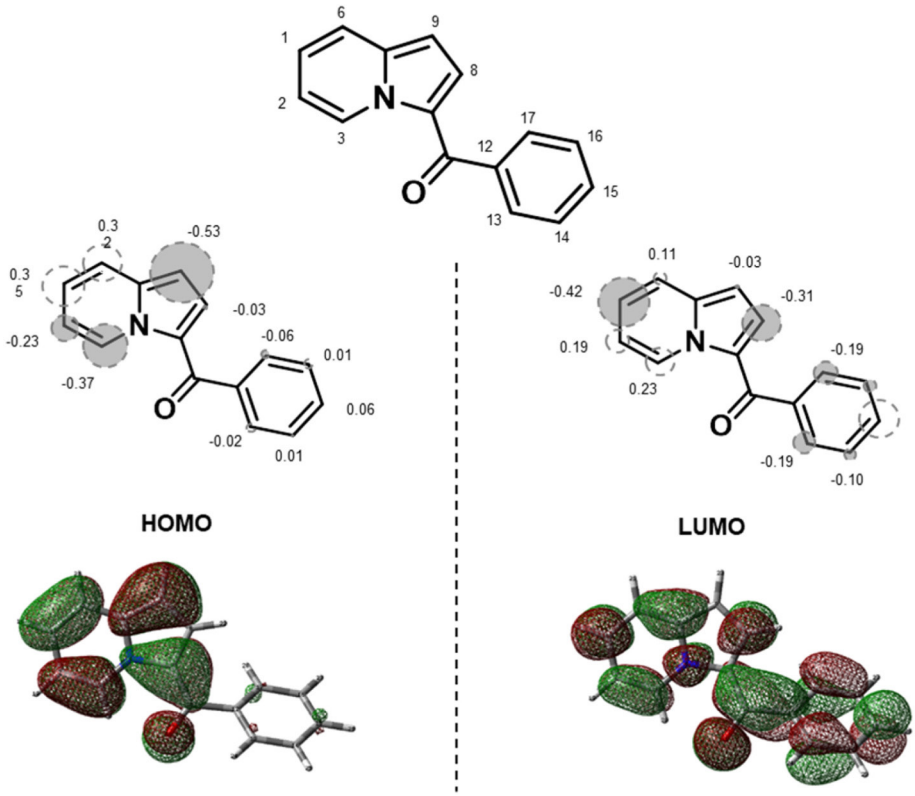
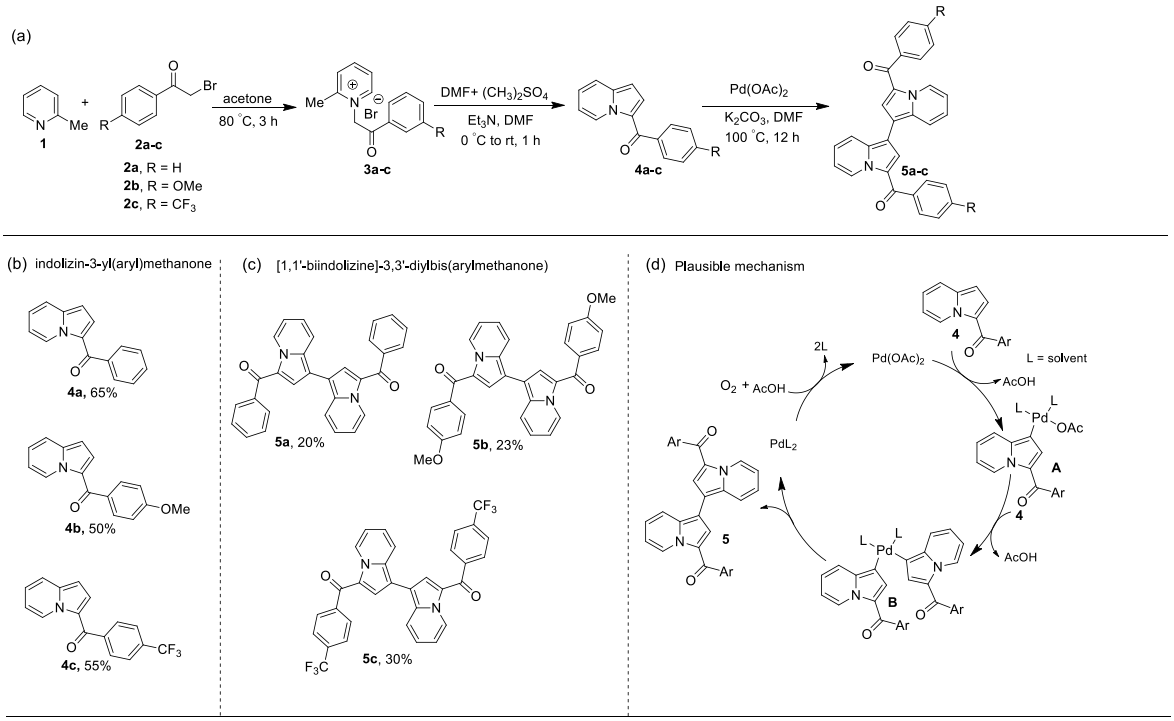


Fig. 2. Electron density distributions in the HOMO and LUMO were predicted by Gaussian 09 at the B3LYP/6-31G (d,p) method.



SCHEME 1. (a) Synthetic route of bisindolizine **5** (b) chemical structure of indolizine monomers, **4a** (65 %), **4b** (50 %), **4c** (55 %) 65 %) (c) chemical structure of bisindolizine products **5a** (20 %), **5b** (23 %), **5c** (30 %) and (d) plausible mechanism depiction for the dimerization process.

the effects of functional groups, we observed that the emission wavelengths of the monomeric indolizines (**4a-c**) remained relatively consistent across neutral, electron-donating and electron-withdrawing groups, ranging from 446 nm to 437 nm. Similarly, the bisindolizine dimers (**5a-c**) showed emission maxima in a narrow range of 537–558 nm, indicating that the nature of the substituents had a limited effect on emission wavelength. To further understand these trends, we analyzed the frontier orbital energy levels of **5a–5c**, using TD-DFT calculations

Table 1
Photophysical property of **4a-c** and **5a-c**.

Cpd[a]	R	λ_{Abs}[b]	λ_{Em}[c]	ε
4a	H	375	446	24,674
4b	OMe	377	442	26,404
4c	CF_3	378	437	22,393
5a	H	397	548	25,181
5b	OMe	394	537	31,683
5c	CF_3	403	558	11,955

[a] Absorbance and emission spectra of indolizine monomers (**4a-c**) and bisindolizines (**5a-c**) were measured in MeCN and in aqueous DMF (10 % water), respectively, at a concentration of 20 μM.
[b] Only the longest absorption maxima are shown.
[c] Excited at the maximum absorption wavelength.

based on their excited-state optimized structures. The results revealed that **5b**, bearing the most electron donating group, exhibited the highest LUMO energy and the largest HOMO–LUMO gap (2.08 eV), followed by **5a** (1.88 eV) and **5c** (1.56 eV) (Table S1). These computational findings correlated well with the experimental emission data, where **5b** showed shortest emission wavelength (537 nm), followed by **5a** (548 nm) and **5c** (558 nm). Notably, all bisindolizine dimers exhibited a significant bathochromic shift in λ_{em} compared to their counterparts, which is likely attributed to the extended π-conjugation between the pyrrole components within the bisindolizine framework.

Interestingly, the emission intensity increased significantly under aqueous DMF conditions. For example, the emission intensities of **5a**, **5b**, and **5c** are 46-, 3- and 6-fold higher, respectively, compared to their corresponding monomers, **4a**, **4b** and **4c** under identical measurement conditions (equal concentration with equal spectroscopic settings).

2.4. Mechanistic understanding of AIE property

Given the known AIEgen properties of the indolizine core skeleton [6], [-7] we hypothesized that the enhanced emission intensities observed for compounds **5a-5c** under aqueous conditions result from their improved AIE behavior compared to their monomeric counterparts. To evaluate this, we examined the emission properties of bisindolizine dimers (**5a-5c**) at a concentration of 20 μM in DMF/water mixtures with increasing water fraction (f_w). Consistent with previous findings, dimer **5a** displayed very weak emission in pure DMF. However, as the f_w gradually increased, a pronounced enhancement in fluorescence intensity was observed (Fig. 4a and j).

A notable increase in emission intensity was observed when the f_w exceeded 70 % (Fig. 4d). This enhancement coincided with the formation of aggregates, with an average particle size of 1272 nm and a polydispersity index of 0.398 (Fig. 4g). When the f_w reached 99 %, dimer **5a** showed a 75-fold increase in fluorescence intensity compared to 0 % (Fig. 4a and j) while the emission maximum remained constant (λ_{em} = 547 nm). For dimer **5b**, weak emission was observed in DMF, but 17-fold enhancement of emission intensity was observed when f_w reached 99 % (Fig. 4b). Similarly, **5C** also exhibited increased emission intensity with f_w increase (Fig. 4c). However, the emission intensities of **5b** and **5c** was significantly weaker than that of **5a** (Fig. 4j vs 4k and 4l). In addition, we observed that dimer **5b** and **5c** formed irregular aggregates with a random size distribution at higher f_w (Fig. 4e and f). As a result, we confirmed that dimer **5a** shows the most robust AIE properties. Therefore, we focused on **5a** for the application of AIE property of bisindolizine.

Based on the previous experimental results and literature proceedings [7], we believe that restricted intramolecular motion (RIM) and other ICT mechanisms may be responsible for the AIE phenomena of the bisindolizines. To confirm ICT mechanism of bisindolizines, solvent-dependent properties (solvatochromism) of the representative bisindolizine, **5a**, were explored (Fig. 5).

A series of solvents with varying polarities [24] were selected, including THF (37.4), ethyl acetate (38.1), DMF (43.8), and MeOH (55.4). As the solvent polarity decreased (from MeOH to THF), dimer **5a** exhibited an increase in fluorescence intensity, with the highest fluorescence intensity observed in the least polar solvent, THF. Notably, no significant shift in emission wavelength was observed across the different solvent polarities (Fig. 5b). These findings suggest that the ICT mechanism may not fully account for the AIE phenomenon of **5a**.

Next, we investigated the RIM mechanism by measuring the emission profiles of **5a-c** in glycerol/methanol mixtures with increasing volume fractions of glycerol (f_g) (Fig. 6).

Notably, a significant 40-fold increase in the emission intensity was observed with increasing f_g (Fig. 6a and d). Similarly, **5b** (Fig. 6b and e) and **5c** (Fig. 6c and f) 4-fold and 10-fold enhancements, respectively. These results suggest that molecular motion between the two indolizine rings at the 3-3′ bond position contributes to non-radiative energy decay. Collectively, these findings clearly demonstrate that the synergistic

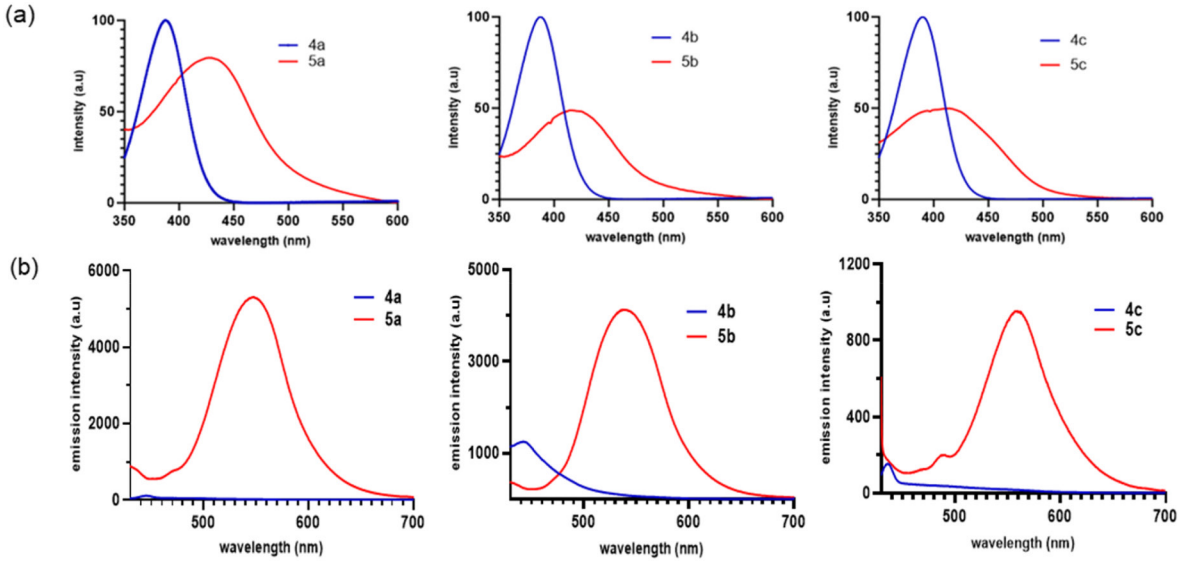


Fig. 3. (a) Absorbance and (b) emission spectra of indolizine monomers (blue line, **4a-c**) and bisindolizine (red line, **5a-c**). Absorbance and emission spectra were measured in MeCN and in aqueous DMF (10 % water), respectively, at a concentration of 20 μM. Excitation wavelength for monomer **4a-c** is 385 nm and for dimer **5a-c** is 407 nm. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

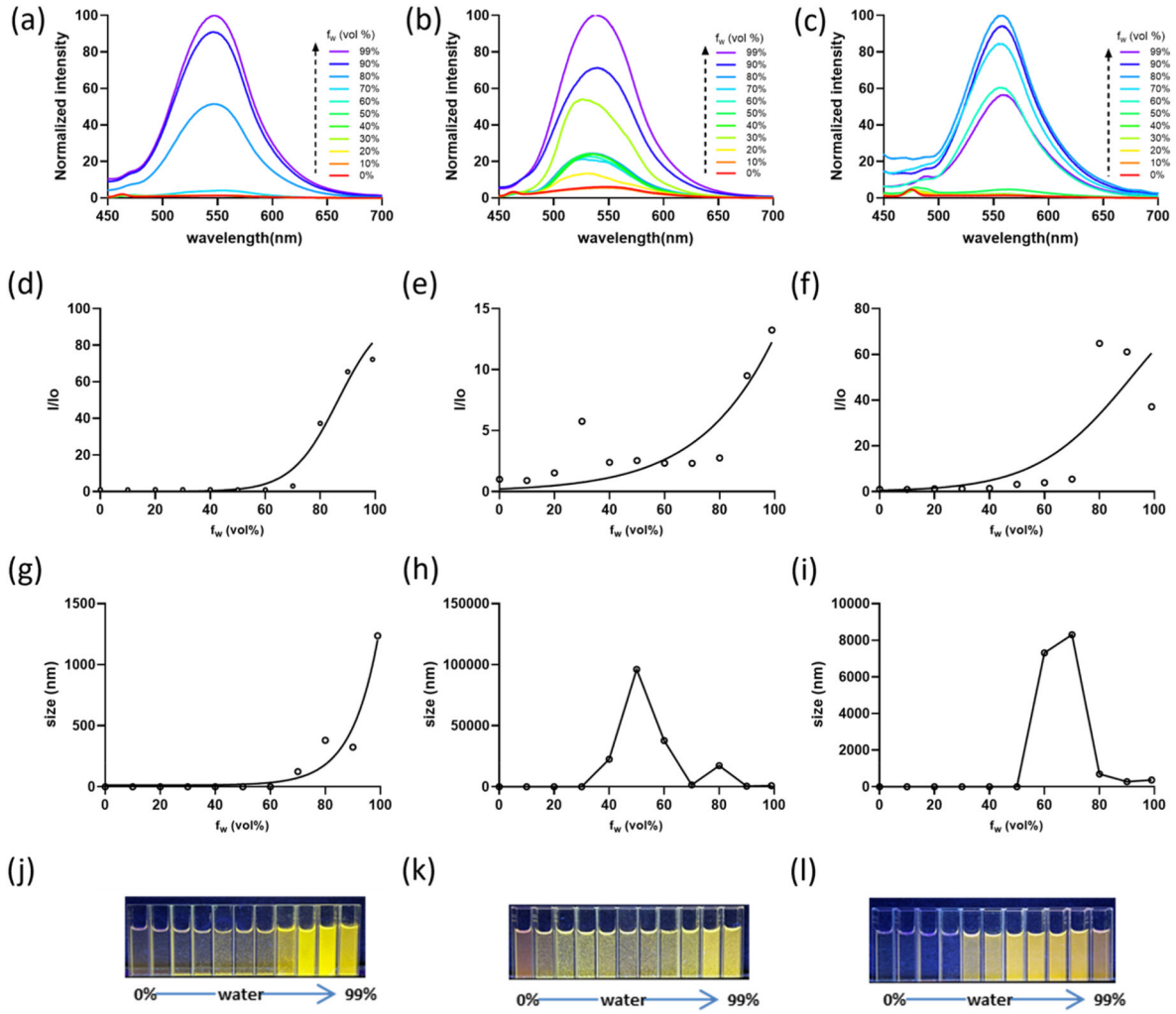


Fig. 4. The AIE property of compounds **5a** (a,d,g,j) **5b** (b,e,h,k), and **5c** (c,f,i,l). a-c) emission spectra of **5a**, **5b** and **5c** (20 μM) in different volume faction of water (f_w) in DMF/water mixture. d-f) intensity ratio of **5a**, **5b** and **5c** between in DMF (I_0) and in different DMF/water mixtures (I) with indicated f_w value. g-i) Particle size measurement of **5a**, **5b** and **5c** in DMF/water mixtures (20 μM) with indicated f_w value. j-l) photographic images of compounds **5a**, **5b** and **5c** (20 μM) in different f_w in DMF/water mixtures under longwave UV irradiation (365 nm).

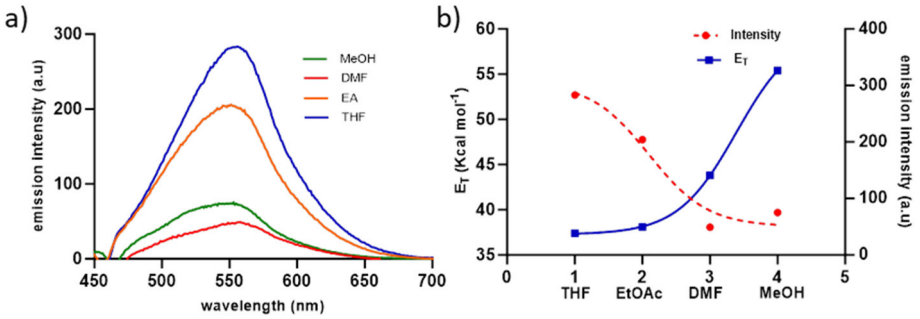


Fig. 5. Solvatochromism study of **5a**. a) Emission spectra of **5a** in different solvents (MeOH, DMF, EtOAc and THF). b) Scatter plot of emission intensity and solvent polarity index (E_T) in different solvents.

interplay between ICT and RIM mechanisms plays a crucial role in the AIE behavior of **5a-c**.

2.5. *Latent fingerprint imaging*

Fingerprints are unique to each individual and serve as a critical tool for personal identification. The detection of latent fingerprints (LFPs) is a routine procedure in forensic investigations. However, because LFPs are invisible to the naked eye, effective visualization methos are essential. In recent years, AIE-based imaging of LFPs has attracted considerable attention due to its high contrast, sensitivity, and simplicity [25–30]. Despite these advancements, many practical still rely on using luminescent or magnetic materials, which can potentially damage the fingerprint patterns and pose health risk to the forensic

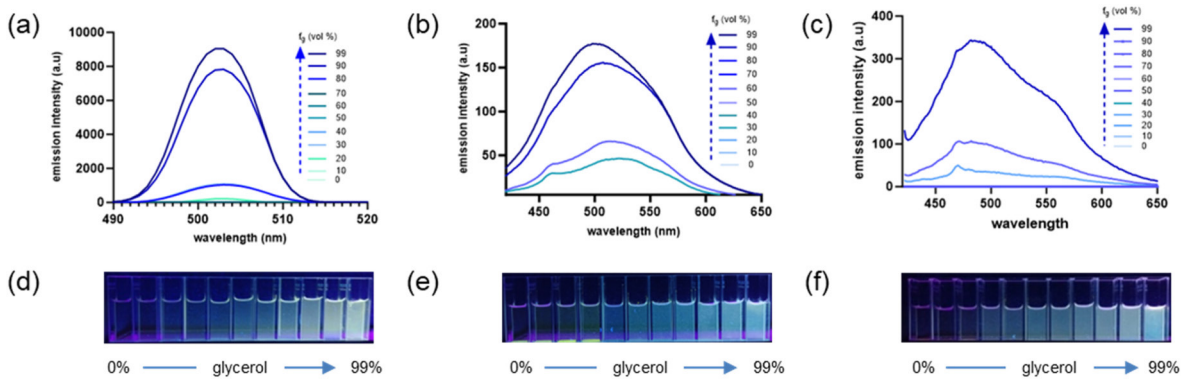


Fig. 6. Mechanism study for the AIE property of **5a-c**, (a–c) Emission spectra of **5a**, **5b** and **5c** (20 μM) in different volume faction of glycerol (f_g) in glycerol/MeOH mixture. d-f) photographic images of compounds **5a**, **5b** and **5c** (20 μM) in different f_g in glycerol/MeOH mixture under longwave UV irradiation (365 nm).

examiners [31]. To address these limitations, the development of safer and more efficient solution-phase detection methods is highly desirable.

In this study, a solution containing bisindolizine **5a** was applied to fingerprint-bearing surfaces, and the patterns were visualized using a smartphone. Under natural light, the fingerprints were invisible, however, upon UV irradiation, the fingerprint patterns became clearly distinguishable (Fig. 7a). This method successfully enabled LFP detection on various substrates (Fig. 7b). A 1 mM solution of **5a** in an organic solvent was applied to the slide glass bearing fingerprints and rinsed with water after 1 min. Fluorescence microscopy confirmed the presence of well-resolved fingerprint patterns. In contrast, treatment with the monomeric **4a**, did not produced a discernible fingerprint pattern (Fig. 7c).

3. Conclusion

In summary, we successfully synthesized a series of new AIE-active bisindolizine dimers (**5a**–**c**) by modulating the molecular symmetry of the monomeric indolizines (**4a**–**c**), resulting in a bathochromic shift in emission wavelengths. Among them, compound **5a** exhibited the most pronounced AIE properties and was applied for latent fingerprint (LFP) imaging. The key advantage of the developed LFP detection protocol lies in its simplicity—dispersing the **5a** solution onto a fingerprint surface followed by a water rinse effectively reveals hidden fingerprint patterns. Compared to its monomeric counterpart **4a**, bisindolizine **5a** demonstrated superior photophysical performance, establishing it as a promising tool for LFP visualization. This study underscores the potential of molecular symmetry modulation as a viable strategy for designing high-performance AIEgens for advanced imaging and biological applications. Inspired by these findings, further development of novel fluorophores is

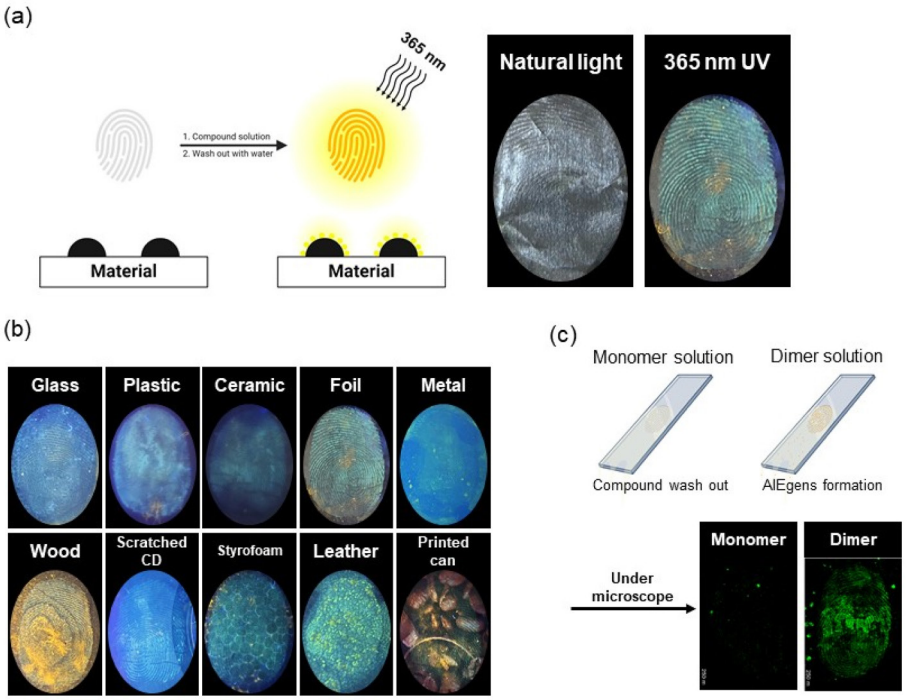


Fig. 7. Representative Bisindolizine (**5a**) and indolizine monomer (**4a**) were used for latent fingerprinting. (a) Schematic illustration of the latent fingerprint development process, highlighting the effects of AIEgens treatment and the resulting patterns observed under natural light and UV light. (b) Results of latent fingerprint visualization on various surfaces, including glaas, plastic, ceramic, foil, metal, wood, scratched CD, styrofoam, leather and printed can. (c) Latent fingerprints revealed through treatment with bisindolizine (**5a**) and monomer (**4a**), alongside a fluorescence imaging comparison. UV lamp (visualization at 365 nm, power 8 W).

currently underway, with results to be reported in due course.

CRediT authorship contribution statement

V. Arun: Writing – original draft, Data curation. **Eunsu Kim:** Writing – original draft, Formal analysis, Data curation. **Hyeon Young Kim:** Data curation. **Hyun-Ji Park:** Formal analysis, Resources, Writing – review & editing. **Wook Kim:** Formal analysis, Resources, Writing – review & editing. **Youngjun Lee:** Writing – review & editing, Writing – original draft, Data curation. **Eunha Kim:** Writing – review & editing, Writing – original draft, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grants (RS-2024-00411474, RS-2023-00208819) funded by the Ministry of Science and ICT. GRRC program of Gyeonggi province (GRRCAjou2023-B03) and by the Ajou University research fund.

Supporting Information

Detailed synthetic procedures for the monomeric indolizines **4a**–**4c** and bisindolizine dimers **5a**–**5c**, detailed procedures for the photophysical data measurements of dimers **5a**–**5c** (AIE, solvatochromism and RIM), DLS experiment procedure, fingerprint imaging procedure, DFT details, characterization of compounds **4a**–**4c** and **5a**–**5c**, copies of ^{1}H and ^{13}C NMR copies and LRMS data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dyepig.2025.112991.

Data availability

Data will be made available on request.

References

[1] Mei J, Leung NL, Kwok RT, Lam JW, Tang BZ. Aggregation-induced emission: together we shine, united we soar. Chem Rev 2015;115(21):11718–940.
[2] Wang S, Zhou K, Lyu X, Li H, Qiu Z, Zhao Z, Tang BZ. The bioimaging story of AIEgens. Chem Biomed Imaging 2023;1(6):509–21.
[3] Chen Y, Lam JW, Kwok RT, Liu B, Tang BZ. Aggregation-induced emission: fundamental understanding and future developments. Mater Horiz 2019;6:428–33.
[4] Duo Y, Luo G, Zhang W, Wang R, Xiao GG, Li Z, Li X, Chen M, Yoon J, Tang BZ. Noncancerous disease-targeting AIEgens. Chem Soc Rev 2023;52:1024–67.
[5] Kim E, Lee Y, Lee S, Park SB. Discovery, understanding, and bioapplication of organic fluorophore: a case study with an indolizine-based novel fluorophore, Seoul-Fluor. Acc Chem Res 2015;48(3):538–47.
[6] Choi SK, Rho J, Yoon SE, Seok JH, Kim H, Min J, Yoon W, Lee S, Yun H, Kwon OP, Kim JH. Full color tunable aggregation-induced emission luminogen for bioimaging based on an indolizine molecular framework. Bioconjug Chem 2020;31(11):2522–32.
[7] Arun V, Choi SK, Han JH, Choi H, Kim HM, Kim W, Choi J, Kim J, Kim E. Harnessing aggregation-induced emission property of indolizine derivative as a fluorogenic bioprobe for endoplasmic reticulum. Dyes Pigments 2022;200:110118.
[8] Wu J, Shi Z, Zhu L, Li J, Han X, Xu M, Hao S, Fan Y, Shao T, Bai H, Peng B. The design and bioimaging applications of NIR fluorescent organic dyes with high brightness. Adv Opt Mater 2022;10:2102514.
[9] Lovell TC, Branchaud BP, Jasti R. An organic chemist's guide to fluorophores–understanding common and newer non-planar fluorescent molecules for biological applications. Eur J Org Chem 2024;27:e202301196.
[10] Wang L, Wu Q, Kang Z, Guo X, Miao W, Li Z, Zuo H, Wang H, Si H, Jiao L, Hao E. Regioselective synthesis of directly connected BODIPY dimers through oxidative coupling of α-amino-substituted BODIPYs. Org Lett 2023;25:5055–60.
[11] Li H, Li Q, Gu Y, Wang M, Tan P, Wang H, Han L, Zhu Y, He F, Tian L. Dimerization extends π-conjugation of electron donor-acceptor structures leading to phototheranostic properties beyond the sum of two monomers. Aggregate 2024;e528.
[12] Kim J, Oh JH, Kim D. Recent advances in single-benzene-based fluorophores: physicochemical properties and applications. Org Biomol Chem 2021;19:933–46.
[13] Jung Y, Jin JH, Kim Y, Oh JH, Moon H, Jeong H, Kim J, Park YK, Oh Y, Park S, Kim D. Development of a fluorescent nanoprobe based on an amphiphilic single-benzene-based fluorophore for lipid droplet detection and its practical applications. Org Biomol Chem 2022;20:5423–33.
[14] Ye C, Huang R, Chiou MF, Wang B, Li D, Bao H. Synthesis of a new fluorophore: wavelength-tunable bisbenzo [f] isoindolylidenes. Chem Sci 2023;14:13151–8.
[15] Mondal S, Ballav T, Mohammad TS, Ganesh V. Bis-benzofulvenes: synthesis and studies on their optoelectronic properties. Org Lett 2023;25:3941–5.
[16] Hsiao PY, Chu JH. Novel bipyrazolo [1, 5-a] pyridine luminogens with aggregation-induced emission enhancement properties. Chem. Commun Now 2021;57:12281–4.
[17] Kim H, Park W, Kim Y, Filatov M, Choi CH, Lee D. Relief of excited-state antiaromaticity enables the smallest red emitter. Nat Commun 2025;202(12).
[18] Yeung CS, Dong VM. Catalytic dehydrogenative cross-coupling: forming carbon− carbon bonds by oxidizing two carbon− hydrogen bonds. Chem Rev 2011;111(r 3):1215–92.
[19] Albano G. Palladium-catalyzed cross-dehydrogenative coupling of (hetero) arenes. Org Chem Front 2024;11(5):1495–622.
[20] Arun V, Mahanty K, De Sarkar S. Nickel-catalyzed dehydrogenative couplings. ChemCatChem 2019;11(9):2243–59.
[21] Priyanka Rani P, Kiran Sindhu J. Indolizine: a promising framework for developing a diverse array of C−H functionalized hybrids. ChemistrySelect 2023;8:e202203531.
[22] Przewloka T, Chen S, Xia Z, Li H, Zhang S, Chimmanamada D, Kostik E, James D, Koya K, Sun L. Application of DMF–methyl sulfate adduct in the regioselective synthesis of 3-acylated indolizines. Tetrahedron Lett 2007;48(33):5739–42.
[23] Hsiao PY, Chang R, Sue AC, Chu JH, Liao GW, Lee YH, Huang JY. Synthesis and mechanistic investigation of bipyrazolo [1, 5-a] pyridines via palladium-catalyzed cross-dehydrogenative coupling of pyrazolo [1, 5-a] pyridines. J Org Chem 2022;87(15):9851–63.
[24] Reichardt C. Solvatochromic dyes as solvent polarity indicators. Chem Rev 1994;94(8):2319–58.
[25] Bera S, Selvakumaraswamy A, Nayak BP, Prasad P. Aggregation-induced emission luminogens for latent fingerprint detection. Chem Commun 2024;60:8314–38.
[26] Lian J, Meng F, Wang W, Zhang Z. Recent trends in fluorescent organic materials for latent fingerprint imaging. Front Chem 2020;8:594864.
[27] Tian R, Zhu FY, Ma R, Wang YL, Huang J, Li C, Zhu MQ. Instant in situ highlighting of latent fingerprints by a green fluorescent probe based on aggregation-induced emission. Biosens Bioelectron 2024;263:116572.
[28] Liu L, Wang J, Wang X, Wang H, Li M, Wu T, Gao G, Zheng X, Liu G, Fan L, Shen W. High-resolution imaging of latent fingerprints through near-infrared organoboron AIEgens. Chin J Chem 2023;41:1465–70.
[29] Zhou Y, Villani E, Kurioka T, Zhang Y, Tomita I, Inagi S. Fabrication of luminescent patterns using aggregation-induced emission molecules by an electrolytic micelle disruption approach. Aggregate 2023;4:e202.
[30] Wang J, Li M, Gao H, Xie L, Zheng X, Liu G, Wu T, Lin L, Liu L. Photoresponsive tetracoordinate arylboron smart molecules: strategies for molecular design and photoresponse mechanisms. Smart Mol 2024;2:e20240041.
[31] Xu L, Li Y, Li S, Hu R, Qin A, Tang BZ, Su B. Enhancing the visualization of latent fingerprints by aggregation induced emission of siloles. Analyst 2014;139:2332–5.

이한길 교수

논문

- Y. Kim, J. Kim, S.-i Choi*, H. Lee*, H. S. Ahn* (2026). Structure Dependent Accessibility of Active Sites Governs Catalytic Activity and Stability of Iridium Oxides in the Acidic Oxygen Evolution Reaction, J. Am. Chem. Soc. 148, 957-966.

- J. Choi, H. Jeon, S. Hwang, H. Lee*, S. Hong*, K. Kang* (2025). Synthetic Neuromelanin Particles for Metal-Free Photocatalysis, Small, 21(52) e10311, 1-7

- S. Hwang, H. Yu, H.-A Ju, H. Kim, J. Baik, Y.-M. Kim*, H. S. Kim*, H. Lee* (2025) In situ Study of Visible Light-Activated Photocatalytic Conversion Reaction Enhancement of Biomass Using a Magnetic Heterostructure, Materials Horizons.12(23) 10270-10280

- H. Park, S. Hwang, J. Choi, K. Kang*, H. Lee*, S. Hong* (2025). Defect engineering in OH-functionalized carbon nanotubes for visible-light-driven ROS-mediated anticancer therapy and high-performance photocatalysis, Nanoscale. 17, 25783-25794.

- S. Hwang, J. Kim, H. Kim, J. Baik, M.-C. Kim*, H. Lee* (2025). Enhanced visible-light photocatalytic conversion of biomass by calcination of Ir-doped SnO2 nanoparticles with a narrowed bandgap using NaBH4, Chem. Eng. J. 523(168457), 1-8.

- S. K. Perumal, H. Yu, H. Jae, H. Lee*, Y. Kim*, H. S. Kim* (2025). Cr-Promoted Pt@$Fe2_{-x}CrxO_3$ Hematite-Type Magnetic Catalysts for Efficient Base- Free HMF Oxidation under Mild Conditions, ACS Sustainable Chemistry & Engineering, 13(41) 17470-17482.

- Y.-H. Kim, M. Shin, E.-B. Park, S. Hwang, J. A. Hong, J. Jang, Y.-M. Kim*, H. Lee* (2025). Enhancing the Multifunctional Photocatalytic Activity of Sustainable Magnetic Nanoparticles, small structures, 6 (2500125), 1-10.

- H. Jeon, S. Hwang, T. Kang, S. Hong*, H. S. Ahn*, H. Lee* (2025). Impact of Oxygen Vacancy Density in Modified Cobalt Oxide Nanoparticles on the Efficiency of Visible-light Photocatalysis, Inorg. Chem. Communications, 177(114287) 1-10.

- H. Jeon, D. T. Hoang, H. Yu, S. Hwang, H. S. Kim*, S. Hong*, H. Lee* (2025). Enhancing Visible-Light-Driven Photocatalysis of Pd- and Pt-Doped WO3 Nanoparticles: Role of Oxygen Vacancies and Bandgap Narrowing, Dalton Transactions, 54(13) 5546-5553.

- icles via Cr and Co ion doping: Insights into surface oxygen defect structures, H. Jeon, D. T. Hoang, I. Y. Kim*, H. Lee*, S. Hong* (2025). Enhanced photocatalytic and electrocatalytic properties of IrO2 nanopart Appl. Surf. Sci. 686(162088), 1-9.

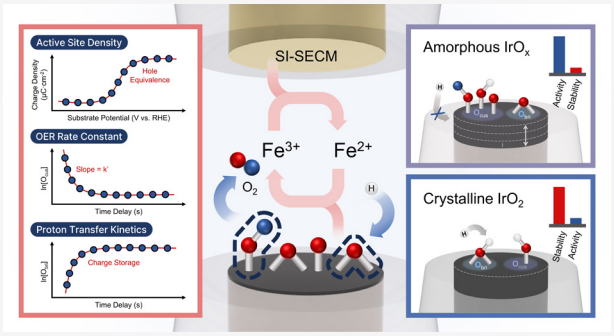
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Structure Dependent Accessibility of Active Sites Governs Catalytic Activity and Stability of Iridium Oxides in the Acidic Oxygen Evolution Reaction

Yeonsu Kim, Jeonghyeon Kim, Sang-Il Choi,* Hangil Lee,* and Hyun S. Ahn*

Cite This: J. Am. Chem. Soc. 2026, 148, 957−966
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Deep understanding of the structural influence on the activity−stability relationship of iridium oxides in acidic oxygen evolution reaction (OER) is desired for better catalyst designs and, consequently, improved atom efficiency of the rare-earth element. However, identifying the variables governing the reaction mechanism remains a challenge in environments where multistep reactions and catalyst degradation occur simultaneously. Surface interrogation scanning electrochemical microscopy (SI-SECM) was employed to directly quantify the transient reactivity of OER intermediates on iridium oxide surfaces (amorphous IrO_x and crystalline IrO_2). Electrochemical measurements revealed a structure dependence of the electrolyte-accessible depth of active site layers, the variation of which controls the kinetics of charge storage and ultimately governs the OER catalysis. Contrary to the common models in the literature, effective charge storage exhibited little influence on the lowering of the OER kinetic barrier. Rather, the activity of iridium oxide catalysts was governed by the electrolyte permeability of the material, which controlled the number of active sites.

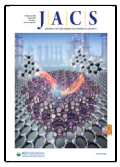


1. INTRODUCTION

Climate change induced by fossil-fuel-driven activities amplified the demand for the development of renewable energy technologies. One of the most attractive large-scale energy storage solutions is the production of green hydrogen via proton exchange membrane water electrolysis (PEMWE);[1,2] however, under highly acidic and oxidizing conditions, iridium oxide remains the sole catalyst capable of achieving a balance between activity and stability for the kinetically sluggish oxygen evolution reaction (OER) at the anode.[3,4] Unfortunately, iridium is scarce and the reported catalytic durability is limited.[5−7] Therefore, a fundamental mechanistic understanding of the acidic OER on iridium oxides is desired for the preparation of catalysts that meet the stability and activity requirements to be competitive at high current scales.

The activity and stability of a catalyst are sharply influenced by its crystal structure.[8] In the case of iridium oxides, amorphous IrO_x is documented to exhibit superior activity for OER, with a trade-off in stability, compared to that of crystalline IrO_2.[9−11] Important OER intermediates are electron-deficient oxygen species (O^{1-}, oxygen species with more electrophilic hole character), which are categorized into two sites with distinct redox behavior: (1) bridging oxygen (O_{bri}, μ_2-O) and (2) coordinatively unsaturated site oxygen (O_{cus}, μ_1-O).[12−15] Recent investigations attempted to differentiate the two species in potential space and ascertain the respective roles in catalysis; however, the exact functions of O_{bri} and O_{cus} in the OER remain unclear. Inspired by Jones and co-workers' establishment of the correlation between O^{1-} charge accumulation (i.e., increased surface hole (h^+) coverage (θ^{h+})) and OER activation energy,[16] significant research efforts were focused on elucidating the crystal structure dependence of the reactivity of the electron-deficient oxygen species.[17,18] In 2022, an *operando* X-ray spectroscopy investigation by Mom and co-workers demonstrated that the degree of deprotonation of the lattice oxygen species depended on sample crystallinity, and the deprotonated oxygen population influences the kinetics of both OER and iridium dissolution.[10] Recent work by Stephens and co-workers revealed that crystal structure dependent active site coverage regulates the balance between *O binding energy and adsorbate−adsorbate interactions, thereby controlling the reaction kinetics.[19] Despite these insightful findings, the fundamental link between the lattice structure of the catalyst and the redox behavior (both the intermediate adsorbate and the active site) remains to be further elucidated. The main challenge lies in the lack of

Received: September 23, 2025
Revised: December 4, 2025
Accepted: December 5, 2025
Published: December 29, 2025

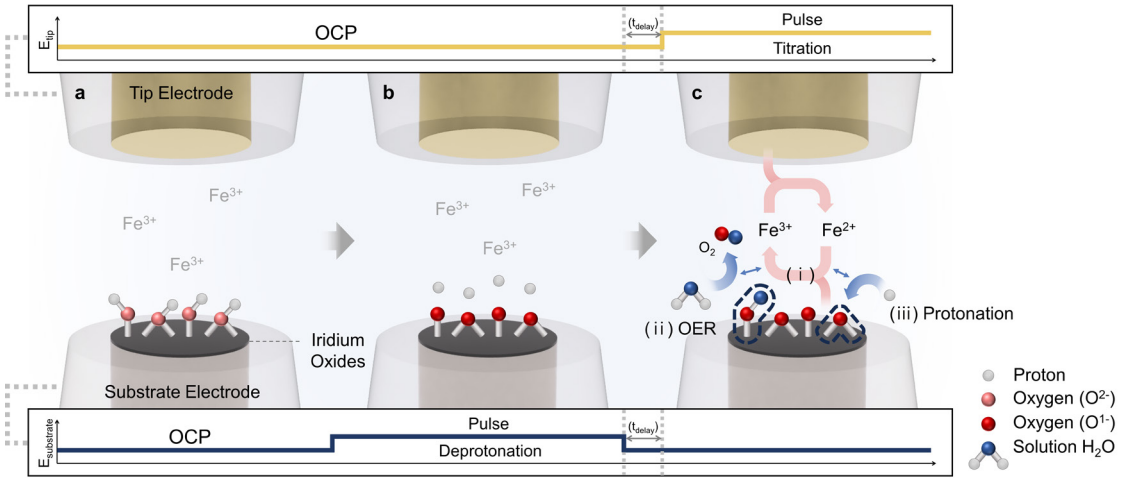


Downloaded via SOOKMYUNG WOMEN'S UNIV on January 15, 2026 at 00:11:09 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

© 2025 American Chemical Society

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

Scheme 1. Representation of the SI-SECM Redox Titration Sequence: (a) Resting at Open Circuit Potential (OCP), (b) Titrand (O^{1-}) Generation by Substrate Pulse, and (c) Titrand Consumption by (i) Titrant (Fe^{2+}) Generation, through (ii) OER, and (iii) Protonation during Delay Time



quantitative analytical means for the detection of the transient surface redox species involved in the catalytic cycle.[20,21]

In this regard, surface interrogation scanning electrochemical microscopy (SI-SECM) offers an attractive approach for catalyst surface analysis.[22−24] The SI-SECM technique enables time-resolved detection of electrolyte-accessible surface sites, independent of the underlying bulk. Notably, this titration technique has provided critical mechanistic insights into various electrochemical energy conversion reactions by accurate quantification of the transient reactivity of surface species, including active site density, intermediate kinetics, and lifetimes.[25−29]

In this work, we systemically identify the key factors influencing the activity and stability of two catalyst structures, amorphous IrO_x and crystalline IrO_2 by SI-SECM. We directly quantified the active site density and activity of electron-deficient oxygen species for each catalyst. In the process, the much-discussed charge storage mechanism of iridium oxide in the acidic OER was empirically demonstrated through time-dependent titrations. Quantitative determination of the proton transfer kinetics of the electron-deficient oxygen species revealed the link between the structure-dependent electrolyte permeability and the ease of oxidative charge accumulation. Facilitated oxidative charge accumulation exhibited little impact on lowering the activation energy of the OER, and the key factor driving the catalytic activity was the electrolyte-accessible active site density. Furthermore, the penetration of electron-deficient oxygen species (μ_2 and μ_1) into the oxide layers lowered the connectivity of the lattice support, leading to facile iridium dissolution.

2. RESULTS AND DISCUSSION

Amorphous IrO_x and crystalline IrO_2 exhibit a complementary relationship in terms of catalytic activity and stability according to literature precedents[9−11] and our experimental data: compared to IrO_2, IrO_x showed *ca.* 150 mV cathodic overpotential for the OER (Figure S1), while its *S*-number (defined as the ratio between the amount of evolved oxygen and dissolved iridium, $n_{oxygen}/n_{iridium}$) was lowered by a factor of 48.8 at current densities of 10 mA·cm^{-2} (Table S1)[30] (see Figures S2−S7 and Tables S2−S4 for samples preparation and remaining characterization).

2.1. The SI-SECM Titration Sequence. The experimental setup and redox titration sequence of SI-SECM are shown in Scheme 1. The SI-SECM experiments utilize a tip electrode (top) that enables *in operando* interrogation of the substrate electrode (bottom, where the catalytic reaction of interest occurs) through an electrochemical reaction of a redox mediator orthogonal to the catalytic reaction of interest. In this work, a gold ultramicroelectrode (UME) was used as the tip electrode, iron(III) sulfate ($Fe_2(SO_4)_3$) as the redox mediator, and iridium oxides electrodeposited on platinum UME as the substrate electrode. As shown in Scheme 1a, two size-matched (25 μm in diameter) UMEs were aligned and approached to a distance of *ca.* 2−3 μm, enabling the tip to directly detect any perturbations at the substrate (see Figure S8 for details on tip−substrate alignment and collection efficiency determination).

The titration sequence proceeded as follows: first, the surface-active species on the catalyst film were in the protonated resting state (O^{2-}-H^+) at the open circuit potential (OCP) (Scheme 1a). The chemical identity of the active species was ascertained as redox states localized on lattice oxygen atoms by hard and soft X-ray absorption analyses (*vide infra*), corroborated by a large volume of literature precedents.[12−15,17,19,31−33] Upon application of anodic potentials to the substrate electrode (anodic pulse time; t_{pulse} = 20 s), the active oxygen species were deprotonated as their electrophilic hole character increased (O^{1-} and H^+) (Scheme 1b). Deprotonated oxygens were then quantified via iron-mediated redox reaction feedback at the tip (Scheme 1c). Upon switching the substrate electrode to open circuit, the cathodic potential applied to the tip electrode reduced the iron-based redox mediator ($Fe^{3+/2+}$ E° = 0.65 V vs RHE), thereby initiating spontaneous redox reactions with surface O^{1-} (positive feedback) ($Fe^{2+} + O^{1-} \rightarrow Fe^{3+} + O^{2-}$) (Scheme 1c (i)) (see Figures S9 and S10 for further details on the redox mediator). The tip titration current relaxed to the negative feedback level as the surface-bound and electrolyte-accessible active oxygen species were fully consumed (Figure S11). The collected amperograms were then integrated over time to yield the number of titrand species per electrochemically active surface area (ECSA) (Figures S12 and S13). Repeating the quantification sequence from the OCP (0.75 V vs RHE) to the

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

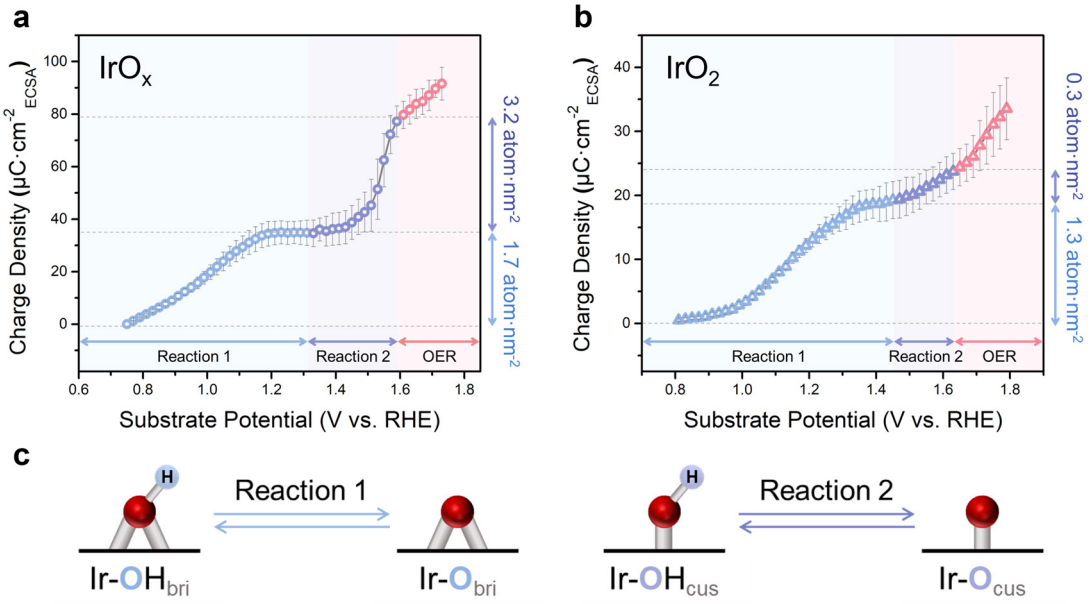


Figure 1. Redox titration curves (charge density as a function of substrate potential) for (a) amorphous IrO_x and (b) crystalline IrO_2. Active site quantification was normalized by the electrochemically active surface area (ECSA). The right *Y*-axis represents the atom density of each oxygen species (sky-blue for O_{bri} and purple for O_{cus}). (c) Schematic representation of substrate reactions within different potential ranges: the left panel illustrates reaction 1 in the sky-blue potential range, while the right panel shows reaction 2 in the purple potential range.

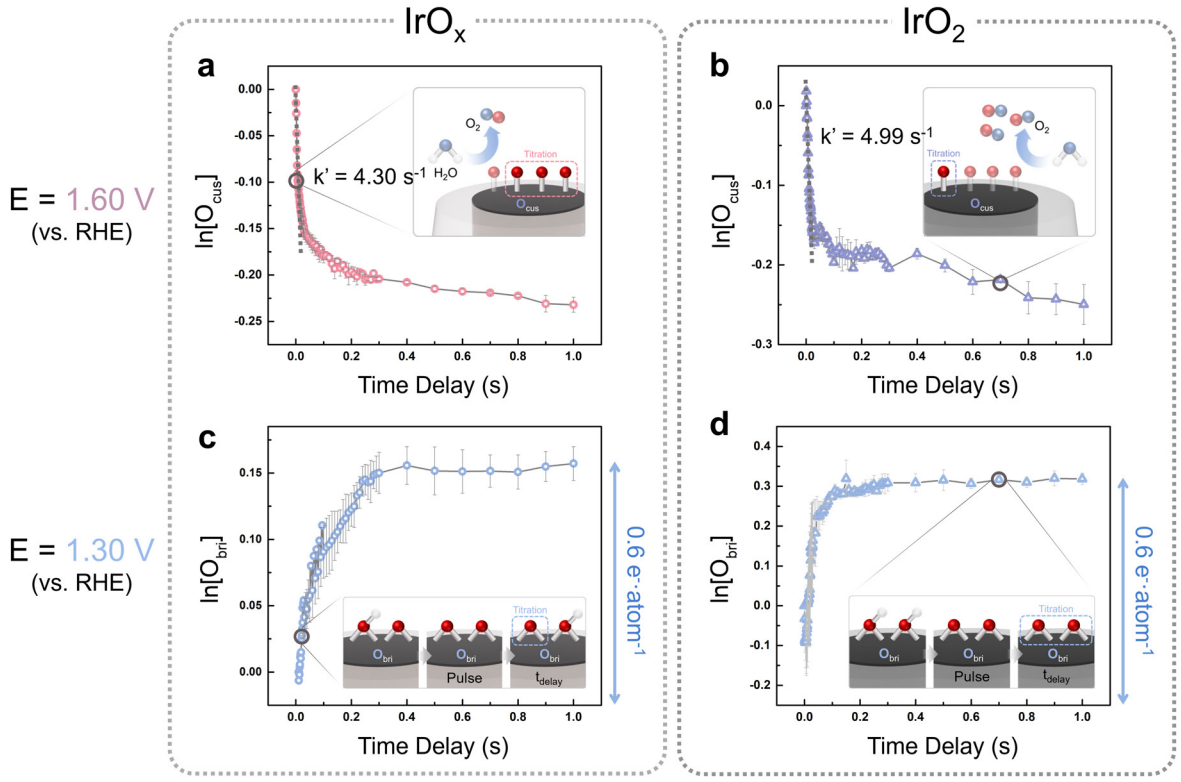


Figure 2. Time-dependent titration curves (natural logarithm of the charge density as a function of time delay) for (a, c) amorphous IrO_x and (b, d) crystalline IrO_2, indicating pseudo-first-order kinetics model (the slope of each graph) at specific reaction potentials. (a) and (b) exhibit results at 1.60 V corresponding to O_{cus} deprotonation, while (c) and (d) show results at 1.30 V where O_{bri} deprotonation occurs. The right *Y*-axis in (c) and (d) represents the charge storage capacity per O_{bri} atom.

OER (1.85 V vs RHE) resulted in the redox titration curve, reflecting the variation in surface charge density as a function of the substrate potential, E_{subs} (Figure 1).

Introducing the time delay (t_{delay}) between titrand (O^{1-}) formation and titrant (Fe^{2+}) generation yielded important kinetic parameters of the OER active species. In the absence of

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

the titrant, the titration current at the tip decays as a function of t_{delay}, due to the active site consumption via the OER (Scheme 1c (ii)) or protonation (Scheme 1c (iii)), which are chemical reactions in competition with the redox titration reaction. Time-dependent titration at specific reaction potentials yielded the rate constants for the chemical reactions by precise control of the time scales of the faster redox reaction (Figure 2). We employed an external rapid switching device that enabled the control of t_{delay} at 1 ms scale, which is a fast enough time resolution for the tracking of the reactive intermediates of OER with typical rates of several turnovers per second (Figure S14).[34]

2.2. The Redox Titration Curve of Iridium Oxides in Acidic OER. The redox titration curves provide both qualitative and quantitative insights into the electrocatalytic reactions of interest (Figure 1). A common trend was observed for IrO_x and IrO_2: both plots exhibited increases in the number of detected redox active species per unit area (ECSA) with increasing anodic substrate potential (Figure 1a,b; see also Figure S11 for the amperograms of the tip and substrate). Based on the presence of plateaus or changes in slopes, each curve was segmented into three distinct redox reaction regions color-coded in sky-blue, purple, and red, respectively. Concurrent to the findings in the literature,[12,13,17,19,31−33] we confirmed that the sky-blue potential range (0.75−1.30 V for IrO_x and 0.85−1.45 V for IrO_2) indicated oxidative deprotonation of bridging lattice oxygen (O_{bri}) (Figure 1c, reaction 1 in the left panel) and the purple potential range (1.30−1.60 V for IrO_x and 1.45−1.63 V for IrO_2) indicated oxidative deprotonation of coordinatively unsaturated site lattice oxygen (O_{cus}) (Figure 1c, reaction 2 in the right panel). The electron-deficient oxygen species designated as O_{cus} in this study includes both deprotonated water molecules absorbed on iridium coordinatively unsaturated sites and the lattice μ_1-O formed via lattice defect (see Supporting Information (page S21) for further details). The constant slope in the red region (above 1.60 V for IrO_x and 1.63 V for IrO_2) reflected the predominant occurrence of the OER.

Typically in redox titration curves, a plateau signals a surface redox equivalence.[26] A plateau in charge accumulation with increasing substrate potentials is indicative of saturation of oxidizable (such as that shown in this work) surface species. The sky-blue region for both IrO_x and IrO_2 revealed the first oxidation equivalence of the surfaces, which we designated as oxidative deprotonation of O_{bri}, and accumulation of O^{1-}, in accordance with literature precedents. The chemical identity of the lattice oxygen species is further corroborated by other experimental means such as hard and soft X-ray spectroscopies and redox kinetic isotope studies (*vide infra*). The subsequent change in slope and climb in the charge density as functions of substrate potential signaled the access of the O^{1-} species in the O_{cus} lattice sites (the purple region) for the two respective catalysts. Through the two plateaus corresponding to the two redox states, we calculated the atom densities for each active oxygen species: 1.7 ± 0.2 atom·nm^{-2} for O_{bri} and 3.2 ± 0.2 atom·nm^{-2} for O_{cus} in IrO_x (Figure 1a, right axis) and 1.3 ± 0.1 atom·nm^{-2} for O_{bri} and 0.3 ± 0.1 atom·nm^{-2} for O_{cus} in IrO_2 (Figure 1b. right axis) (see Figures S15 and S16, Tables S5 and S6, and Supporting Information (page S24) for detailed calculations and statistical analysis).

The findings collectively led to the following analysis: the calculated atom density for each oxygen species revealed that IrO_x exhibits comparable (a factor of 1.3) O_{bri} density and 10.7-fold higher O_{cus} density compared to those of IrO_2. This indicated that the structural differences are more pronounced in expression for O_{cus} than for that for O_{bri}. Moreover, O_{cus} undergoes deprotonation at more anodic potentials than that associated with O_{bri}, only slightly shy of that for OER. The significantly higher O_{cus} density in IrO_x and its proximity in the potential space to that of OER strongly suggest that O_{cus} is the OER active oxygen species.[35−37] Owing to the high reactivity of the OER active O_{cus} (see also discussions in the subsequent sections), direct spectroscopic characterization of the species has been scarce.[10,19] Our experimental findings and data analysis exhibited similar trends to that reported by Stephens and co-workers,[19] that comparable amounts of O_{cus} and O_{bri} are present in IrO_x and IrO_2 and that more active sites are expressed in IrO_x due to the structural ease of proton permeation. Quantification and discrimination in the potential space of the lattice oxygen species were feasible owing to the unique capabilities of the SI-SECM method, enabling selective probing of the electrode−electrolyte interface, which successfully overcomes the limitations of conventional surface science techniques in the investigation of complex dynamics at catalyst−solution interfaces.[21,24−29]

2.3. Time-Dependent Titrations: The OER Rate Constant Measurement. We employed time-dependent titration to investigate the reactivity of each electron-deficient oxygen species at specific reaction potentials (Figure 2). The pseudo-first-order reaction rate constants were determined by varying the t_{delay} between electron-deficient oxygen generation (O^{1-}) at the substrate and the titrant formation (Fe^{2+}) at the tip. The rate constant is determined by the slope of the graph plotting the natural logarithm of the integrated charge density (i.e., remaining O^{1-} during t_{delay}) versus t_{delay} (see Supporting Information (page S25) for details). The t_{delay} was controlled between 0 and 1 s, with intervals of 1, 5, 10, and 100 ms, using an external rapid switching device (Figure S14).[34]

As shown in Figure 2a (IrO_x) and Figure 2b (IrO_2), the amount of titrated charge decreased with increasing t_{delay} at 1.60 V, near the saturation potential of the O_{cus} in the redox titration curve (Figure 1). The decay in remaining O_{cus} indicates active sites consumption via OER during t_{delay} in the absence of other oxidizable species: oxygen can be evolved primarily via nucleophilic attack of solution H_2O.[38,39] Linear decay of the natural logarithm of the concentration of surface-active species has been typically observed for electrocatalytic reactions,[25,26] and the similar results observed here confirmed that O_{cus} is OER-active.

The observed slope of each graph represents the O_{cus} OER rate constant (k'), 4.30 ± 0.82 s^{-1} for IrO_x (Figure 2a) and 4.99 ± 0.67 s^{-1} for IrO_2 (Figure 2b; see Figures S17 and S18 and Tables S7 and S8 for statistical analysis), and indicated that an active lattice oxygen site turned over the OER four to five times per second. The similar k' values indicate that the intrinsic activity of O_{cus} is comparable across both catalyst structures at a given potential. As shown in Figure 1 and Figure S1, IrO_x reached a current density of 10 mA·cm^{-2} at 1.58 V coinciding with O_{cus} saturation, while IrO_2 exhibited a potential gap between O_{cus} charge saturation (1.63 V) and that related with 10 mA·cm^{-2} current density (1.73 V). Despite comparable O_{cus} OER activity per atom at 1.60 V, *ca.* 150 mV overpotential difference between the two catalysts is attributed to the O_{cus} density quantified by the redox titration curve (Figure 1; see the Supporting Information (page S24) for further explanation and calculations).

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

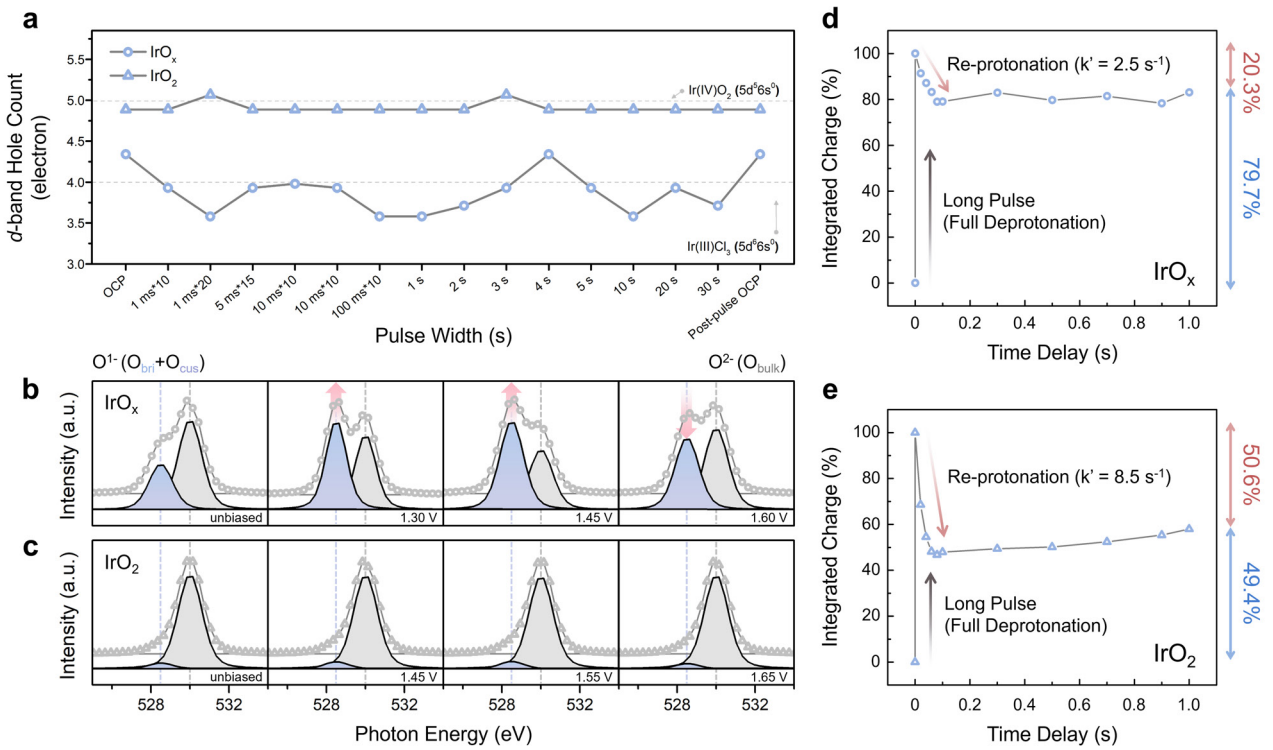


Figure 3. (a) d-band hole count variations in response to continuous pulse (IrO_x circle, IrO_2 triangle). *Operando* Ir L_3-edge spectra were recorded under the potential of reaction 1 (Figure 1c, left panel) in 0.5 M H_2SO_4. O K-edge spectra of (b) IrO_x and (c) IrO_2, measured using total electron yield (TEY) detection mode. Peak intensity variations under increasing anodic potentials are shown from left to right with the applied potential labeled in each graph. Gray-shaded peak at 530 eV denotes O^{2-} (O_{bulk}), while the sky-blue-purple colored peak at 528.5 eV corresponds to O^{1-} (O_{bri} and O_{cus}).[14,15] Note that some O^{1-} concentration was observed in the unbiased samples due to the preconditioning of the electrode to potentials where O_{bri} occurs. Time-dependent titration curves representing reprotonation kinetics following long-pulse (1.30 V) treatment for (d) IrO_x and (e) IrO_2.

An interesting anomaly (Figure 2c,d) was observed when time-dependent titration experiments were performed at 1.30 V (O_{bri} saturation potential in Figure 1): titrated charge density increased linearly with t_{delay} until reaching saturation, unlike the behavior observed at 1.60 V with O_{cus} (see Figure S19). The data signaled that O_{bri} was not consumed by OER after deprotonation (no spontaneous reaction with water); therefore, O_{bri} is not an OER-active site (see Table S9 for further explanation).[16,19,35] Typically with chemically inactive redox states in SI-SECM, no time dependence in the concentration is observed;[26] that is, the natural logarithm of the concentration remains constant at all t_{delay}, due to the preservation of the surface redox state (Figure S20). However, in the present case, we observed an increase in the redox concentration with continued anodic stimuli on the electrode. Therefore, we attributed this to the accumulation of the metastable O^{1-} lattice oxygen species (for spectral evidence refer to the subsequent section) as a function of the number of substrate potential pulse. To the best of our knowledge, this investigation is a unique explicit experimental documentation of the much-proposed charge storage mechanism in iridium oxide catalyzed acidic OER.[10,12,13,16,36,40−43] Moreover, we carried out quantitative analysis of the relationship between the stored charge and the applied potential, providing comprehensive insight into the charge storage mechanism and the implications of the stored charge on the OER kinetics and material stability (see sections below). Similar charge accumulation profiles were obtained regardless of the order of applied t_{delay} (whether starting from short t_{delay} and increasing or vice versa) (Figure S21), revealing that the charge accumulated was strictly a function of the number of applied potential pulses (also time of oxidative potential application; *vide infra*).

2.4. Analysis of the Electronic Structure of the Electron-Deficient Oxygen Species. Regarding reaction 1 centered on the O_{bri} (left panel in Figure 1c), X-ray absorption near edge structure (XANES) spectroscopy was employed to shed light on the chemical identity of the redox moiety related to the charge storage mechanism. The oxidation state of Ir was analyzed based on the white line position in the Ir L_3-edge XANES spectra. Ir(0) black ($5d^76s^2$), Ir(III)Cl_3 ($5d^66s^0$), and Ir(IV)O_2 ($5d^56s^0$), which have one electron difference in the d-band, were utilized as the reference materials (Figure S22). The Ir L_3-edge spectra of the catalyst electrodes under continuous pulses (1.35 V for IrO_x, 1.40 V for IrO_2) resulted in negligible variations in the d-band hole: IrO_x at 3.9 ± 0.3 and IrO_2 at 4.9 ± 0.1 (Figure 3a). In addition, electrodes thinned to near-nanoscale thickness, designed to increase the proportion of active-surface to core ratio, exhibited similar behavior (Table S10 and Figures S23 and S24). The apparent iridium oxidation state of 3+ in the IrO_x film results from the averaged probing, including the metallic iridium in the underlying bulk near the platinum support. Linear combination fit of the IrO_x XANES spectra (Figure S25) suggested the predominance of 4+ oxidation states in the oxide layer (see also angle-resolved XPS (AR-XPS) spectra in Tables S11 and

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

S12). More importantly, no significant oxidation state change in iridium suggested a weak link between the iridium metal centers and the time-dependent charge accumulation in Figure 2c,d. Also, for potential ranges related to reaction 2 and OER (Figure S26), little change in the oxidation state of iridium was observed for both amorphous IrO_x and crystalline IrO_2, similar to a handful of research precedents;[12−14,31,44] which led us to assume that the OER active sites are centered around electron-deficient oxygen species rather than Ir atoms. A number of literature precedents suggested that the strong overlap in the Ir 5d and O 2p orbitals allows oxidation of the lattice oxygen species to O^{1-} upon application of an anodic bias, rather than accessing higher oxidation states in iridium.[16,40,44−46] A number of research accounts reported detection of the Ir(IV) to Ir(V) transition during the water oxidation cycle;[29,31,47−50] however, presumably due to the transience of the chemical species, it was not explicitly observed in our experimental time and film thickness scales. Fitting of the AR-XPS spectra (Ir 4f; Tables S11 and S12) suggested partial involvement of the iridium(V) at the OER potentials (with more pronounced contribution at the surface—larger tilt angle), which led us to believe that Ir(V) states are involved in acidic OER; however, our experiments suggest that surface charge storage primarily occurs via the O^{1-} species (see Supporting Information (pages S33−S34) for further discussion).

The pre-edge regions of the O K-edge spectra in Figure 3b (IrO_x) and Figure 3c (IrO_2) further provided information regarding the chemical identity of the redox species in the catalytic cycle, in good alignment with the electrochemical findings from the SI-SECM data (see Figure S27 for full-range spectra). Figure 3b,c represents the O K-edge spectra of the surface states at the denoted applied potentials. In contrast to the Ir L_3-edge, changes in signal intensity were observed in response to anodic stimuli: application of 1.30 V (and 1.45 V) to IrO_x increased the intensity of the O^{1-} peak (O_{bri} + O_{cus}, sky-blue-purple peak in Figure 3b). This trend strongly suggests that the redox species responsible for charge storage in IrO_x is the electron-deficient lattice oxygens;[12,36,51] oxidizable species in the system were limited to iridium, oxygen, and solution H_2O, and we have shown (*vide supra*) that iridium and water were not oxidized at the given potential from hard X-ray absorption and SI-SECM experiments. The O K-edge spectral analysis further confirmed the chemical distinction between O_{cus} and O_{bri}. At 1.60 V the O^{1-} peak intensity decreases due to O_{cus} consumption via OER, in contrast to peak increase at 1.30 V through deprotonated O_{bri} accumulation, strongly supporting that O_{cus} is OER active and O_{bri} is OER inactive, in accordance to our SI-SECM findings (Table S9).

In the proposed charge storage mechanism (of iridium oxide acidic OER), the strong hybridization between iridium and oxygen orbitals results in positive charge sharing between the lattice cations and anions,[40,45] lowering the energy level of the oxygen p-band relative to the metal d-band: therefore, the lattice oxygens become the OER active sites via the electron-deficient states (O^{1-}).[9] First proposed by Grimaud and co-workers,[46] numerous studies have supported the mechanism;[10,12,13,36,41,42] however, experimental validation has been scarce. Our time-dependent titration clearly demonstrated the process of positive charge accumulation on the electron-deficient lattice O_{bri} (Figure 2). As the catalyst electrode was continuously exposed to anodic pulses of reaction 1 (Figure 1c), an increasing amount of O_{bri} (confirmed by the O K-edge in Figure 3b) remained deprotonated during t_{delay} (maximum 1 s), resulting in greater feedback at the tip (Figure 2c,d inset).

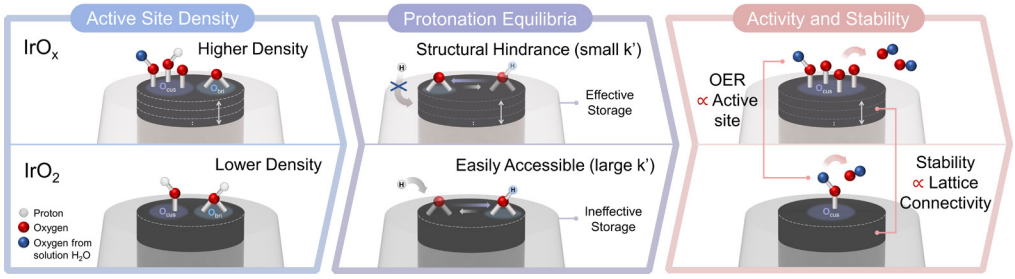
To further verify that charge accumulation on the surface of the elongated O_{bri} corresponds to the proton transfer in reaction 1 (Figure 1c), time-dependent titration experiments were repeated with D_2O as the proton source. As shown in Figures S28 and S29 and Tables S13 and S14, the kinetic plot revealed similar overall shape to those with H_2O at both 1.30 and 1.60 V but exhibited a normal kinetic isotope effect (KIE) value of 2.0 ± 0.5.[52,53] The slower charge accumulation in D_2O indicated that the observed oxidative transformation involves proton transfer and strongly suggests (with limited combinatory possibility invoking only iridium, water, and lattice oxygens) the deprotonation of O_{bri} (at 1.30 V). To the best of our knowledge, this is the first empirical verification and chemical identification of the charge storage process in iridium oxide catalyzed acidic OER. To rule out the catalyst detachment scenario, the effect of substrate (platinum) and dissolved O_2, identical experiments were performed on a bare Pt UME substrate and Ar-purging condition respectively: the results further confirm that the observed effects were specific to IrO_x and IrO_2 (Figures S20 and S30).

2.5. Measurement of the Lattice Oxygen Deprotonation Rate Constants and the Implications of the Protonation Equilibria on the OER Activity−Stability Trend. Numerous studies suggested the correlation between accumulated positive charge on iridium oxide surfaces and the subsequent OER catalysis.[10,16] As experimentally verified by our time-dependent titration, O_{bri} deprotonation (and slow reprotonation; *vide infra*) leads to oxidative charge accumulation.[15,16] Therefore, electrochemical investigation of the proton transfer kinetics on O_{bri} and O_{cus} is of interest, and more importantly, complete elucidation of the implications of the protonation equilibria on acidic OER catalysis is desired.

Quantitative data from Figure 2c,d reveal clear differences in the proton transfer process between the two catalysts. Dividing the saturated charge by the number of participating atoms yielded the number of redox equivalence accumulated per lattice oxygen atom: 0.6 ± 0.2 $e^-\cdot atom^{-1}$ for IrO_x, and 0.6 ± 0.1 $e^-\cdot atom^{-1}$ for IrO_2 (Figure 2c,d, right axis; see Figures S31 and S32 and Tables S15−S20 for calculation details). The per atom charge storage capacity of O_{bri} is comparable in both catalysts; however, the key difference lies in the kinetic accessibility of the deprotonated state. The IrO_x electrode (Figure 2c) required 55 anodic pulses (total time of 273 ± 30 s) to reach redox saturated surface (Table S17), whereas IrO_2 (Figure 2d) exhibited a significantly faster deprotonation rate by requiring only 27 anodic pulses (total time of 136 ± 26 s) to reach redox saturation (Table S20). To further delve into the protonation equilibria of the electron-deficient lattice oxygens, we applied long pulses matching the time required for each catalyst to reach its O_{bri} charge saturation (275 s for IrO_x and 140 s for IrO_2) and then measured the reprotonation (backward reaction in Figure 1c reaction 1) rate constants (by variation of t_{delay}; see Supporting Information (page S25) for details). As shown in Figure 3d,e, IrO_2 exhibited a much faster reprotonation rate: 2.5 ± 1.0 s^{-1} for IrO_x and 8.5 ± 0.9 s^{-1} for IrO_2 (see Table S21 for statistical analysis). This indicated that while IrO_x and IrO_2 accumulate a similar amount of charge per atom, the proton transfer rate constant (k') differed significantly: IrO_2 exhibited a much faster reprotonation rate compared to that for IrO_x.

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

Scheme 2. Representation Illustrating Structure-Dependent Penetration Depth of Electron-Deficient Oxygen Species, Governing Charge Storage Kinetics, OER Activity, and Catalyst Durability



The structure dependent proton transfer kinetics can be explained by the O K-edge spectra. As shown in Figure 3b,c, the relative intensities of the O^{2-} (gray-colored peak, bulk lattice) and the O^{1-} (sky blue-purple colored peak, O_{bri} and O_{cus}) reflect the subsurface penetration of the O^{1-} species. In the case of IrO_x (Figure 3b), the intensity of O_{bri} and O_{cus} peaks varied with applied potentials, whereas no significant change (presumably due to the small population) was observed for IrO_2. This observation suggests that while O^{1-} extended to the subsurface regions of IrO_x due to electrolyte permeability, those in IrO_2 were limited to the topmost layer (Table S4 and Figure S33).[10,15,19,54] Consequently, the structure-dependent penetration depth of the elongated state of O^{1-} controlled the proton transfer kinetics and the equilibrium between the protonated and the deprotonated states. Beyond kinetic rate constants for the reprotonation of the electron-deficient lattice oxygen species, the kinematic experiments in Figure 3d,e shed light on the protonation/deprotonation equilibria. After achieving fully deprotonated O^{1-} state at 1.30 V, the system was allowed to reach equilibrium by reaction with water at OCP. At equilibrium, IrO_x retained 79.7% of its deprotonated O_{bri}, with 20.3% undergoing reprotonation, while IrO_2 retained only 49.4% of O_{bri}, with 50.6% reprotonated (Figure 3d,e). The protonation equilibrium lied heavily in favor of the deprotonated state for IrO_x, whereas that for IrO_2 favored the protonated state.

Scheme 2 summarizes the experimental findings from redox titration, electrochemical kinetics, and X-ray structural measurements in establishing the comprehensive picture of the structure-dependent active site penetration depth (lattice oxygen mediated charge storage) and its implication on the catalytic activity and stability of the materials. Amorphous IrO_x contained a higher density of active species ($O_{bri} + O_{cus}$, 4.9 ± 0.2 atom·nm^{-2}) stemming from the access of subsurface, resulting in slower deprotonation and reprotonation (k' = 2.5 ± 1.0 s^{-1}) and equilibrium favoring the oxidative charge (79.7% of O_{bri}). In contrast, crystalline IrO_2 exhibited a lower density of active species (1.6 ± 0.1 atom·nm^{-2}) confined to the catalyst surface. As a result, IrO_2 undergoes rapid deprotonation and reprotonation (k' = 8.5 ± 0.9 s^{-1}), which limited effective oxidative charge storage at equilibrium (maintaining 49.4% of O_{bri}).

Numerous studies have suggested that as oxidative charge accumulates on electron-deficient oxygen species, the activation free energy for OER decreases;[10,12,13,16,36,41,42] however, our electrochemical kinetic study revealed that the effective charge accumulation on O_{bri} does not lower the activation energy of OER at O_{cus}. From the redox titration curves in Figure 1, the OER potential ranges are comparable for the two structures (at 1.60 and 1.63 V, respectively, for IrO_x and IrO_2), despite the much larger apparent overpotential gap of *ca.* 150 mV found in conventional linear sweep voltammetry (Figure S1), which resulted from the amplification effect fueled by the 10-fold difference in the active site (O_{cus}) densities (Figure S34). Also from the slopes of Figure 2a,b, both IrO_x and IrO_2 exhibit comparable OER rate constants per atom, indicating identical intrinsic activity for OER at a given potential regardless of the crystal structure.[19] Therefore, the higher activity of IrO_x is primarily attributed to its higher O_{cus} density quantified by a redox titration curve (Figure 1, due to the subsurface propagation of O^{1-}); instead, the effective charge accumulation on O_{bri} can reduce the activation energy for O_{cus} deprotonation. Although other contributing factors, such as the surface adsorbate−adsorbate interactions,[19] need also be considered for the comprehensive understanding of the acidic OER mechanism, the difference in the active site density is a major contributor in the structure-distinct apparent reactivities in IrO_x and IrO_2 (Figure S34).

Meanwhile, the electron-deficient oxygen species also serve as key intermediates in iridium dissolution,[9,20] contributing to catalyst degradation; therefore, the active sites in IrO_x that effectively retain their oxidative deprotonation state may be more prone to dissolution. Moreover, the correlation between *S*-number (defined as the ratio between the amount of evolved oxygen and dissolved iridium, $n_{oxygen}/n_{iridium}$; see Table S1 for calculation details) measured via ICP-MS and O K-edge spectra (Figure 3) suggests that the structural robustness of the lattice supporting the active sites plays a crucial role in catalyst durability. In the case of IrO_x, most active sites are located within the electrolyte-permeable subsurface with low lattice connectivity (μ_1-O, μ_2-O; represented by the higher $O_{bri} + O_{cus}$ density), whereas IrO_2 exhibits stronger lattice connectivity (μ_3-O) directly beneath the active sites.[10] As a result, under similar O_{cus} activity, IrO_2 is more likely to retain its active sites due to robust lattice support, leading to a 151.2-fold increase in catalyst durability (Table S1). Thus, the penetration depth of electron-deficient oxygen species into the oxide layer determines proton transfer to O^{1-}, ultimately leading to the structure dependence of activity and stability of the acidic OER catalyst.

3. CONCLUSIONS

In this work, we employed the SI-SECM technique to quantify the transient reactivity of active oxygen species in amorphous IrO_x and crystalline IrO_2, correlating with structural analysis to comprehensively elucidate the factors influencing the catalytic activity and stability. The titration sequence accurately quantified significant differences from the O^{1-} penetration depth across different structures. While the intrinsic OER activity of O_{cus} (*ca.* 4.6 s^{-1}) and charge storage capacity of O_{bri}

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

(*ca.* 0.6 e^{-}·$atom^{-1}$) were comparable in both catalysts, redox titration curves revealed that IrO_x exhibits 10.7-fold higher density of the OER-active O_{cus}, compared to IrO_2. Moreover, the electrolyte permeability facilitated charge accumulation of IrO_x, exhibiting 3-fold more gradual reprotonation slope. However, efficient charge storage had a marginal effect on lowering the OER barrier itself, and the catalytic activity is primarily determined by the number of active sites. Furthermore, catalyst durability is linked to the electrolyte permeability of the material. The robust connectivity of the substrate support stabilizes the active sites, thereby enhancing the catalyst's durability. The discrete quantitative comparison using the SI-SECM technique elucidates the structure-dependent activity−stability trend, offering valuable insights for catalyst design aimed at efficient green hydrogen production. This technique and its insights can be applied to a broader range of energy conversion reactions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.5c16721.

Experimental details, materials and method, additional experimental and characterization data (LSV, CV, ICP-MS, SEM/EDS, XPS, XRD, BET, XAS), device setup, further discussions, and statistical analysis (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Sang-Il Choi − *Department of Chemistry and Green-Nano Materials Research Center, Kyungpook National University, Daegu 41566, Republic of Korea;* orcid.org/0000-0002-8280-3100; Email: sichoi@knu.ac.kr

Hangil Lee − *Department of Chemistry, Sookmyung Women's University, Seoul 04310, Republic of Korea;* orcid.org/0000-0002-0657-5660; Email: easyscan@sookmyung.ac.kr

Hyun S. Ahn − *Department of Chemistry, Yonsei University, Seoul 03722, Republic of Korea;* orcid.org/0000-0002-6014-0916; Email: ahnhs@yonsei.ac.kr

Authors

Yeonsu Kim − *Department of Chemistry, Yonsei University, Seoul 03722, Republic of Korea*

Jeonghyeon Kim − *Department of Chemistry and Green-Nano Materials Research Center, Kyungpook National University, Daegu 41566, Republic of Korea*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.5c16721

Author Contributions

All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Basic Science Research Program through National Research Foundation (NRF) of Korea (RS-2024-00337549 for H.S.A., RS-2024-00346153 for H.L. and S.-I.C., and RS-2025-02215028 for S.-I.C.), Global—Learning & Academic Research Institution for Master's·PhD students, and Postdocs (LAMP) Program of the NRF grant funded by the Ministry of Education (RS-2024-00442483), and by the Technology Innovation Program (20022479) funded by the Ministry of Trade Industry & Energy (MOTIE, Korea). This work was also supported by the NRF grant funded by the Korea government (MSIT) (RS-2025-14383258 for H.S.A., and RS-2025-21072982 for H.L.).

■ REFERENCES

(1) Nikolaidis, P.; Poullikkas, A. A comparative overview of hydrogen production processes. *Renew. Sustain. Energy Rev.* **2017**, *67*, 597−611.

(2) Carmo, M.; Fritz, D. L.; Mergel, J.; Stolten, D. A comprehensive review on PEM water electrolysis. *Int. J. Hydrogen Energy* **2013**, *38*, 4901−4934.

(3) Kibsgaard, J.; Chorkendorff, I. Considerations for the scaling-up of water splitting catalysts. *Nat. Energy* **2019**, *4*, 430−433.

(4) Chen, Z.; Duan, X.; Wei, W.; Wang, S.; Ni, B.-J. Iridium-based nanomaterials for electrochemical water splitting. *Nano Energy* **2020**, *78*, 105270.

(5) Minke, C.; Suermann, M.; Bensmann, B.; Hanke-Rauschenbach, R. Is iridium demand a potential bottleneck in the realization of large-scale PEM water electrolysis? *Int. J. Hydrogen Energy* **2021**, *46*, 23581−23590.

(6) Cherevko, S.; Zeradjanin, A. R.; Topalov, A. A.; Kulyk, N.; Katsounaros, I.; Mayrhofer, K. J. Dissolution of noble metals during oxygen evolution in acidic media. *ChemCatChem* **2014**, *6*, 2219−2223.

(7) Jovanovic, P.; Hodnik, N.; Ruiz-Zepeda, F.; Arcon, I.; Jozinovic, B.; Zorko, M.; Bele, M.; Sala, M.; Selih, V. S.; Hocevar, S.; Gaberscek, M. Electrochemical dissolution of iridium and iridium oxide particles in acidic media: transmission electron microscopy, electrochemical flow cell coupled to inductively coupled plasma mass spectrometry, and X-ray absorption spectroscopy study. *J. Am. Chem. Soc.* **2017**, *139* (36), 12837−12846.

(8) Wang, H.; Zhang, K. H. L.; Hofmann, J. P.; de la Pena O'Shea, V. A.; Oropeza, F. E. The electronic structure of transition metal oxides for oxygen evolution reaction. *J. Mater. Chem. A* **2021**, *9* (35), 19465−19488.

(9) Lončar, A.; Escalera-López, D.; Cherevko, S.; Hodnik, N. Inter-relationships between oxygen evolution and iridium dissolution mechanisms. *Angew. Chem., Int. Ed.* **2022**, *61* (14), No. e202114437.

(10) Mom, R. V.; Falling, L. J.; Kasian, O.; Algara-Siller, G.; Teschner, D.; Crabtree, R. H.; Knop-Gericke, A.; Mayrhofer, K. J.; Velasco-Vélez, J.-J. s.; Jones, T. E. Operando structure-activity-stability relationship of iridium oxides during the oxygen evolution reaction. *ACS Catal.* **2022**, *12* (9), 5174−5184.

(11) Danilovic, N.; Subbaraman, R.; Chang, K.-C.; Chang, S. H.; Kang, Y. J.; Snyder, J.; Paulikas, A. P.; Strmcnik, D.; Kim, Y.-T.; Myers, D.; et al. Activity-stability trends for the oxygen evolution reaction on monometallic oxides in acidic environments. *J. Phys. Chem. Lett.* **2014**, *5* (14), 2474−2478.

(12) Pfeifer, V.; Jones, T. E.; Velasco Vélez, J. J.; Arrigo, R.; Piccinin, S.; Hävecker, M.; Knop-Gericke, A.; Schlögl, R. In situ observation of reactive oxygen species forming on oxygen-evolving iridium surfaces. *Chem. Sci.* **2017**, *8* (3), 2143−2149.

(13) Pfeifer, V.; Jones, T. E.; Wrabetz, S.; Massue, C.; Velasco Velez, J. J.; Arrigo, R.; Scherzer, M.; Piccinin, S.; Havecker, M.; Knop-Gericke, A.; Schlogl, R. Reactive oxygen species in iridium-based OER catalysts. *Chem. Sci.* **2016**, *7* (11), 6791−6795.

(14) Pfeifer, V.; Jones, T. E.; Velasco Vélez, J. J.; Massué, C.; Arrigo, R.; Teschner, D.; Girgsdies, F.; Scherzer, M.; Greiner, M. T.; Allan, J.; et al. The electronic structure of iridium and its oxides. *Surf. Interface Anal.* **2016**, *48* (5), 261−273.

(15) Saveleva, V. A.; Wang, L.; Teschner, D.; Jones, T.; Gago, A. S.; Friedrich, K. A.; Zafeiratos, S.; Schlögl, R.; Savinova, E. R. Operando evidence for a universal oxygen evolution mechanism on thermal and electrochemical iridium oxides. *J. Phys. Chem. Lett.* **2018**, *9* (11), 3154−3160.

(16) Nong, H. N.; Falling, L. J.; Bergmann, A.; Klingenhof, M.; Tran, H. P.; Spöri, C.; Mom, R.; Timoshenko, J.; Zichittella, G.; Knop-Gericke, A.; et al. Key role of chemistry versus bias in electrocatalytic oxygen evolution. *Nature* **2020**, *587* (7834), 408−413.
(17) Kwon, S.; Stoerzinger, K. A.; Rao, R.; Qiao, L.; Goddard, W. A., III; Shao-Horn, Y. Facet-dependent oxygen evolution reaction activity of IrO2 from quantum mechanics and experiments. *J. Am. Chem. Soc.* **2024**, *146* (17), 11719−11725.
(18) BalaKrishnan, A.; Blanc, N.; Hagemann, U.; Gemagami, P.; Wonner, K.; Tschulik, K.; Li, T. Direct detection of surface species formed on iridium electrocatalysts during the oxygen evolution reaction. *Angew. Chem., Int. Ed.* **2021**, *60* (39), 21396−21403.
(19) Liang, C.; Rao, R. R.; Svane, K. L.; Hadden, J. H.; Moss, B.; Scott, S. B.; Sachs, M.; Murawski, J.; Frandsen, A. M.; Riley, D. J.; et al. Unravelling the effects of active site density and energetics on the water oxidation activity of iridium oxides. *Nat. Catal.* **2024**, *7* (7), 763−775.
(20) Kasian, O.; Grote, J. P.; Geiger, S.; Cherevko, S.; Mayrhofer, K. J. The common intermediates of oxygen evolution and dissolution reactions during water electrolysis on iridium. *Angew. Chem., Int. Ed.* **2018**, *57* (9), 2488−2491.
(21) Li, T.; Kasian, O.; Cherevko, S.; Zhang, S.; Geiger, S.; Scheu, C.; Felfer, P.; Raabe, D.; Gault, B.; Mayrhofer, K. J. J. Atomic-scale insights into surface species of electrocatalysts in three dimensions. *Nat. Catal.* **2018**, *1* (4), 300−305.
(22) Bard, A. J.; Mirkin, M. V. *Scanning Electrochemical Microscopy*, 3rd ed.; CRC Press, 2022.
(23) Amemiya, S.; Bard, A. J.; Fan, F.-R. F.; Mirkin, M. V.; Unwin, P. R. Scanning electrochemical microscopy. *Annu. Rev. Anal. Chem.* **2008**, *1* (1), 95−131.
(24) Rodríguez-López, J.; Alpuche-Avilés, M. A.; Bard, A. J. Interrogation of surfaces for the quantification of adsorbed species on electrodes: oxygen on gold and platinum in neutral media. *J. Am. Chem. Soc.* **2008**, *130* (50), 16985−16995.
(25) Ahn, H. S.; Bard, A. J. Electrochemical surface interrogation of a MoS2 hydrogen-evolving catalyst: in situ determination of the surface hydride coverage and the hydrogen evolution kinetics. *J. Phys. Chem. Lett.* **2016**, *7* (14), 2748−2752.
(26) Ahn, H. S.; Bard, A. J. Surface interrogation scanning electrochemical microscopy of Ni1-x Fe x OOH (0< x< 0.27) oxygen evolving catalyst: kinetics of the "fast" iron sites. *J. Am. Chem. Soc.* **2016**, *138* (1), 313−318.
(27) Ahn, H. S.; Bard, A. J. Surface interrogation of CoPi water oxidation catalyst by scanning electrochemical microscopy. *J. Am. Chem. Soc.* **2015**, *137* (2), 612−615.
(28) Barforoush, J. M.; Seuferling, T. E.; Jantz, D. T.; Song, K. R.; Leonard, K. C. Insights into the active electrocatalytic areas of layered double hydroxide and amorphous nickel-iron oxide oxygen evolution electrocatalysts. *ACS Appl. Energy Mater.* **2018**, *1* (4), 1415−1423.
(29) Arroyo-Currás, N. y.; Bard, A. J. Iridium oxidation as observed by surface interrogation scanning electrochemical microscopy. *J. Phys. Chem. C* **2015**, *119* (15), 8147−8154.
(30) Geiger, S.; Kasian, O.; Ledendecker, M.; Pizzutilo, E.; Mingers, A. M.; Fu, W. T.; Diaz-Morales, O.; Li, Z.; Oellers, T.; Fruchter, L.; et al. The stability number as a metric for electrocatalyst stability benchmarking. *Nat. Catal.* **2018**, *1* (7), 508−515.
(31) Bozal-Ginesta, C.; Rao, R. R.; Mesa, C. A.; Liu, X.; Hillman, S. A.; Stephens, I. E.; Durrant, J. R. Redox-state kinetics in water-oxidation IrO x electrocatalysts measured by operando spectroelectrochemistry. *ACS Catal.* **2021**, *11* (24), 15013−15025.
(32) Bozal-Ginesta, C.; Rao, R. R.; Mesa, C. A.; Wang, Y.; Zhao, Y.; Hu, G.; Antón-García, D.; Stephens, I. E.; Reisner, E.; Brudvig, G. W.; et al. Spectroelectrochemistry of water oxidation kinetics in molecular versus heterogeneous oxide iridium electrocatalysts. *J. Am. Chem. Soc.* **2022**, *144* (19), 8454−8459.
(33) Pfeifer, V.; Jones, T.; Velasco Vélez, J.; Massué, C.; Greiner, M.; Arrigo, R.; Teschner, D.; Girgsdies, F.; Scherzer, M.; Allan, J.; et al. The electronic structure of iridium oxide electrodes active in water splitting. *Phys. Chem. Chem. Phys.* **2016**, *18* (4), 2292−2296.
(34) Ahn, H. S.; Bard, A. J. Switching transient generation in surface interrogation scanning electrochemical microscopy and time-of-flight techniques. *Anal. Chem.* **2015**, *87* (24), 12276−12280.
(35) Velasco Vélez, J. J.; Bernsmeier, D.; Mom, R. V.; Zeller, P.; Shao-Horn, Y.; Roldan Cuenya, B.; Knop-Gericke, A.; Schlögl, R.; Jones, T. E. Iridium oxide coordinatively unsaturated active sites govern the electrocatalytic oxidation of water. *Adv. Energy Mater.* **2024**, *14* (19), 2303407.
(36) Velasco-Vélez, J.-J. s.; Carbonio, E. A.; Chuang, C.-H.; Hsu, C.-J.; Lee, J.-F.; Arrigo, R.; Hävecker, M.; Wang, R.; Plodinec, M.; Wang, F. R.; et al. Surface electron-hole rich species active in the electrocatalytic water oxidation. *J. Am. Chem. Soc.* **2021**, *143* (32), 12524−12534.
(37) Frevel, L. J.; Mom, R.; Velasco-Vélez, J.-J. s.; Plodinec, M.; Knop-Gericke, A.; Schlögl, R.; Jones, T. E. In situ X-ray spectroscopy of the electrochemical development of iridium nanoparticles in confined electrolyte. *J. Phys. Chem. C* **2019**, *123* (14), 9146−9152.
(38) Feng, Z.; Dai, C.; Shi, P.; Lei, X.; Guo, R.; Wang, B.; Liu, X.; You, J. Seven mechanisms of oxygen evolution reaction proposed recently: A mini review. *Chem. Eng. J.* **2024**, *485*, 149992.
(39) Chen, L.; Zhao, W.; Zhang, J.; Liu, M.; Jia, Y.; Wang, R.; Chai, M. Recent Research on Iridium-Based Electrocatalysts for Acidic Oxygen Evolution Reaction from the Origin of Reaction Mechanism. *Small* **2024**, *20* (43), 2403845.
(40) Ooka, H.; Yamaguchi, A.; Takashima, T.; Hashimoto, K.; Nakamura, R. Efficiency of oxygen evolution on iridium oxide determined from the pH dependence of charge accumulation. *J. Phys. Chem. C* **2017**, *121* (33), 17873−17881.
(41) Willinger, E.; Massué, C.; Schlögl, R.; Willinger, M. G. Identifying key structural features of IrOx water splitting catalysts. *J. Am. Chem. Soc.* **2017**, *139* (34), 12093−12101.
(42) Dickens, C. F.; Montoya, J. H.; Kulkarni, A. R.; Bajdich, M.; Nørskov, J. K. An electronic structure descriptor for oxygen reactivity at metal and metal-oxide surfaces. *Surf. Sci.* **2019**, *681*, 122−129.
(43) Hao, S.; Liu, M.; Pan, J.; Liu, X.; Tan, X.; Xu, N.; He, Y.; Lei, L.; Zhang, X. Dopants fixation of Ruthenium for boosting acidic oxygen evolution stability and activity. *Nat. Commun.* **2020**, *11* (1), 5368.
(44) Liang, C.; Katayama, Y.; Tao, Y.; Morinaga, A.; Moss, B.; Celorrio, V.; Ryan, M.; Stephens, I. E.; Durrant, J. R.; Rao, R. R. Role of electrolyte pH on water oxidation for iridium oxides. *J. Am. Chem. Soc.* **2024**, *146* (13), 8928−8938.
(45) Ooka, H.; Takashima, T.; Yamaguchi, A.; Hayashi, T.; Nakamura, R. Element strategy of oxygen evolution electrocatalysis based on in situ spectroelectrochemistry. *Chem. Commun.* **2017**, *53* (53), 7149−7161.
(46) Grimaud, A.; Hong, W. T.; Shao-Horn, Y.; Tarascon, J.-M. Anionic redox processes for electrochemical devices. *Nat. Mater.* **2016**, *15* (2), 121−126.
(47) Sanchez Casalongue, H. G.; Ng, M. L.; Kaya, S.; Friebel, D.; Ogasawara, H.; Nilsson, A. In situ observation of surface species on iridium oxide nanoparticles during the oxygen evolution reaction. *Angew. Chem., Int. Ed.* **2014**, *53* (28), 7169−7172.
(48) Mo, Y.; Stefan, I. C.; Cai, W.-B.; Dong, J.; Carey, P.; Scherson, D. A. In situ iridium LIII-edge X-ray absorption and surface enhanced Raman spectroscopy of electrodeposited iridium oxide films in aqueous electrolytes. *J. Phys. Chem. B* **2002**, *106* (14), 3681−3686.
(49) Pedersen, A. F.; Escudero-Escribano, M.; Sebok, B.; Bodin, A.; Paoli, E.; Frydendal, R.; Friebel, D.; Stephens, I. E.; Rossmeisl, J.; Chorkendorff, I.; Nilsson, A. Operando XAS study of the surface oxidation state on a monolayer IrOx on RuOx and Ru oxide based nanoparticles for oxygen evolution in acidic media. *J. Phys. Chem. B* **2018**, *122* (2), 878−887.
(50) Minguzzi, A.; Locatelli, C.; Lugaresi, O.; Achilli, E.; Cappelletti, G.; Scavini, M.; Coduri, M.; Masala, P.; Sacchi, B.; Vertova, A.; et al. Easy accommodation of different oxidation states in iridium oxide nanoparticles with different hydration degree as water oxidation electrocatalysts. *ACS Catal.* **2015**, *5* (9), 5104−5115.

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

(51) Velasco-Vélez, J.; Jones, T.; Streibel, V.; Hävecker, M.; Chuang, C.-H.; Frevel, L.; Plodinec, M.; Centeno, A.; Zurutuza, A.; Wang, R.; et al. Electrochemically active Ir NPs on graphene for OER in acidic aqueous electrolyte investigated by in situ and ex situ spectroscopies. *Surf. Sci.* **2019**, *681*, 1−8.
(52) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books, 2006.
(53) Nishimoto, T.; Shinagawa, T.; Naito, T.; Takanabe, K. Microkinetic assessment of electrocatalytic oxygen evolution reaction over iridium oxide in unbuffered conditions. *J. Catal.* **2020**, *391*, 435−445.
(54) Kasian, O.; Geiger, S.; Li, T.; Grote, J.-P.; Schweinar, K.; Zhang, S.; Scheu, C.; Raabe, D.; Cherevko, S.; Gault, B.; Mayrhofer, K. J. J. Degradation of iridium oxides via oxygen evolution from the lattice: correlating atomic scale structure with reaction mechanisms. *Energy Environ. Sci.* **2019**, *12* (12), 3548−3555.

https://doi.org/10.1021/jacs.5c16721
J. Am. Chem. Soc. 2026, 148, 957−966

Check for updates

RESEARCH ARTICLE

NANO · MICRO small
www.small-journal.com

Synthetic Neuromelanin Particles for Metal-Free Photocatalysis

Joowon Choi, Hyeri Jeon, Sunyoung Hwang, Hangil Lee, Seungwoo Hong,* and Kyungtae Kang**

Neuromelanin, a brain pigment comprising a pheomelanin core encased in a eumelanin shell, offers a unique bioinspired blueprint for functional nanomaterials. Here, synthetic neuromelanin particles (NMPs) with tunable shell thicknesses are reported, prepared via co-polymerization of dopamine and L-cysteine under alkaline oxidative conditions. Electron microscopy and spectroscopy reveal a core–shell nanostructure in which sulfur-rich, spin-active pheomelanin cores are progressively shielded by eumelanin shells of variable thickness. The thinnest-shell NMPs exhibit the highest sulfur-to-carbon ratio and EPR spin concentration (mass-normalized double-integrated signal); their spectra show a single line at g = 2.005, characteristic of semiquinone/carbon-centered radicals. The spin concentration closely correlates with the photocatalytic activity in the visible-light oxidation of 5-hydroxymethylfurfural to 2,5-furandicarboxylic acid. Transient photocurrent measurements confirm superior charge-separation efficiency, consistent with radical-mediated photophysics intrinsic to the core. This neuromelanin-inspired design establishes a new class of metal-free organic photocatalysts, in which nanoscale architecture dictates radical density and catalytic performance.

1. Introduction

Melanin pigments are widely present in various organisms, serving diverse functions such as coloration,[1–4] hormonal regulation,[5] antimicrobial protection,[6] thermal regulation,[7] photoprotection,[8–10] and radioprotection,[11] depending on the biological context in which they are produced. Despite being known for centuries, the chemical structures of the simplest variant of melanin—i.e., eumelanin—remain incompletely understood, as they are composed of largely heterogeneous oligomeric species (e.g., 5,6-dihydroxyindole and its derivatives).[12,13] These structures, held together by non-covalent interactions, rendered eumelanin inherently challenging to handle in aqueous systems, purify, and analyze at the molecular level. Recent evidences emphasize that not only the structural diversity of these oligomers, but the way they are assembled is a critical factor for determining the macroscopic properties of eumelanin.[14–16] Despite these chemical ambiguities, chemical synthesis of eumelanin derivatives into various forms (e.g., particles, films, and composites) is simple, and a vast amount of applications of eumelanin derivatives have been reported.[17–21]

In humans, the two primary types of melanin are eumelanin and pheomelanin, synthesized through the oxidative polymerization of L-dihydroxyphenylalanine (L-dopa) and L-dopa combined with L-cysteine, respectively.[22] While pheomelanin is a simple variant of eumelanin, it has received far less attention in studies on its chemical properties, material characteristics, and biological roles. A particularly intriguing aspect of pheomelanin is its photoreactivity under UV radiation (occasionally even without irradiation[23]), as it generates reactive oxygen species and is considered a major contributor to skin cancer.[24–26] Unlike eumelanin, which has well-documented protective functions, no widely accepted biological benefits of pheomelanin have been identified beyond its role in pigmentation. Although some hypotheses propose that pheomelanin may play a role in antimicrobial defense,[27] environmental sensing, neuroprotection,[28,29] and thermoregulation,[30] these remain speculative with few experimental evidences.

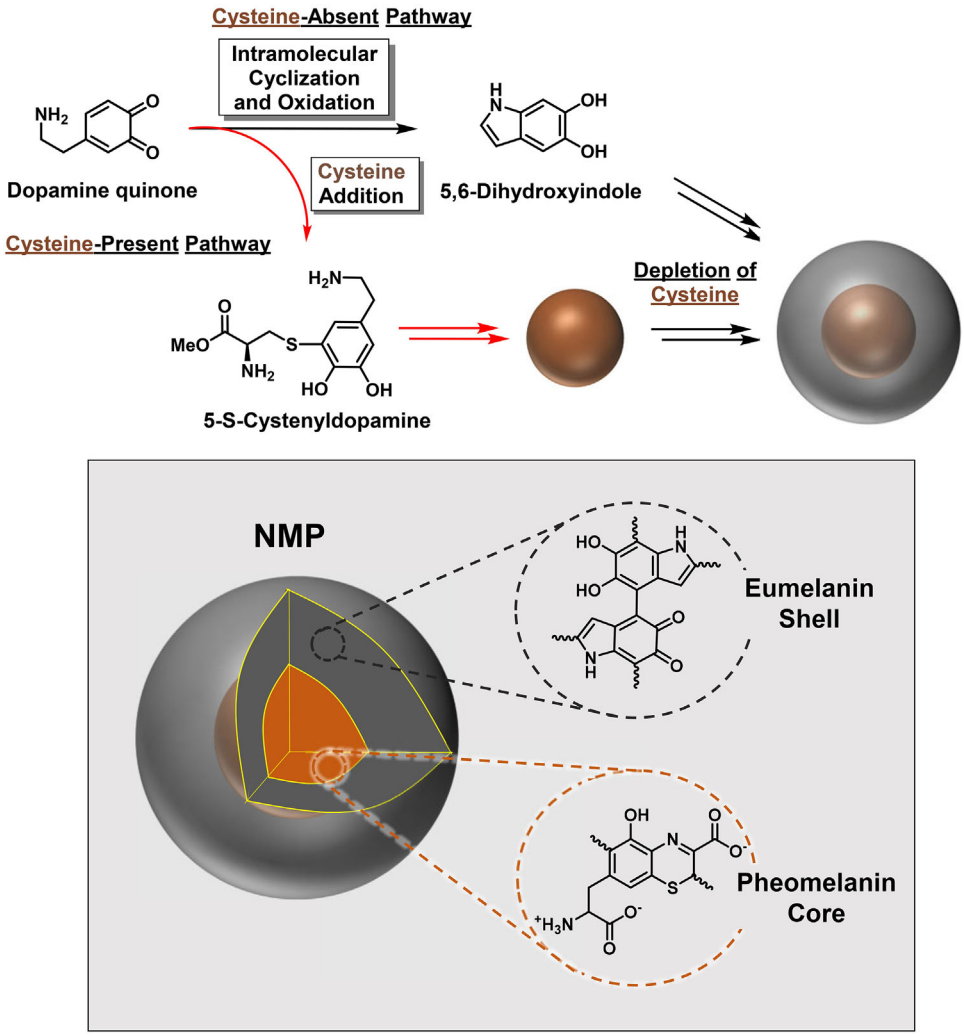
In the brain, neuromelanin is an intriguing variant of melanin pigments, comprising a hybrid of eumelanin and pheomelanin. It is abundantly present in dopaminergic neurons of the substantia nigra.[31–33] Like other types of melanin pigments, the chemical structures of neuromelanin remain relatively underexplored, yet spectroscopic evidence suggests that neuromelanin

J. Choi, K. Kang
Department of Applied Chemistry
Kyung Hee University
Yongin, Gyeonggi 17104, Republic of Korea
E-mail: kkang@khu.ac.kr
H. Jeon, S. Hong
Department of Chemistry & Nanoscience
Ewha Womans University
Seoul 03760, Republic of Korea
E-mail: hsw@ewha.ac.kr
H. Jeon, S. Hong
Graduate Program in Innovative Biomaterials Convergence
Ewha Womans University
Seoul 03760, Republic of Korea
S. Hwang, H. Lee
Department of Chemistry
Sookmyung Women's University
Seoul 04310, Republic of Korea
E-mail: easyscan@sookmyung.ac.kr

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/smll.202510311

DOI: 10.1002/smll.202510311

 © 2025 Wiley-VCH GmbH



Scheme 1. Oxidation pathways of dopamine under cysteine-absent and cysteine-present conditions (schematic). In the cysteine-absent route, dopamine is oxidized to dopamine quinone, which cyclizes to leucodopaminochrome and oxidizes to dopaminochrome; subsequent rearrangement affords 5,6-dihydroxyindole, shown here as a representative eumelanin repeating unit in the shell. In the cysteine-present route, dopamine quinone undergoes 1,4-Michael addition with cysteine to yield 5-S-cysteinyldopamine, a key pheomelanin precursor illustrated in the core. Intermediates are indicated in the caption for clarity but not drawn explicitly to maintain focus on the NMP core–shell architecture.

may adopt a core–shell configuration, with pheomelanin forming the core and eumelanin surrounding it as a protective shell.[34] This structural arrangement appears to be unique to neuromelanin and is not observed in pheomelanin pigments found in epidermal cells, which undergo turnover every few weeks, whereas brain cells do not regenerate. These findings suggest that pheomelanin in brains, despite its vulnerability and reactivity, may serve chemical functions beyond what is currently understood. Furthermore, this insight provides a basis for harnessing its reactivity in the design of artificially synthesized melanin pigments.

Recently, pheomelanin started gaining attention as a cost-effective and environmentally friendly alternative for various applications thanks to its properties distinguishable from those of eumelanin, including photoresponsiveness and oxidative potential.[35–39] However, fewer pheomelanin-like materials have been reported compared to eumelanin-like counterparts, which can be attributed to the usage of cysteine in synthetic procedures. Cysteine or its derivatives tend to complicate the process of melanin formation leading to materials with lower stability against environmental factors such as pH and light.[40] These challenges, combined with the unpredictable properties of the resulting materials, have hindered the exploration of pheomelanin's photocatalytic potential.

In this study, we drew inspiration from the natural structure of neuromelanin to synthesize neuromelanin particles (NMPs) with a pheomelanin core surrounded by an eumelanin shell (**Scheme 1**). Such core–shell structures could be simply obtained by adding L-cysteine during the chemical oxidation of dopamine. Introducing L-cysteine during dopamine oxidation directs the early pathway toward pheomelanin-like products by preferentially trapping oxidized dopamine species. After L-cysteine is consumed, subsequent oxidative polymerization gives rise to the surrounding eumelanin-like shell. By adjusting the ratio of

© 2025 Wiley-VCH GmbH

16136829, 2025, 52, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202510311 by Sookmyung Womens University, Wiley Online Library on [29/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

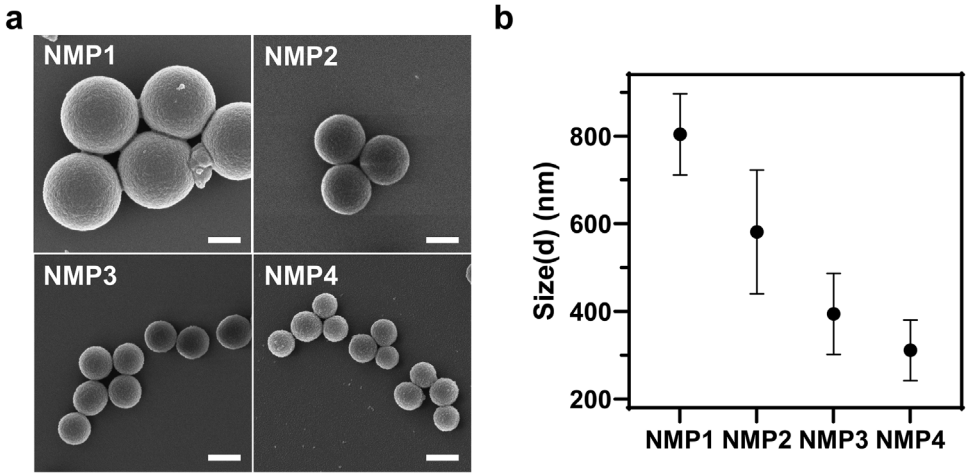


Figure 1. a) Representative scanning electron microscopy images of NMPs synthesized at different amounts of ammonia solution, NMP1: 0.5 mL, NMP2: 1.0 mL, NMP3: 1.5 mL, NMP4: 2.0 mL. Scale bar: 300 nm. b) Size difference of NMP*i* (*i*=1-4).

dopamine to L-cysteine, we were able to precisely control the thickness of the eumelanin shell while maintaining the general properties of the NMPs. These particles were employed as metal-free photocatalysts for the degradation of organic pollutants and the selective oxidation of 2,5-hydroxymethylfurfural (HMF).[41] Our findings indicate that the photocatalytic performance of NMPs was inversely related to the thickness of the eumelanin shell, supporting our hypothesis that the pheomelanin core serves as the reactive center, while the eumelanin layer functions as a protective barrier.

2. Results and Discussion

2.1. Synthesis of Size-Dependent NMPs

The synthesis of NMPs was performed via chemical oxidation of an aqueous solution containing dopamine and cysteine (dopamine: cysteine = 1, mol/mol) and aqueous ammonia solution (28 wt.%) as a catalyst at room temperature under stirring for 96 h. The presence of cysteine interferes with dopamine oxidation, thereby slowing the kinetics of melanin formation.[39] As a result, the solution turned black abruptly after a significant amount of quiescent period in which the solution remained clear (Figure S1 and Video S1, Supporting Information). Particles were only barely observable before the color change (Figure S2, Supporting Information). According to previous reports on melanin particle formation,[42] we attempted to control the size of NMPs by varying the concentration of the base, NH_4OH, which deprotonates dopamine's phenolic groups, thereby accelerating O_2-driven autoxidation and polymerization. As expected, the amount of NH_4OH was inversely proportional to the final particle size (**Figure 1a**). By adjusting the relative concentration of NH_4OH to dopamine, we successfully obtained NMPs with a broad size distribution, ranging from 250 to 800 nm in diameter (Figure 1b). Such morphological features were comparable to conventional eumelanin particles (EMPs) synthesized under conditions omitting cysteine (Figure S3, Supporting Information).

The internal structure of NMPs was examined using transmission electron microscopy (TEM). **Figure 2a** and Figure S4a (Supporting Information) reveal that NMPs exhibit internal contrast differences corresponding to electron density variation, suggesting a core–shell architecture. To substantiate the core–shell

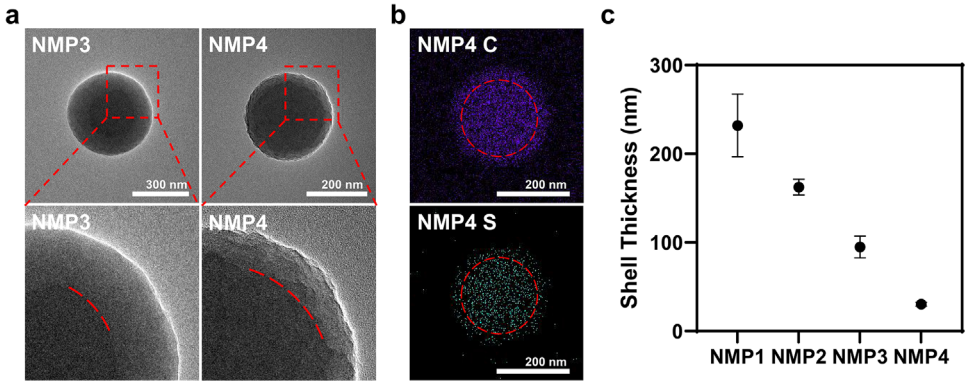


Figure 2. a) (Top) Representative TEM images and (Bottom) enlarged region inside the red dashed box TEM images of NMP3 and NMP4. b) (Top) Representative TEM EDS elemental mapping images of Carbon and (Bottom) Sulfur of NMP4. The interface between the pheomelanin-core and the eumelanin-shell was marked by a red dashed line in both (a) and (b). c) Shell thickness of NMPs.

© 2025 Wiley-VCH GmbH

16136829, 2025, 52, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202510311 by Sookmyung Womens University, Wiley Online Library on [29/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

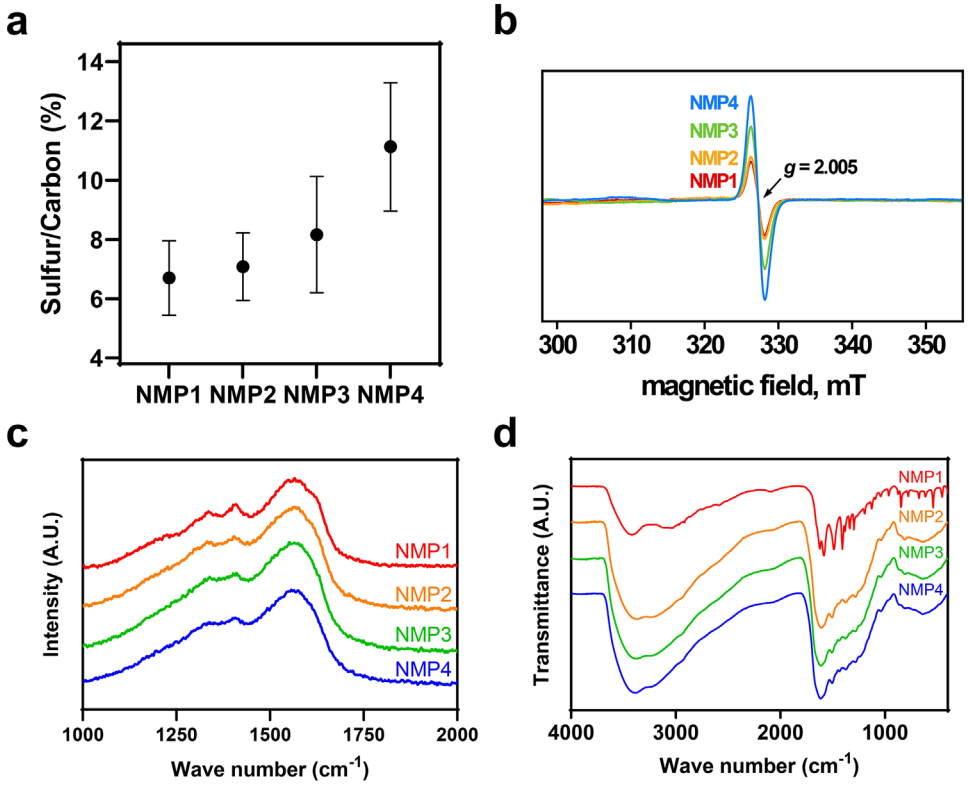


Figure 3. Characterization of NMPs. a) Sulfur-to-carbon ratio of NMPs. b) EPR spectra, c) Raman spectra, d) FTIR spectra of NMPs.

structure and analyze the spatial distribution characteristics of the elements, TEM-Energy Dispersive X-ray Spectroscopy (EDS) elemental mapping is presented in Figure 2b. The analysis confirmed that carbon (C) exhibited a uniform and dense distribution across both the core and shell regions, whereas sulfur (S) was observed to have a significantly lower density in the shell region compared to the core region. In contrast, EMPs lacked such contrast, suggesting that the core–shell structure is specifically induced by the presence of cysteine (Figure S4b, Supporting Information). Notably, the relative thickness of the shell slightly decreased as the overall particle size decreased (Figure 2b).

To further investigate their chemical composition, we conducted X-ray photoelectron spectroscopy (XPS), electron paramagnetic resonance (EPR), Raman, and Fourier-transform infrared (FTIR) spectroscopy. In consistent with the TEM data, XPS analysis revealed that NMPs with thinner shells contained higher sulfur-to-carbon (S/C) ratios, implying that the core region is rich in pheomelanin-like components, while the shell is more characteristic of eumelanin (**Figure** 3a). EPR spectra revealed characteristic single-line signals centered at $g = 2.005$, consistent with stable organic radicals commonly observed in melanin-like materials (Figure 3b). Among the samples, NMP4 exhibited the highest EPR signal intensity, indicating a significantly greater population of unpaired electrons. This higher radical density suggests an increased exposure or concentration of pheomelanin-like components in NMP4, which is in agreement with the thinner eumelanin shell observed by TEM and the higher S/C ratio from XPS analysis (Figure S5, Supporting Information). Taken together, these results indicate that NMP4 harbors a larger proportion of its structure as a reactive pheomelanin core.

The Raman spectroscopy shows the major bands characteristic of sp^2 carbon materials. The D band (1370 cm^{-1}) is related to the stretching of the hexagonal carbon rings, while the G band (1580 cm^{-1}) is related to the stretching of three of the six C—C bonds within the rings.[43,44] NMPs exhibited two prominent bands at 1370 cm^{-1} (D band) and 1580 cm^{-1} (G band) (Figure 3c), which were attributable to catechol group stretching vibrations and ring deformations, as shown in other melanin-like materials.[45,46]

The FTIR spectroscopy showed a broad peak ≈3400 cm^{-1} assigned to the stretching vibration of N—H and O—H, the double sharp peaks at 1616 and 1505 cm^{-1} were attributed to the stretching vibration of C═C in aromatic ring, and another double sharp peaks at 1375 and 1282 cm^{-1} related to the bonding vibration of C—O—H in the catechol moieties (Figure 3d). The observed peaks were consistent with the previous reports.[38,47] These results indicate that the chemical composition of the shell regions remains similar across different particle sizes.

2.2. Photocatalytic Performances of NMPs

Figure 4 illustrates the photocatalytic conversion efficiency of various NMPs under visible light over time. Figure 4a demonstrated that the NMP4 achieved significantly superior HMF photocatalytic activity, with performance ordered as follows:

© 2025 Wiley-VCH GmbH

16136829, 2025, 52, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202510311 by Sookmyung Womens University, Wiley Online Library on [29/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

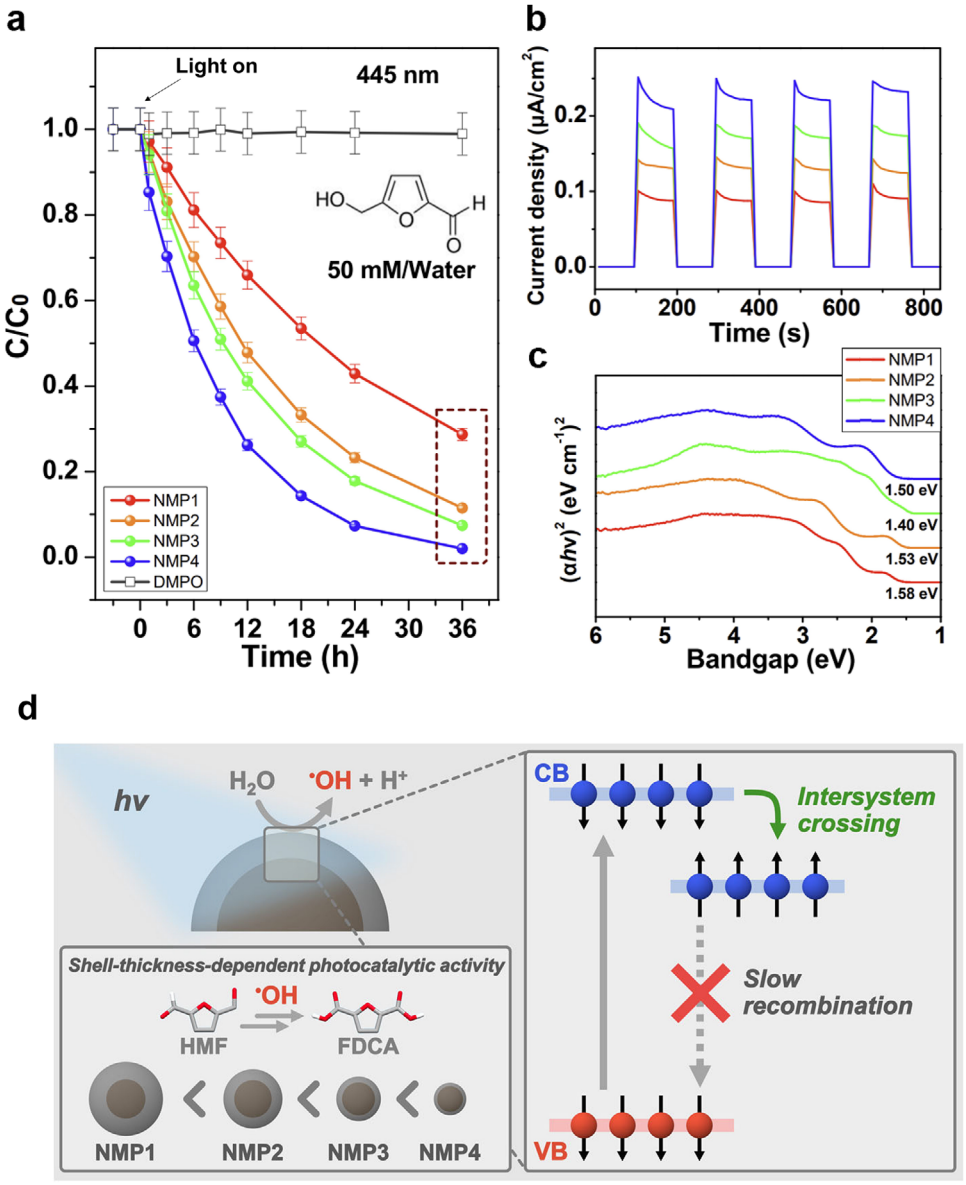


Figure 4. a) Photocatalytic conversion activities of NMP*i* (*i*=1-4) toward HMF as functions of time when irradiated at 445 nm. Experimental conditions: [catalyst] = 0.5 g L^{-1}, $\lambda \geq 445$ nm, $[HMF]_0 = 50$ mM. b) photocurrent density data, c) UV–vis DRS spectra. d) Schematic illustration of charge separation in the photocatalytic oxidation of HMF catalyzed by NMPs.

NMP4 > NMP3 > NMP2 > NMP1. Transient photocurrent responses of four NMPs generated under 445 nm irradiation were measured over four on–off cycles to examine charge-separation efficiencies of photogenerated electron–hole pairs within NMPs (Figure 4b). The four NMPs generated photocurrents upon irradiation, with current densities dropping to zero when the light was off. NMP4 exhibited ≈2.5-fold higher current density compared to NMP1. UV–vis diffuse reflectance spectroscopy and Kubelka–Monk plots estimated bandgaps of 1.58 eV (NMP1), 1.53 eV (NMP2), 1.40 eV (NMP3), and 1.50 eV (NMP4), indicating negligible impact of size differences on bandgap values.

Organic products originated from the photocatalytic oxidation of HMF in the presence of four NMPs in aqueous media were analyzed by high-performance liquid chromatography, and the product concentrations were plotted over time (Figure S6, Supporting Information). **Table 1** summarizes data from photocat-

Table 1. C/C_0 values for the photocatalytic degradation of HMF and yield of FDCA from HMF over the four NMP under irradiation at 445 nm for 36 h.

	NMP1	NMP2	NMP3	NMP4
HMF	0.287 ± 0.014	0.115 ± 0.024	0.074 ± 0.004	0.02 ± 0.001
FDCA (%)	18	23	33	45

C and C_0 are the time-dependent and initial concentrations of HMF, respectively.

© 2025 Wiley-VCH GmbH

16136829, 2025, 52, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202510311 by Sookmyung Womens University, Wiley Online Library on [29/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

alytic oxidation of HMF over the four NMPs after 36 h of irradiation at 445 nm.

Since —OH and —CHO oxidation reactions are critical for the synthesis of 2,5-furandicarboxylic acid (FDCA) from HMF, it is noteworthy to highlight the need to identify the functional groups involved in HMF oxidation in the presence of four NMPs. Interestingly, all NMPs exhibited an induction period during which the yield of FDCA remained near zero as HMF was selectively oxidized to 2,5-diformylfuran (DFF) within the first 6 h, followed by a gradual decrease in DFF concentration due to its oxidation and concurrent formation of FDCA, as the final product (see Scheme S1, Supporting Information). The formation efficiency of FDCA mirrored the HMF photocatalytic conversion efficiency trend; NMP4 > NMP3 > NMP2 > NMP1. Across NMP1–NMP4, photocatalytic performance tracks transient photocurrent and EPR spin density, whereas optical gaps are similar, indicating that interfacial charge separation modulated by shell thickness governs activity.

In view of the persistent radicals inherent to NMPs, we propose that the photocatalytic oxidation of HMF proceeds through the mechanism illustrated in Figure 4d. The presence of radical species with unpaired electrons in eumelanin and pheomelanin could play an important role in electron-hole charge separation because unpaired electrons can introduce spin dynamics,[48] which influence the efficiency of charge separation and the recombination rate by stabilizing separated charges.[49] Photoexcitation generates ion radical pairs whose presence stabilizes the formation of the excited state via intersystem crossing. As recombination to the ground state is a spin-forbidden process, this excited state exhibits substantially prolonged lifetimes and a correspondingly low recombination yield. Inclusion of 5,5-dimethyl-1-pyrroline N-oxide (DMPO), a well-established radical scavenger, entirely suppressed photocatalytic activity, unequivocally demonstrating the involvement of radical species in the process (Figure 4a). To further evaluate the radical species produced in the photocatalytic process, we employed terephthalic acid (TA) as a selective molecular probe for hydroxyl radical (·OH).[50] Upon reaction with ·OH, TA underwent hydroxylation to yield 2-hydroxyterephthalic acid (TA-OH), a fluorescent species. In the absence of light, no appreciable fluorescence was observed. However, when an aqueous dispersion of NMPs (with TA present) was continuously irradiated at 445 nm for 6 h, a distinct fluorescent emission at 428 nm–a characteristic feature of TA-OH–emerged, indicating the formation of ·OH (Figure S7, Supporting Information). These results confirmed that the non-metallic NMP photocatalyst actively produced hydroxyl radicals under visible light, which were implicated in the oxidative transformation of biomass.

To illustrate generality, we examined additional substrates under the standard conditions: toluene was oxidized sequentially to benzyl alcohol and then benzaldehyde (Figure S8, Supporting Information).

3. Conclusion

Recent studies have shown that melanin-like polymers can act as visible-light photocatalysts for organic oxidations and pollutant remediation, and that hybrid melanin–inorganic constructs can further enhance activity and stability.[51–54] Building on this body of work, we have developed a set of neuromelanin-inspired core–shell particles, mimicking the natural configuration of pheomelanin and eumelanin in dopaminergic neurons, to achieve efficient metal-free photocatalysis. The design allows precise tuning of eumelanin shell thickness while preserving a reactive pheomelanin-rich core. EPR spectroscopy confirmed that the radical content increases with decreasing shell thickness, with NMP4 displaying the highest spin density, correlating strongly with its superior photocatalytic performance. Complementary XPS data further support the enrichment of sulfur-containing components in thinner-shelled particles, consistent with a pheomelanin-like composition. Photocatalytic studies demonstrate that shell thickness modulates the accessibility of active sites and electron transfer, offering a new strategy to optimize melanin-based catalysts. By leveraging the intrinsic redox activity of pheomelanin within a structurally stabilized framework, this work provides a new avenue for sustainable photocatalysis using entirely organic, bioinspired materials. In particular, our results position architecture—a pheomelanin core enveloped by a tunable eumelanin shell—as a practical lever to regulate radical content and charge separation in metal-free, bioinspired photocatalysts. Rather than competing for record FDCA yields, our study provides a design rule that can be translated to other melanin chemistries, geometries (films/flow), and substrates.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

J.C. and H.J. contributed equally to this work. This work was supported by the Technology Innovation Program (RS-2024-00419769 to K.K.), funded by the Ministry of Trade, Industry & Energy (MOTIE, Republic of Korea), the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (RS-2025-16071328 for K.K., RS-2024-00346153 and RS-2025-21072982 for H.L., RS-2023-00272031 for H.J. and 2021R1A6A1A10039823, 2020R1A6C101B194 and RS-2025-20942968 for S.H.), and Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Ministry of Education (2020R1A6C101B194 for S.H.).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

core–shell nanostructures, melanin polymers, pheomelanin, photocatalysis, spin-active organic materials

Received: August 24, 2025
Revised: November 3, 2025
Published online: November 19, 2025

© 2025 Wiley-VCH GmbH

16136829, 2025, 52, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202510311 by Sookmyung Womens University, Wiley Online Library on [29/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[1] L. Siefferman, G. E. Hill, *Behav. Ecol.* **2003**, *14*, 855.
[2] K. J. McGraw, *Pigm. Cell Melanoma Res.* **2008**, *21*, 133.
[3] Y. Matsuoka, A. Monteiro, *Cell Rep.* **2018**, *24*, 56.
[4] R. C. Duarte, A. A. V. Flores, M. Stevens, *Philos. Trans. R. Soc., B* **2017**, *372*, 20160342.
[5] A. Slominski, D. J. Tobin, S. Shibahara, J. Wortsman, *Physiol. Rev.* **2004**, *84*, 1155.
[6] A. Zerrad, J. Anissi, J. Ghanam, K. Sendide, M. El Hassouni, *J. Biotechnol.Lett.* **2014**, *5*, 87.
[7] S. Pinkert, D. Zeuss, *Curr. Biol.* **2018**, *28*, R887.
[8] N. Kollias, R. M. Sayre, L. Zeise, M. R. Chedekel, *J. Photochem. Photobiol., B* **1991**, *9*, 135.
[9] M. Brenner, V. J. Hearing, *Photochem. Photobiol.* **2008**, *84*, 539.
[10] B.-L. L. Seagle, K. A. Rezai, Y. Kobori, E. M. Gasyna, K. A. Rezaei, J. R. Norris Jr, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 8978.
[11] E. Dadachova, R. A. Bryan, R. C. Howell, A. D. Schweitzer, P. Aisen, J. D. Nosanchuk, A. Casadevall, *Pigm. Cell Melanoma Res.* **2008**, *21*, 192.
[12] W. Cao, X. Zhou, N. C. McCallum, Z. Hu, Q. Z. Ni, U. Kapoor, C. M. Heil, K. S. Cay, T. Zand, A. J. Mantanona, A. Jayaraman, A. Dhinojwala, D. D. Deheyn, M. D. Shawkey, M. D. Burkart, J. D. Rinehart, N. C. Gianneschi, *J. Am. Chem. Soc.* **2021**, *143*, 2622.
[13] L. Panzella, G. Gentile, G. D'Errico, N. F. Della Vecchia, M. E. Errico, A. Napolitano, C. Carfagna, M. d'Ischia, *Angew. Chem., Int. Ed. Engl.* **2013**, *52*, 12684.
[14] M. d'Ischia, A. Napolitano, A. Pezzella, P. Meredith, M. Buehler, *Angew. Chem., Int. Ed. Engl.* **2020**, *59*, 11196.
[15] Y. Shen, R. Su, D. Hao, X. Xu, M. Reches, J. Min, H. Chang, T. Yu, Q. Li, X. Zhang, Y. Wang, Y. Wang, W. Qi, *Nat. Commun.* **2023**, *14*, 3054.
[16] X. Wang, L. Kinziabulatova, M. Bortoli, A. Manickoth, M. A. Barilla, H. Huang, L. Blancafort, B. Kohler, J.-P. Lumb, *Nat. Chem.* **2023**, *15*, 787.
[17] Y. Liu, K. Ai, L. Lu, *Chem. Rev.* **2014**, *114*, 5057.
[18] P. Yang, F. Zhu, Z. Zhang, Y. Cheng, Z. Wang, Y. Li, *Chem. Soc. Rev.* **2021**, *50*, 8319.
[19] M. Xiao, M. D. Shawkey, A. Dhinojwala, *Adv. Opt. Mater.* **2020**, *8*, 2000932.
[20] D. T. Sun, L. Peng, W. S. Reeder, S. M. Moosavi, D. Tiana, D. K. Britt, E. Oveisi, W. L. Queen, *ACS Cent. Sci.* **2018**, *4*, 349.
[21] S. Chen, Y. Cao, J. Feng, *ACS Appl. Mater. Interfaces* **2014**, *6*, 349.
[22] R. Micillo, L. Panzella, K. Koike, G. Monfrecola, A. Napolitano, M. d'Ischia, *Int. J. Mol. Sci.* **2016**, *17*, 746.
[23] D. Mitra, X. Luo, A. Morgan, J. Wang, M. P. Hoang, J. Lo, C. R. Guerrero, J. K. Lennerz, M. C. Mihm, J. A. Wargo, K. C. Robinson, S. P. Devi, J. C. Vanover, J. A. D'Orazio, M. McMahon, M. W. Bosenberg, K. M. Haigis, D. A. Haber, Y. Wang, D. E. Fisher, *Nature* **2012**, *491*, 449.
[24] S. Ito, K. Wakamatsu, T. Sarna, *Photochem. Photobiol.* **2018**, *94*, 409.
[25] H. Z. Hill, G. J. Hill, *Pigm. Cell Res.* **2000**, *13*, 140.
[26] G. Szewczyk, A. Zadlo, M. Sarna, S. Ito, K. Wakamatsu, T. Sarna, *Pigm. Cell Melanoma Res.* **2016**, *29*, 669.
[27] J. A. Mackintosh, *J. Theor. Biol.* **2001**, *211*, 101.
[28] A. Moreno-García, A. Kun, M. Calero, O. Calero, *Antioxidants* **2021**, *10*, 124.
[29] L. Zecca, L. Casella, A. Albertini, C. Bellei, F. A. Zucca, M. Engelen, A. Zadlo, G. Szewczyk, M. Zareba, T. Sarna, *J. Neurochem.* **2008**, *106*, 1866.
[30] A. Roulin, *Global Change Biol.* **2014**, *20*, 3344.
[31] F. A. Zucca, E. Basso, F. A. Cupaioli, E. Ferrari, D. Sulzer, L. Casella, L. Zecca, *Neurotoxic. Res.* **2014**, *25*, 13.
[32] L. Zecca, D. Tampellini, M. Gerlach, P. Riederer, R. G. Fariello, D. Sulzer, *Mol. Pathol.* **2001**, *54*, 414.
[33] F. A. Zucca, G. Giaveri, M. Gallorini, A. Albertini, M. Toscani, G. Pezzoli, R. Lucius, H. Wilms, D. Sulzer, S. Ito, K. Wakamatsu, L. Zecca, *Pigm. Cell Res.* **2004**, *17*, 610.
[34] W. D. Bush, J. Garguilo, F. A. Zucca, A. Albertini, L. Zecca, G. S. Edwards, R. J. Nemanich, J. D. Simon, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 14785.
[35] E. Kim, L. Panzella, R. Micillo, W. E. Bentley, A. Napolitano, G. F. Payne, *Sci. Rep.* **2015**, *5*, 18447.
[36] M. Kang, E. Kim, Z. Temoçin, J. Li, E. Dadachova, Z. Wang, L. Panzella, A. Napolitano, W. E. Bentley, G. F. Payne, *Chem. Mater.* **2018**, *30*, 5814.
[37] W. Cao, N. C. McCallum, Q. Z. Ni, W. Li, H. Boyce, H. Mao, X. Zhou, H. Sun, M. P. Thompson, C. Battistella, M. R. Wasielewski, A. Dhinojwala, M. D. Shawkey, M. D. Burkart, Z. Wang, N. C. Gianneschi, *J. Am. Chem. Soc.* **2020**, *142*, 12802.
[38] W. Cao, H. Mao, N. C. McCallum, X. Zhou, H. Sun, C. Sharpe, J. Korpanty, Z. Hu, Q. Z. Ni, M. D. Burkart, M. D. Shawkey, M. R. Wasielewski, N. C. Gianneschi, *Chem. Sci.* **2023**, *14*, 4183.
[39] J. S. Lee, J. R. Jeong, M. H. Lee, K. Kang, *ACS Appl. Mater. Interfaces* **2024**, *16*, 31768.
[40] A. Zadlo, G. Szewczyk, M. Sarna, T. G. Camenisch, J. W. Sidabras, S. Ito, K. Wakamatsu, F. Sagan, M. Mitoraj, T. Sarna, *Pigm. Cell Melanoma Res.* **2019**, *32*, 359.
[41] W. P. Dijkman, D. E. Groothuis, M. W. Fraaije, *Angew. Chem., Int. Ed.* **2014**, *53*, 6515.
[42] K. Ai, Y. Liu, C. Ruan, L. Lu, G. M. Lu, *Adv. Mater.* **2013**, *25*, 998.
[43] I. Galván, A. Jorge, K. Ito, K. Tabuchi, F. Solano, K. Wakamatsu, *Pigm. Cell Melanoma Res.* **2013**, *26*, 917.
[44] Z. Huang, H. Lui, X. K. Chen, A. Alajlan, D. I. McLean, H. Zeng, *J. Biomed. Opt.* **2004**, *9*, 1198.
[45] J. Zhang, P. Singh, Z. Cao, S. Rahimi, S. Pandit, I. Mijakovic, *Appl. Surf. Sci.* **2023**, *626*, 157221.
[46] I. Galván, A. Jorge, F. Solano, K. Wakamatsu, *Spectrochim. Acta, Part A* **2013**, *110*, 55.
[47] D. Fuentes-López, D. Ortega-Zambrano, M. A. Fernández-Herrera, H. Mercado-Uribe, *PLoS One* **2022**, *17*, 0265277.
[48] P. Meredith, T. Sarna, *Pigm. Cell Res.* **2006**, *19*, 572.
[49] I. Vlassiouk, S. Smirnov, O. Kutzki, M. Wedel, F.-P. Montforts, *J. Phys. Chem. B* **2002**, *106*, 8657.
[50] H. Park, S. Hwang, J. Choi, K. Kang, H. Lee, S. Hong, *Nanoscale* **2025**, https://doi.org/10.1039/D5NR04008F.
[51] Y. Zou, Z. Wang, Z. Chen, Q.-P. Zhang, Q. Zhang, Y. Tian, S. Ren, Y. Li, *J. Phys. Chem. C* **2019**, *123*, 5345.
[52] W. Xie, E. Pakdel, Y. Liang, D. Liu, L. Sun, X. Wang, *J. Photochem. Photobiol. A Chem.* **2020**, *389*, 112292.
[53] M. Sridharan, P. Kamaraj, R. Vennila, Y. S. Huh, M. Arthanareeswari, *New J. Chem.* **2020**, *44*, 15223.
[54] Z. Yu, F. Li, Q. Yang, H. Shi, Q. Chen, M. Xu, *ACS Sustainable Chem. Eng.* **2017**, *5*, 7840.

© 2025 Wiley-VCH GmbH

16136829, 2025, 52, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202510311 by Sookmyung Womens University, Wiley Online Library on [29/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Materials Horizons

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal | View Issue

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

Check for updates

Cite this: *Mater. Horiz.*, 2025, **12**, 10270

Received 17th June 2025,
Accepted 13th August 2025

DOI: 10.1039/d5mh01147g

rsc.li/materials-horizons

In situ study of visible light-activated photocatalytic conversion reaction enhancement of biomass using a magnetic heterostructure

Sunyoung Hwang,†[a] Hyejin Yu,†[b] Hyeon-Ah Ju, †[cd] Heewon Kim,[a] Jaeyoon Baik,[e] Young-Min Kim, *[cd] Hyun Sung Kim *[b] and Hangil Lee *[a]

Although cadmium sulfide (CdS) is a promising visible light-active photocatalyst, it is highly toxic. Therefore, in addition to improving the photocatalytic properties of CdS, enhancing its recoverability would allow for more environmentally friendly photocatalytic reactions. In this study, ferrimagnetic Fe_3O_4 nanoparticles (NPs) were added to CdS to form a heterostructure, and photocatalytic CdS NPs were efficiently recovered. The photocatalytic reaction efficiency of this heterostructure for biomass conversion was measured. Using the CdS/Fe_3O_4 heterostructure system, the photocatalytic conversion rates of 5-hydroxymethylfurfural to 5-formylfuran-2-carboxylic acid and toluene to benzaldehyde were approximately 30% and 63% higher, respectively, than those achieved using pristine CdS NPs. Furthermore, the magnetic properties were reduced by only ~6.8% compared to those of Fe_3O_4 NPs, confirming their applicability. The CdS/Fe_3O_4 heterostructure had better photocatalytic properties than CdS, and the magnetic properties were on par with those of Fe_3O_4. Changes in the electronic structure of the heterostructure system were monitored using *in situ* X-ray photoelectron spectroscopy during visible light irradiation and characterized using electron energy-loss spectroscopy in scanning transmission electron microscopy to analyze the defect structure changes occurring at the interface of the heterostructure system.

New concepts

The environmental impact of cadmium-based photocatalysts has long posed a paradox in solar-driven catalysis: high photocatalytic efficiency comes at the cost of toxic material use and limited recoverability. In this work, we introduce a magnetically recoverable CdS/Fe_3O_4 heterostructure that not only addresses the sustainability challenge of CdS reuse but also enhances photocatalytic activity *via* interfacial defect engineering. Unlike conventional heterojunction designs, this system exploits the formation of S–O bonds and oxygen vacancies at the CdS/Fe_3O_4 interface to induce local electronic structure modulation, promoting effective charge separation and reactive oxygen species (ROS) generation. *In situ* XPS and nanoscale STEM-EELS analyses reveal the dynamic interfacial charge redistribution under visible light, establishing a new paradigm for defect-modulated, multifunctional heterostructures with simultaneous activity and recyclability in biomass-derived molecule oxidation.

Introduction

Currently, visible light-activated photocatalysis research focuses on developing highly efficient and low-cost photocatalysts, mainly CdS nanoparticles (NPs) and their derivatives.[1,2] However, CdS photocatalysts suffer from high electron–hole recombination rates, a major weakness of semiconductor NPs, which limits the scope for efficiency enhancement.[3,4] Therefore, several strategies have been developed to compensate for this serious problem. Among them, heterostructures composed of two materials with complementary properties have attracted considerable attention. For enhanced photocatalytic activity, holes and electrons must be readily generated and separated on the NP surface under visible light irradiation. Thus, efficiently generating and separating photogenerated holes and electrons on the surface of the photocatalyst system under visible light is essential to enhance photocatalytic activity. In the design of CdS-based heterostructured photocatalysts, the valence band (VB) and conduction band (CB) edges of CdS and secondary semiconductor NPs are aligned in a staggered configuration. This arrangement facilitates the transfer of photogenerated holes to CdS with a higher VB edge, whereas photogenerated

[a] *Department of Chemistry and Research Institute of Natural Sciences, Sookmyung Women's University, Seoul, 04310, Republic of Korea. E-mail: easyscan@sookmyung.ac.kr*
[b] *Department of Chemistry, Pukyong National University, Busan, 48513, Republic of Korea. E-mail: kimhs75@pknu.ac.kr*
[c] *Department of Energy Science, Sungkyunkwan University (SKKU), Suwon, 16419, Republic of Korea. E-mail: youngmk@skku.edu*
[d] *Center for 2D Quantum Heterostructures, Institute for Basic Science (IBS), Suwon, 16419, Republic of Korea*
[e] *Beamline Research Division, Pohang Accelerator Laboratory (PAL), Pohang, 37673, Republic of Korea*
† All authors contributed equally to this work.

This journal is © The Royal Society of Chemistry 2025

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

electrons are transferred to secondary semiconductor NPs with a lower CB edge.[5] This allows for selective, facile oxidation reactions of energy-relevant molecules because CdS-based semiconductors can form suitable heterojunctions when combined with materials with lower CB edges. In a type-II heterojunction, the VB and CB edges of CdS and metal oxide (MO) NPs adopt a staggered alignment.[6] This configuration facilitates the migration of photogenerated electrons from CdS to the MO NPs owing to the lower CB edge of the MO NPs relative to that of the CdS NPs. Consequently, the type-II heterojunction promotes effective charge separation between the two components, enhancing their photocatalytic performance. Therefore, fabricating type-II heterojunctions is a viable approach for enhancing photocatalytic activity. In this study, we focused on type-II heterojunctions for oxidizing biomass molecules.

In addition to increasing the efficiency of visible light-activated CdS-based photocatalytic NPs, another challenge that should be considered is the recovery rate to minimize the detrimental environmental impact of photocatalytic processes. CdS is inherently toxic, and its indiscriminate use is potentially catastrophic to the global environment. Therefore, developing eco-friendly alternatives with similar photocatalytic performance or new, efficient recovery methods is essential.[7] Owing to their magnetic properties and band gap of $\sim$1.8 eV, Fe_3O_4 NPs are good candidates for fabricating visible light-active CdS-based eco-friendly heterostructured photocatalysts.[8] However, Fe_3O_4 NPs exhibit intrinsically poor photocatalytic activity and must form heterostructures to significantly enhance their performance.[9] For this purpose, CdS/Fe_3O_4 heterojunction photocatalysts were synthesized using a simple method to enhance their visible light-active photocatalytic performance and maximize the sample recovery rate. The properties of the CdS/Fe_3O_4 heterostructure system were also analyzed; particularly, the interfacial interactions between heterojunction photocatalysts were investigated. Additionally, their photocatalytic performance was compared with those of pristine CdS and Fe_3O_4 NPs to confirm that the heterojunction composition affected the photocatalytic activity. Furthermore, various characterization studies indicated that the CdS/Fe_3O_4 heterostructures exhibited photocatalytic properties that were superior to those of CdS. The recovery rates were also notable, where the CdS/Fe_3O_4 heterostructures showed a good recovery rate owing to the ferrimagnetic nature of the Fe_3O_4 NPs. Therefore, forming heterostructures consisting of Fe_3O_4 NPs, which have intrinsic magnetic properties but poor photocatalytic performance, and CdS, which has good photocatalytic performance, is expected to yield effective, environmentally friendly photocatalysts.[10]

To achieve the dual goal of increasing photocatalytic activity and recovery, Fe_3O_4 NPs were combined with CdS NPs to create CdS-based type-II heterostructures.[11] Owing to the interfacial contact between CdS and Fe_3O_4 NPs, a simple strategy to increase the photocatalytic efficiency is to enhance the formation of defect structures at this interface. Furthermore, charge transfer in the heterostructures is expected owing to the bonding between the sulfur in CdS and the oxygen in the MO NPs (–S–O–).[6,12] This phenomenon enhances the photocatalytic activity of the heterostructure by modulating the local valence states of neighbouring cations on the atomic scale, allowing the defect structures to act as mobile electron donors.[13]

In this study, we focused on heterojunction systems to increase the efficiency of oxidation of biomass molecules.[14] To address the versatility of these enhanced photocatalysts, two potential applications were investigated: the photocatalytic conversion reaction (PCR) of 5-hydroxymethylfurfural (HMF) and toluene.[15,16] HMF is a biomass-derived platform chemical that is oxidized into 2,5-furandicarboxylic acid (FDCA), providing a green alternative to petroleum-based terephthalic acid.[17] Toluene is a biomass-derived platform chemical that is oxidized into benzaldehyde (BAD), which must be degraded owing to its high volatility, environmental persistence, and toxic effects on ecosystems. The CdS/Fe_3O_4 heterostructure demonstrated multifunctionality as a photocatalyst, achieving superior oxidation of HMF (96% conversion rate) and toluene (82% conversion rate). Therefore, the enhanced photocatalytic activity of the CdS/Fe_3O_4 heterostructure can be attributed to forming a defect structure at its interface.

Previously, several research groups reported the concept for magnetic heterostructures.[18,19] In other words, the concept of heterostructures containing Fe_3O_4 is not entirely new, but we have systematized it using a new analytical method. In this study, we focused on *in situ* X-ray photoelectron spectroscopy (XPS) to investigate the photocatalytic activity of the heterostructure system under visible light irradiation.[20] Specifically, while monitoring the electronic structures of the heterostructure system using XPS, we irradiated the structures with visible light (λ = 445 nm). This allowed us to analyze the changes in the active photocatalytic sites of these systems, providing a mechanistic understanding of the photocatalytic systems. Additionally, scanning probe-based electron energy loss spectroscopy (EELS) was used to investigate the change in electronic structure occurring at the interface of the two NPs (CdS and Fe_3O_4) that comprise the heterostructure and to determine the cause of photocatalytic enhancement.

The results of the PCR of HMF and toluene under visible light irradiation revealed significantly enhanced activities of the CdS/Fe_3O_4 heterostructure, which was thought to result from an increase in the defect structure owing to the formation of S–O bonds at the heterostructure interface. Conversely, CdS/Fe_3O_4 showed no significant degradation in magnetic properties compared with pristine Fe_3O_4 NPs. This result indicates that the formation of the heterostructure does not cause any structural changes in Fe_3O_4 NPs. These compelling results indicate that the various approaches used in this study are promising for developing high-performance photocatalysts capable of addressing numerous environmental challenges.

Results and discussion

Crystal structure characterization

A series of CdS/Fe_3O_4 heterostructures, denoted as $Fe_3O_4(x)$ (x = 25, 50, and 75 wt%, where x represents the weight percentage of Fe_3O_4),

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

were prepared to optimize photocatalytic performance. The HRTEM images in Fig. 1 show the crystal structure of CdS/Fe_3O_4(50) (Fig. 1a and b) as the representative heterostructure system. The corresponding atomic-scale fast Fourier transform (FFT) images were simultaneously obtained (Fig. 1b, right panels). As shown in the magnified images of the CdS/Fe_3O_4(50) system (Fig. 1b, the region marked by the yellow box in Fig. 1a), the two samples exhibit distinct single-crystal orientations at the interface (marked by the green dashed line), which is also evident in the corresponding FFT images.[21] The CdS (101) interplanar distance in the CdS/Fe_3O_4(50) heterostructure was 3.110 Å, which is similar to that of pure CdS (3.10 Å). Additionally, the Fe_3O_4 (311) interplanar distance in the heterostructures was 2.653 Å, which is comparable to that of pure Fe_3O_4 (2.65 Å); in other words, this distance was not significantly changed by heterostructure formation.[22–24] Therefore, in this study, we focused on the interfacial activity at the contact surface between the two samples to investigate the changes in their photocatalytic properties (see Fig. 4). Fig. S1 shows the HRTEM images of the Fe_3O_4 and CdS NPs, which were obtained to confirm the structures of the two types of NPs before forming the heterostructure. As shown in Fig. S1c and f, the Fe_3O_4 (311) and CdS (100) interplanar distances were 2.56 and 3.60 Å, respectively, which are consistent with those of the previously reported structures.[25,26] The elemental distribution of the synthesized CdS/Fe_3O_4(50) heterostructure was analyzed using energy-dispersive X-ray spectroscopy (EDX), which revealed that Cd, S, Fe, and O were homogeneously distributed throughout the heterostructure (Fig. S2).

Powder X-ray diffraction (PXRD) was used to assess the crystallinity and crystal structure of the CdS/Fe_3O_4 series and individual NP samples (Fig. 1c and Fig. S3). The XRD pattern of the pristine CdS NPs (orange) is consistent with that of the hexagonal wurtzite phase of CdS (JCPDS No. 41-1049).[27] CdS NP diffraction peaks appeared at 2θ angles of 24.8, 26.5, 28.17, 36.63, 43.72, and 47.84°, which were attributed to the (002), (100), (101), (102), (110), and (103) planes of the pristine CdS NPs, respectively.[28] Similar results with high crystallinity were obtained for the Fe_3O_4 NPs, as shown in the uppermost traces in Fig. 1c. The 30.16, 35.52, and 43.14° peaks in the XRD profile of the pristine Fe_3O_4 NPs (Fig. 1c, dark red trace) are assigned to the (220), (311), and (400) reflections of Fe_3O_4 (PDF 01-1111).[29] The PXRD patterns of the CdS/Fe_3O_4 series clearly exhibit the characteristic diffraction peaks corresponding to both CdS and Fe_3O_4 nanoparticles, indicating that the formation of the composites occurred without significant alteration in the original crystal structures of the individual components.

The PXRD patterns of bare CdS and the CdS/Fe_3O_4 series exhibit comparable FWHM values for the CdS diffraction peaks,

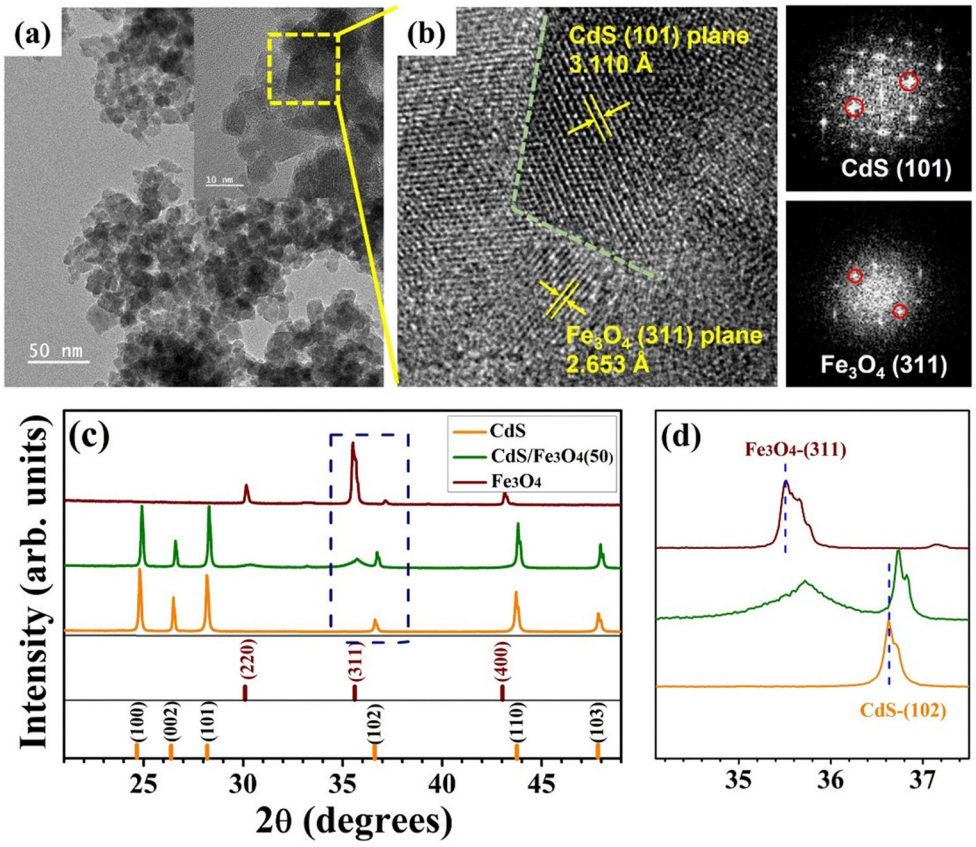


Fig. 1 (a) and (b) HRTEM images and the corresponding FFT images of the CdS/Fe_3O_4 heterostructure system; (c) and (d) XRD patterns of (c) CdS/Fe_3O_4 and (d) CdS/Fe_3O_4 (50) heterostructures at low scattering angles.

 This journal is © The Royal Society of Chemistry 2025

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

indicating that the crystallite size of CdS remains essentially unchanged upon formation of the heterostructure. A closer comparison of crystallite sizes estimated by the Scherrer equation reveals a slight increase from ~57.6 nm for bare CdS to ~64 nm for the CdS/Fe_3O_4 series as the Fe_3O_4 content increases. In contrast, for the Fe_3O_4-related diffraction peaks, the crystallite size decreases markedly from ~30.4 nm in bare Fe_3O_4 to ~11 nm in the CdS/Fe_3O_4 series, despite the clear decrease in FWHM values compared to bare Fe_3O_4. These results are summarized in Table S1. This trend suggests that while CdS crystallite size is only marginally affected, the incorporation into the heterostructure induces significant size reduction in Fe_3O_4 domains, likely due to growth suppression or structural modulation during composite formation.

In detail, after assigning the XRD patterns of the heterostructure system, the changes in the low-angle XRD patterns of the synthesized heterostructures were analyzed (Fig. 1d).

Differences in the defect structures at the interface of the heterostructure can be inferred from the patterns. The CdS/Fe_3O_4(50) heterostructure produced peaks at larger angles, which is an expected result of forming more defect structures at the interface.[30,31] These results indicate the successful formation of the CdS/Fe_3O_4 system with heterojunctions.

However, while the TEM images and XRD patterns alone do not offer direct evidence of defect structure formation, the successful defect construction of the CdS/Fe_3O_4 heterostructure was confirmed through detailed electronic structure analyses and STEM-EELS, as discussed in a later section.

Photocatalytic performance

To assess the photocatalytic performance of the series of CdS/Fe_3O_4 heterostructures, the PCR was performed by exposing a liquid suspension of two candidate molecules (HMF and toluene) to visible light irradiation (λ = 445 nm) with 230 mW cm^{-2} of illumination intensity. As shown in Fig. S4, among the CdS/Fe_3O_4(x) heterostructures, CdS/Fe_3O_4(50) exhibited the highest photocatalytic performance in the photoconversion reactions (PCRs) of both HMF and toluene.

To highlight the superior activity of the CdS/Fe_3O_4 heterostructure, Fig. 2 compares the photocatalytic conversion rates and corresponding product yields for HMF and toluene oxidation using CdS/Fe_3O_4(50) with those obtained from the individual CdS and Fe_3O_4 nanoparticles under visible light irradiation.[32,33] Under visible light irradiation, the PCR activity of HMF and toluene was enhanced when the heterostructure system was used instead of pristine CdS and Fe_3O_4 (Fig. 2a and b). As a result, CdS/Fe_3O_4(50) exhibited better PCR efficiencies than the CdS NPs, indicating that the heterostructures are sufficient to promote significant photocatalytic activity. This finding implies that the superior activity of the heterostructures underscores their crucial role in maximizing the concentration

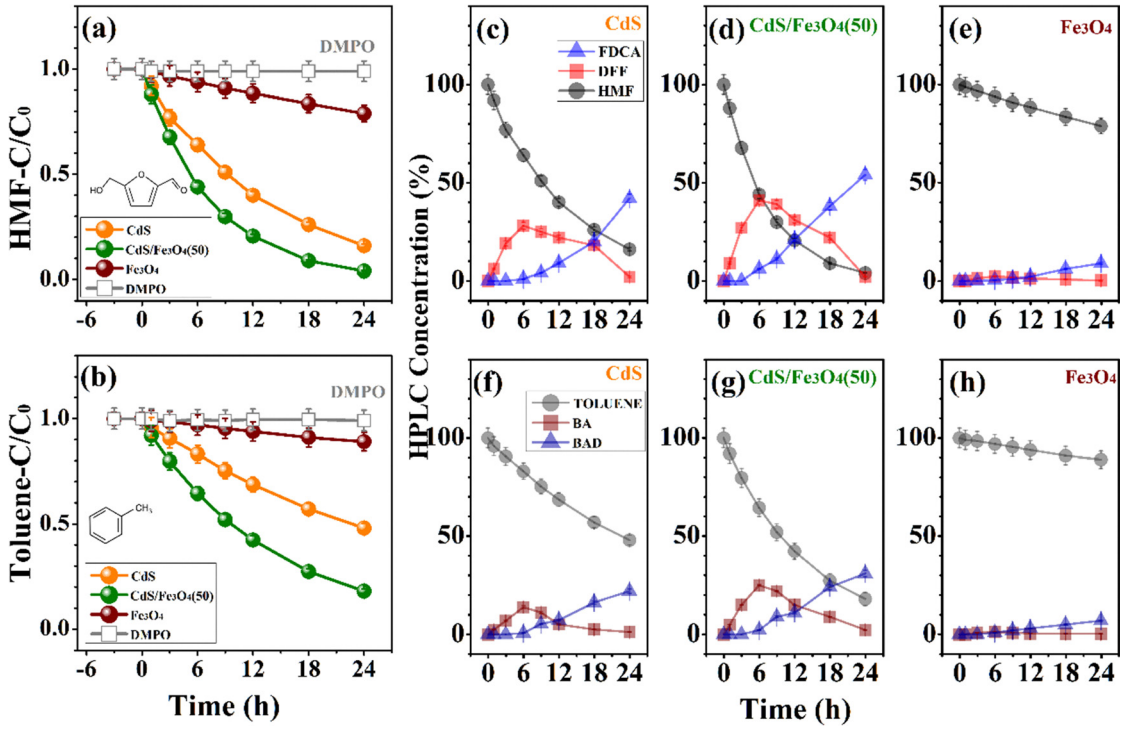


Fig. 2 Photocatalytic conversion reactions of (a) HMF and (b) toluene using the CdS/Fe_3O_4(50) heterostructure under 445 nm irradiation over 24 h. Experimental conditions: [catalyst] = 0.015 g/30 mL, $[HMF]_0$ = $[toluene]_0$ = 25 mM, and λ = 445 nm with an illumination intensity of 230 mW cm^{-2}. HPLC product analysis for the oxidation of (c)–(e) HMF and (f)–(h) toluene using the CdS/Fe_3O_4 heterostructure system. Photocatalytic conversion reactions of (a) HMF and (b) toluene using the CdS/Fe_3O_4 heterostructure under 445 nm irradiation over 24 h. Experimental conditions: [catalyst] = 0.015 g/30 mL, $[HMF]_0$ = $[toluene]_0$ = 25 mM, and λ = 445 nm with an illumination intensity of 230 mW cm^{-2}. HPLC product analysis for the oxidation of (c)–(e) HMF and (f)–(h) toluene using the CdS/Fe_3O_4 heterostructure system.

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

of V_os and optimizing the electronic properties for efficient photocatalysis under visible light irradiation.

Fig. 2c–h show the variation in the yields of FDCA and BAD, the end products of the HMF and toluene oxidation reactions, respectively, as a function of time under visible light irradiation using CdS/Fe_3O_4(50). The yields of FDCA and BAD followed the same trend as the HMF and toluene PCRs (Fig. 2c–h), with the highest yields of FDCA and BAD obtained using the heterostructure system. HMF contains alcohol (–OH) and aldehyde (–CHO) functional groups. Selectively oxidizing –OH or –CHO leads to forming 2,5-diformylfuran (DFF) or 5-hydroxymethyl-2-furancarboxylic acid (HMFCA) as intermediates. Because selectively oxidizing the –OH or –CHO groups of HMF is crucial for the synthesis of FDCA *via* photocatalysis, identifying the functional groups involved in HMF oxidation by the CdS/Fe_3O_4(50) heterostructure (Fig. 2c–e) is crucial.[34] In these catalytic systems, the –OH group of HMF was selectively oxidized to a –CHO group in all samples, resulting in the formation of 2,5-diformylfuran (DFF) as the sole intermediate, which was subsequently converted to 2,5-furandicarboxylic acid (FDCA) as the final product.[35] In the case of toluene, which contains a single functional group (–CH_3), oxidation initially produced benzyl alcohol (BA) as the primary intermediate, which was further oxidized to benzaldehyde (BAD) as the final product (Fig. 2f–h).

In conclusion, the selective oxidation of biomass-derived molecules (HMF and toluene) using a CdS/Fe_3O_4 heterostructure system dispersed in aqueous solution confirmed that under all conditions, the reaction proceeds from HMF *via* DFF to the final product FDCA and from toluene *via* BA to BAD.[36,37] Table 1 lists the 24-hour PCD profiles of HMF, toluene, and the final products (FDCA and BAD) formed using the CdS/Fe_3O_4(50) heterostructure under 445 nm irradiation.

The proposed mechanism for the photocatalytic oxidation of HMF to FDCA and of toluene to benzaldehyde using the heterostructure system can be described as follows (Scheme S1). The defect structure formed at the interface of the heterostructure system can trap electrons and O_2 to promote the formation of $^{\bullet}O_2^-$. As a result, HMF (or toluene) is converted to FDCA (or BAD) *via* photocatalytic oxidation under the action of $^{\bullet}O_2^-$ and h^+. To confirm the involvement of reactive oxygen species such as $^{\bullet}O_2^-$ in the photocatalytic system, we used 5,5-dimethyl-1-pyrroline *N*-oxide (DMPO) as an $^{\bullet}O_2^-$ radical scavenger, which completely inhibited the photocatalytic activity (Fig. 2a and b).[38] Furthermore, to verify the involvement of active O_2 species such as $^{\bullet}O_2^-$ and $^1O_2$ in the CdS/Fe_3O_4(50) photocatalytic system, electron spin resonance spectroscopy was performed using DMPO and 2,2,6,6-tetramethylpiperidine (TMP) as trapping agents to detect $^{\bullet}O_2^-$ and $^1O_2$ (Fig. S5).[39] The signal corresponding to $^{\bullet}O_2^-$ is evident, indicating the contribution of the CdS/Fe_3O_4(50) heterostructure to $^{\bullet}O_2^-$ formation (Fig. S5a). Similarly, signals corresponding to $^1O_2$ were observed with slight enhancement under visible light irradiation, highlighting the favourable effect of CdS/Fe_3O_4(50) on $^1O_2$ generation. (Fig. S5b). Notably, no $^{\bullet}O_2^-$ or $^1O_2$ signals were detected under dark conditions, confirming their origin from the photogenerated carriers of the CdS/Fe_3O_4(50) heterostructure. Consequently, the rapid *in situ* generation of abundant $^{\bullet}O_2^-$ and $^1O_2$ represents a powerful strategy for the photocatalytic targeted oxidation of HMF to FDCA and toluene to BAD.

A linear correlation between $\ln(C/C_0)$ and the reaction time was obtained for the heterostructure system (CdS/Fe_3O_4(50)), and the PCR processes were fitted to first-order kinetics to determine the first-order rate constants, k, for the photocatalytic reaction of each heterostructure system (Fig. S6). As shown in Fig. S6a, the k values for the PCR of HMF using CdS (orange), CdS/Fe_3O_4(50) (green), and Fe_3O_4 (dark red) were calculated to be 0.076, 0.134, and 0.0098 h^{-1}, respectively. The CdS/Fe_3O_4(50) heterostructure accelerated the PCR rate of HMF by 76.4% compared with that of the same reaction using CdS. As shown in Fig. S6b, the k values for the PCR of toluene using CdS (orange), CdS/Fe_3O_4(50) (green), and Fe_3O_4 (dark red) were calculated to be 0.031, 0.071, and 0.0049 h^{-1}, respectively. The CdS/Fe_3O_4(50) accelerated the PCR rate of toluene by 129% compared with that of the same reaction using CdS. Additionally, the photocatalytic activities of the three tested materials were maintained (>97%), even after 10 reuse cycles (Fig. S7).

To thoroughly address environmental concerns regarding Cd leaching and confirm the stability of the CdS/Fe_3O_4 heterostructure photocatalyst, additional stability analyses were performed. Elemental composition analysis by X-ray fluorescence (XRF) revealed no significant change in the Cd/Fe elemental ratio before and after the photocatalytic reaction, confirming negligible Cd loss during the reaction and subsequent magnetic recovery (see Table S2). Furthermore, inductively coupled plasma mass spectrometry (ICP-MS) provided quantitative insight into dissolved Cd, showing that the bare CdS photocatalyst exhibited approximately 3% Cd leaching relative to its total Cd content. In sharp contrast, the CdS/Fe_3O_4(50) heterostructure displayed significantly reduced Cd leaching of less than 1% under identical reaction conditions. These findings clearly illustrate that integrating Fe_3O_4 nanoparticles enhances the structural integrity and environmental safety of CdS-based photocatalysts, highlighting the advantage of this heterostructure approach in terms of catalyst recovery and minimizing environmental risks.

The magnetic properties of the heterostructure systems were evaluated using VSM to determine whether they could be recovered using magnets, which was one of the objectives of this study (Fig. S8a).[40] Fe_3O_4 NPs are known to have ferrimagnetic

Table 1 Relative concentrations (C/C_0) of HMF and toluene and quantities of FDCA and BAD obtained after 24 h of HMF and toluene photocatalytic conversion under 445 nm irradiation using the CdS/Fe_3O_4 heterostructure

	Fe_3O_4	CdS/Fe_3O_4(50)	CdS
HMF (C/C_0)	0.79 ± 0.039	0.04 ± 0.002	0.16 ± 0.008
Yield of FDCA (%)	9.0 ± 0.45	54.1 ± 2.70	42.1 ± 2.10
Toluene (C/C_0)	0.89 ± 0.045	0.18 ± 0.009	0.48 ± 0.024
Yield of BAD (%)	6.9 ± 0.35	31.1 ± 1.55	21.9 ± 1.10

*C and C_0 represent the time-dependent and initial concentrations of HMF and toluene, respectively.

 This journal is © The Royal Society of Chemistry 2025

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

properties at room temperature. The magnetic properties of the Fe_3O_4 NPs and heterostructures prepared under the same conditions were measured.[41] The magnetization of the CdS/Fe_3O_4(50) system decreased by only ~6.8%, from 79.32 emu g^{-1} for the pristine Fe_3O_4 system to 73.62 emu g^{-1} for the CdS/Fe_3O_4(50) heterostructure, even though it only contained 50% CdS. This indicated that the recovery rate was not significantly affected, confirming the applicability of this system. EPR analyses were also performed to determine the differences in the number of V_os between the heterostructure system and Fe_3O_4 NPs (Fig. S8b).[42] As shown in the figure, the EPR spectra of CdS/Fe_3O_4(50) show a pronounced increase in signal intensity at g-factors of ~1.99 each, indicating a Zeeman effect induced by a single electron trapped in V_os.[43] These results confirm that the number of V_os increased in the CdS/Fe_3O_4(50) heterostructure, which is consistent with our observation of the enhanced photocatalytic properties of the CdS/Fe_3O_4(50) heterostructure.[44] Therefore, the CdS/Fe_3O_4(50) heterostructure gave rise to a higher V_os peak intensity than the Fe_3O_4 NPs, which can be explained by an increase in the number of V_os owing to forming a heterojunction with CdS at the interface. As mentioned, structural analysis and PCR experiments are still insufficient to clearly explain that the CdS/Fe_3O_4(50) heterostructure we synthesized has a clear heterostructure. To clarify it, we investigated structural changes at the interface using electron structure analysis using XPS and STEM-EELS techniques.

Electronic structure characterization

Fig. 3a–d illustrate the electronic structure changes captured in the four core-level XPS profiles of the non-irradiated CdS/Fe_3O_4 heterostructure. XPS analyses confirmed that charge carriers were transferred from CdS to Fe_3O_4 in the CdS/Fe_3O_4(50) heterostructure.[45] Therefore, the binding energy shift in this system follows a specific trend: the binding energies of CdS increase, whereas those of Fe_3O_4 decrease. As shown in Fig. 3a and b, the binding energies of the two core levels in the pristine CdS NPs (404.8 eV for Cd $3d_{5/2}$ and 161.1 for S $2p_{3/2}$) shifted to significantly higher positions in the spectra of the CdS/Fe_3O_4(50) heterostructure (405.3 eV for Cd $3d_{5/2}$ and 161.65 for S $2p_{3/2}$).[46] Conversely, Fig. 3c and d shows that the binding energies in the two core-level XPS spectra of the pristine Fe_3O_4 NPs (711.1 eV for Fe $2p_{3/2}$ and 530.2 for O 1s) shifted to lower binding energies in CdS/Fe_3O_4(50) (710.6 eV for Fe $2p_{3/2}$ and 529.7 for O 1s).[47] Thus, we confirmed that charge transfer from CdS to Fe_3O_4 occurred in this heterostructure. Interestingly, we noted a Cd_{2+x} peak (defect-induced structure) in the Cd 3d spectrum of the CdS/Fe_3O_4 heterostructure along with a peak corresponding to –S–O– in the S 2p spectrum, which is evidence of heterostructure electron transfer owing to changes in CdS (S $2p_{3/2}$ = 167.9 eV). These peaks are attributed to forming a new electronic structure at the interface between the two NPs.[48] Further evidence of interfacial changes was observed for Fe_3O_4.

As shown in Fig. 3c, the intensity ratio of Fe^{2+} to Fe^{3+} in the Fe $2p_{3/2}$ peak of CdS/Fe_3O_4(50) was remarkable. For pristine

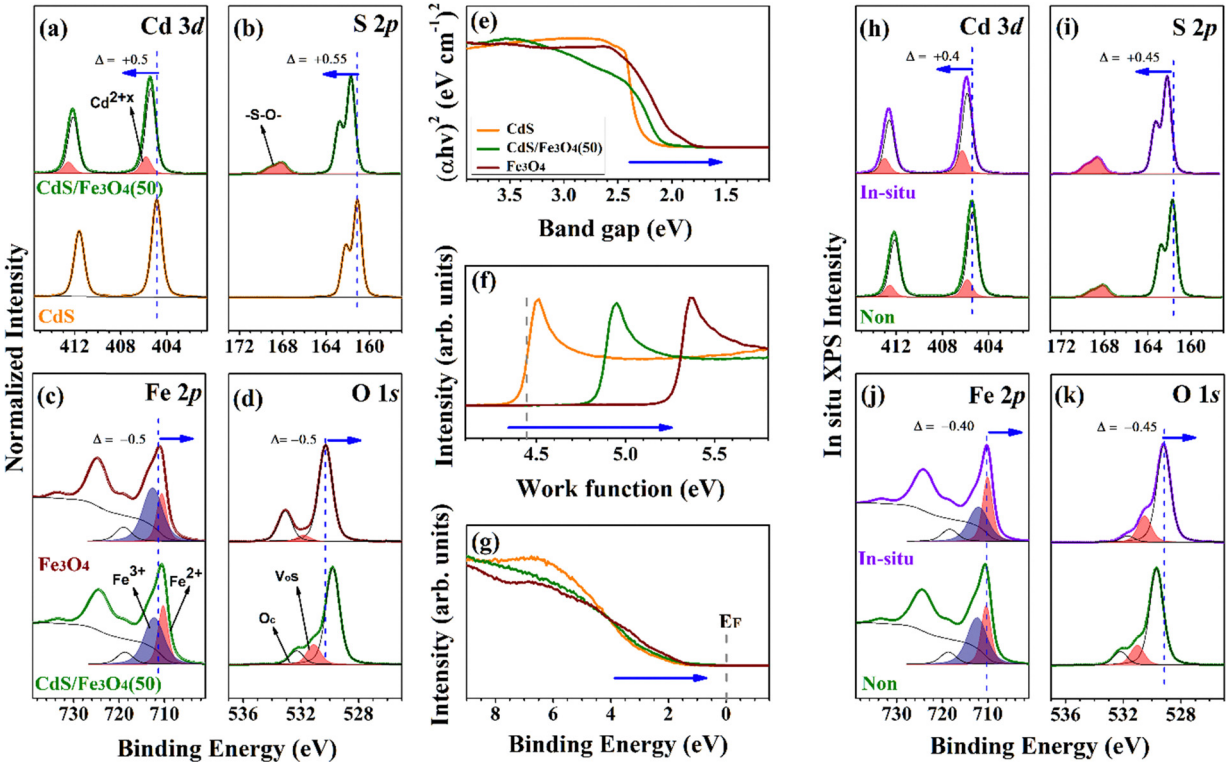


Fig. 3 (a) Cd 3d, (b) S 2p, (c) Fe 2p, and (d) O 1s XPS spectra for the CdS, Fe_3O_4 and CdS/Fe_3O_4(50). (e) UV-vis diffuse reflectance spectra and (f) work functions, and (g) valence band spectra of CdS, Fe_3O_4 and CdS/Fe_3O_4(50) (orange – CdS; green – CdS/Fe_3O_4; dark red – Fe_3O_4). (h)–(k) *In situ* XPS spectra of CdS/Fe_3O_4(50) acquired under 445 nm irradiation: (h) Cd 3d, (i) S 2p, (j) Fe 2p, and (k) O 1s (green – non-CdS/Fe_3O_4; violet – *in situ* CdS/Fe_3O_4).

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

Fe_3O_4, the relative intensity of the Fe^{2+}-induced peak is lower than that of the Fe^{3+}-induced peak. However, the structure of Fe_3O_4 in CdS/Fe_3O_4(50) did not change, but the relative intensity of the Fe^{2+} peak increased owing to an increase in the surface defect structure formed by changes in the electronic structure at the interface. As shown in Fig. 3d, the decomposed O 1s XPS spectra of both samples include three components: Fe–O (529.7 eV), an oxygen vacancy peak (V_os) (531.7 eV), and a chemisorbed adsorption peak –O_c (532.5 eV). The CdS/Fe_3O_4(50) heterostructure gave rise to a higher V_os peak intensity than the Fe_3O_4 NPs, which can be explained by an increase in the number of V_os owing to the formation of a heterojunction with CdS at the interface.[49] Table 2 lists the changes in the binding energies obtained *via* XPS analysis of the heterostructure system.

Pristine CdS NPs exhibit strong light absorption in the visible wavelength range owing to their narrow bandgaps.[50] The bandgap of the synthesized CdS NPs, as shown in Fig. 3e, was determined to be 2.30 eV from Tauc plots obtained from the DRS results. The bandgap of the Fe_3O_4 NPs, as shown in Fig. 3e, was measured to be 1.8 eV.[51] The optical properties and bandgap energies of the heterostructured photocatalysts were confirmed by UV-vis DRS. The bandgap of CdS/Fe_3O_4(50) was found to be 2.1 eV, which lies between those of the constituent NPs of the heterostructure, and the bandgap remained within the visible wavelength range, indicating that the CdS/Fe_3O_4(50) heterostructure exhibited visible light photocatalytic activity. The work function measurements of the heterostructure (Fig. 3f) confirm that charge was transferred between CdS and Fe_3O_4.[52] In the CdS/Fe_3O_4(50) heterostructure, the work functions of CdS and Fe_3O_4 were measured to be 4.45 and 5.30 eV, respectively, indicating that the relatively small work function of CdS compared with that of Fe_3O_4 allows charge carriers to move smoothly from CdS to Fe_3O_4.[53,54] Fig. 3g shows valence band spectra to measure the Fermi level (EF) of the three samples. Each EF of the three samples (CdS, CdS/Fe_3O_4(50), and Fe_3O_4) is measured to be 1.75, 1.52, and 1.40 eV, which lies between those of the constituent NPs of the heterostructure.

The photocatalytic properties of the heterogeneous system were predicted to improve owing to the increased number of defect structures formed at the interface. In other words, the electrons and O_2 trapped in these defect structures were converted into $^{\bullet}O_2^-$ and h^+ under visible light irradiation. To gain a deeper understanding of this phenomenon, observing the light-induced changes in the electronic structures of the heterostructure system is necessary. Accordingly, in this study, *in situ* XPS was used to determine the changes occurring during photocatalysis. Fig. 3h–k shows the *in situ* XPS results for the CdS/Fe_3O_4(50) heterostructure. As expected, the electronic structure changed upon visible light irradiation: under 445 nm irradiation, the CdS peaks shifted towards higher binding energies, and the Fe_3O_4 peaks shifted towards lower binding energies. These observations confirm the light-induced photocatalytic properties of the heterostructures. In particular, a more significant change is observed for the CdS/Fe_3O_4(50) heterostructure, which is consistent with its superior photocatalytic properties. We also observed an increase in the intensity of the –S–O– component of the S 2p peak, indicating charge transfer at the interface, as well as an increase in the intensity of the V_os component of the O 1s peak. Thus, the *in situ* XPS results clearly explain the photocatalytic activity of the heterostructure and reveal the underlying reason for the superior visible-light photocatalytic activity of this system compared with that of CdS. The increase in defect structures at the interface between the two materials in the heterostructure significantly enhanced the photocatalytic properties.

Table 2 Cd $3d_{5/2}$, S $2p_{3/2}$, Fe $2p_{3/2}$, and O 1s binding energies in CdS, CdS/Fe_3O_4(50) and Fe_3O_4

Core level	CdS (eV)	CdS/Fe_3O_4(50) (eV)	Fe_3O_4 (eV)
Cd $3d_{5/2}$	404.8	405.3	—
S $2p_{3/2}$	161.1	161.65	—
Fe $2p_{3/2}$	—	710.6	711.1
O 1s	—	529.7	530.2

Surface structure characterization

To confirm the change in the electronic states at the interface of the CdS/Fe_3O_4(50) heterostructure, we performed nanoscale STEM-EELS SI analysis, as shown in Fig. 4.[55,56] We selected an interfacial region between the CdS and Fe_3O_4 NPs for spectral analysis of the heterostructure (highlighted by the yellow box in Fig. 4a). Fig. S9 shows the EDX mapping results of the four elements comprising the heterostructure interface. The elemental distribution of the synthesized CdS/Fe_3O_4(50) heterostructure revealed that Cd, S, Fe, and O were distributed in the interfacial region of the heterostructure. The ABF STEM image of the magnified area in the yellow box in Fig. 4a shows that the heterostructures of Fe_3O_4 and CdS are in close contact (Fig. 4b). Therefore, changes in the electronic structure can be observed at the interface between the two NPs, as confirmed in this study. FFT was used to verify the lattice structures of the two NP types in the heterostructure (*i.e.*, Fe_3O_4 and CdS), as shown in Fig. 4c and d. No significant structural changes are apparent, confirming the crystallographic structure of each phase. Therefore, we can assume that the improvement in the photocatalytic properties of the CdS/Fe_3O_4 heterojunction is due to the enhancement of the interfacial defect structures and the change in the interfacial electronic structure. Additionally, Fig. S10 shows the Fe L-edge and O K-edge spectra in the three regions, distinguishing between the interface of the heterogeneous structure, the CdS region, and the Fe_3O_4 region. The two spectra clearly show a reduction in the peak at the interface of the CdS/Fe_3O_4 heterostructure and its disappearance in the CdS region, allowing us to observe changes in the electronic structure at the interface.[57]

The multiple linear least squares (MLLS) fitting method, using the selected EELS components as references, was used to map the local distributions of different electronic and charge states within the EELS SI dataset.[58,59] This approach, which assumes that the total spectrum can be represented as a linear

 This journal is © The Royal Society of Chemistry 2025

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

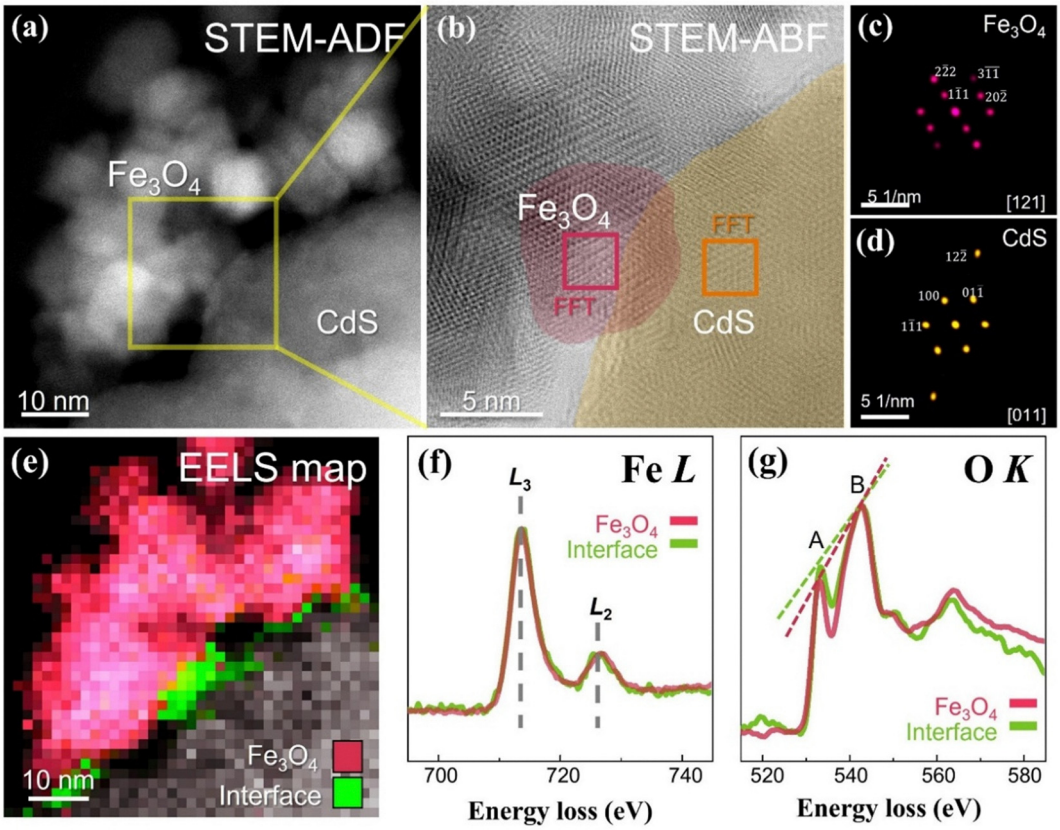
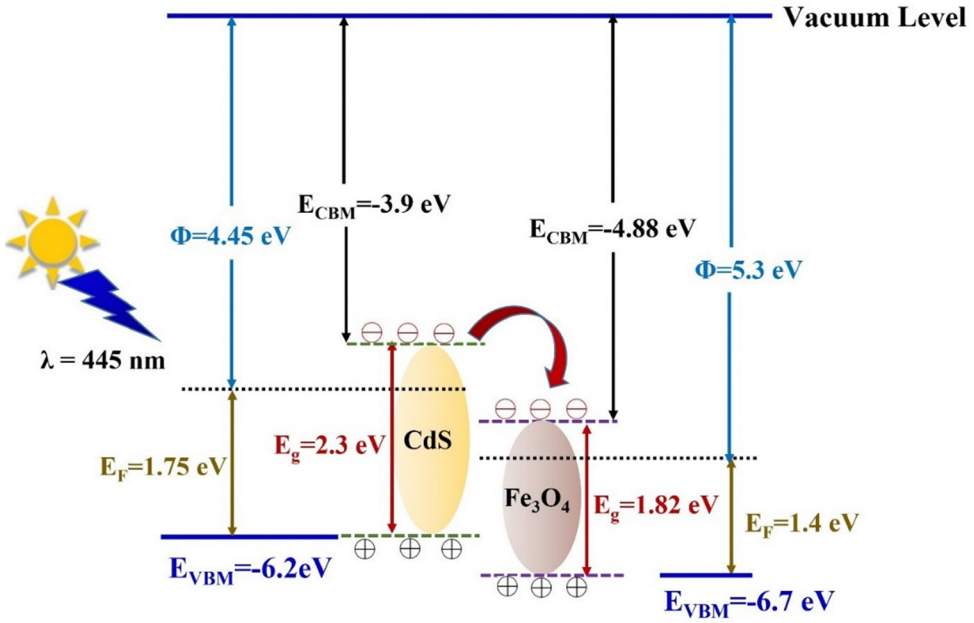


Fig. 4 Structural and chemical characterization of CdS/Fe_3O_4 (50). (a) Annular dark-field (ADF) STEM image and (b) annular bright-field (ABF) STEM image of the CdS/Fe_3O_4 heterostructure, with magnification of the region marked by the yellow box in (a). (c) and (d) FFT patterns extracted from the Fe_3O_4 (magenta) and CdS (orange) regions in (b). (e) EELS map distinguishing Fe_3O_4 (magenta), the interface region (green), and CdS (grey) based on chemical distribution differences. Energy loss near-edge structure (ELNES) of (f) Fe L-edge and (g) O K-edge spectra.

combination of the local component spectra, is traditionally used to distinguish electronically distinct phases at heterostructure interfaces. Fig. 4e shows the chemical phase mapping results for the EELS SI data in the Fe_3O_4 region at the heterostructure interface shown in Fig. 4b. A heterostructure (green) is formed at the interface between Fe_3O_4 (magenta) and CdS (grey), which is expected to have a different electronic structure. Additionally, the Fe L-edge (Fig. 4f) and O K-edge (Fig. 4g) profiles were obtained from the EELS maps. The Fe L-edge spectra of Fe_3O_4 and CdS/Fe_3O_4 were almost identical (Fig. 4f). This indicates that the change in the photocatalytic properties was not due to a change in the Fe_3O_4 structure itself, which is consistent with the subtle change in the two charge states (Fe^{2+} and Fe^{3+}) in the Fe 2p spectrum (Fig. 3c). By contrast, in the case of the O K-edge spectrum shown in Fig. 4g, although the overall shape change is similar, the intensities of the A and B peaks differ, which can explain the change in the defect structure. In other words, the increase in peak A can be explained by the increase in defect structures, and the A/B value (0.74) of the heterostructure increased more than the A/B value (0.68) of Fe_3O_4.[60,61] This result explains the increase in the number of defect structures at the interface. This result agrees well with the change in the oxygen defect structures in the O 1s spectrum shown in Fig. 3d. As a result, the two spectra obtained from the STEM-EELS mapping can explain the changes at the interface. It is worth noting that a small shift of the A peak towards a higher energy loss by $\sim$0.25 eV was observed. This peak shift may be linked to the probable presence of hydroxylated metal at the interface.[62] However, careful caution should be taken for interpreting the observed small peak shift because dehydroxylation progresses fast under electron beam irradiation under high-vacuum conditions of STEM.

Finally, we examined whether simple physically mixed samples of the two NPs could form heterostructures, in contrast to the methodology described in this work, in which the CdS/Fe_3O_4 heterostructure was synthesized *via* base treatment. The physically mixed sample, having the same composition as CdS/Fe_3O_4(50), was prepared by grinding CdS and Fe_3O_4 NPs for 10 min in an agate mortar. This comparison allowed us to determine whether the observed hetero-contact originated from the intrinsic nature of the sample or from the base-treatment process, thereby providing direct evidence for our synthesis methodology. As expected, the XPS and PCR experimental results for the physically mixed CdS–Fe_3O_4 sample, shown in Fig. S11 and S12, revealed that the photocatalytic properties varied depending on the amount of CdS. Furthermore, the XPS analysis confirmed that no chemical bonds formed between the

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.



Scheme 1 Illustration of the type-II mechanism for the CdS/Fe_3O_4 heterostructure system.

two components, indicating only mechanical mixing; the CdS and Fe_3O_4NPs remained independent and exhibited nearly identical electronic structures to their pristine counterparts.

Herein, we propose a type-II mechanism for the photooxidation of HMF and toluene using the CdS/Fe_3O_4 heterojunctions, as shown in Scheme 1. The movement of electrons in heterojunction photocatalysts can also be described in terms of the work function of the individual semiconductors. Free electrons migrate from a semiconductor with a smaller work function to one with a larger work function, thereby storing electron density in the latter. Because the work function of Fe_3O_4 is larger than that of CdS (5.3 *vs.* 4.45 eV, respectively), free electrons move from the positively charged CdS to the negatively charged Fe_3O_4. This makes CdS highly electropositive and Fe_3O_4 highly electronegative. Consequently, photogenerated electrons and holes can exist under conditions that allow visible light activity. When visible light irradiates the surface of the heterostructure, electrons are excited to the conduction band (CB), and holes remain in the valence band (VB) of the heterostructure. In other words, the recombination of electron–hole pairs is strongly suppressed owing to the efficient separation of the charge carriers.[63] Therefore, the photogenerated electrons and holes were efficiently separated under visible light irradiation. Photogenerated electrons accumulate in the CB of Fe_3O_4, and photogenerated holes accumulate in the VB of CdS under visible light irradiation. To confirm the reaction mechanism, the migration directions of the photogenerated electrons and holes were investigated. The photogenerated electrons were found to initially migrate from the highly negative CB of CdS to the low-energy CB of Fe_3O_4, whereas the holes migrated from the VB of Fe_3O_4 to the VB of CdS. This results in excess electrons in the Fe_3O_4 CB and excess holes in the CdS VB, that is, effective separation of the charge carriers.

Conclusions

In this study, CdS and Fe_3O_4 NPs were synthesized independently, and CdS/Fe_3O_4 heterostructures were prepared as photocatalysts under basic conditions. The results showed that after visible-light irradiation, the heterostructure system exhibited a significant increase in the HMF and toluene oxidation reaction yields by CdS. The CdS/Fe_3O_4 heterostructure, which follows a type-II mechanism, exhibited better PCR efficiency owing to its improved defect structure. A breakthrough was also achieved in terms of the recovery rate of these samples; the magnetic properties of the CdS/Fe_3O_4 heterostructure, containing ferrimagnetic Fe_3O_4, did not significantly degrade compared to those of Fe_3O_4 and exhibited a 100% recovery rate when using a magnet. Thus, the environmental problems that may arise from using photocatalytic NPs can be addressed. A CdS/Fe_3O_4 heterojunction structure with significantly improved photocatalytic properties was developed owing to an increase in the defect structure at the interface between the two materials. Moreover, owing to its magnetic properties, this heterosystem provides a solution for recovering NP catalysts. Its photocatalytic properties are comparable to those of other CdS-based heterojunctions reported in the literature, making the heterostructure highly promising for future applications.

Author contributions

Sunyoung Hwang: conceptualization, investigation and formal analysis; H. Yu: investigation and formal analysis; Hyeon-Ah Ju: investigation and formal analysis; Heewon Kim: formal analysis; Jaeyoon Baik: formal analysis for high-resolution XPS with synchrotron radiation; Young-Min Kim: formal analysis and writing – original draft; H. S. Kim: supervision, funding

 This journal is © The Royal Society of Chemistry 2025

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

acquisition, and writing – review & editing. Hangil Lee: conceptualization, supervision, funding acquisition, and writing – review & editing. All authors contributed to discussion and manuscript preparation.

Conflicts of interest

The authors declare no competing financial interest.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

All data supporting this study are available in the SI. The SI includes experimental procedures for CdS, Fe_3O_4 and CdS/Fe_3O_4 synthesis; structural and spectroscopic characterisation (HRTEM, SEM, EDX, XRD, XRF, EELS/ELNES, elemental mapping, XPS, and EPR); photocatalytic performance data (kinetics, cycling stability, trapping experiments, product analysis); magnetic property measurements (VSM); and proposed reaction pathways. See DOI: **https://doi.org/10.1039/d5mh01147g**

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean government (MSIP) (NRF-2023R1A2C1007484) and core research institute (CRI) program, and the basic science research program through the NRF under program number (2022R1A6A1A03051158). Y.-M. Kim acknowledges the support from the National Research Foundation of Korea (NRF) through a grant funded by the Korean government (MSIT) (RS-2023-NR076943). H. Lee acknowledges the National Research Foundation of Korea (NRF) grant, funded by the Korean government (MSIT) (RS-2024-00346153).

Notes and references

1 X. Liu, M. Sayed, C. Bie, B. Cheng, B. Hu, J. Yu and L. Zhang, *J. Materiomics*, 2021, **7**, 419.
2 K. Yang, Z. Yang, C. Zhang, Y. Gu, J. Wei, Z. Li, C. Ma, X. Yang, K. Song, Y. Li, Q. Fang and J. Zhou, *Chem. Eng. J.*, 2021, **418**, 129344.
3 J. Zhang, G. Yang, B. He, B. Cheng, Y. Li, G. Liang and L. Wang, *Chin. J. Catal.*, 2022, **43**, 2530.
4 N. Venkatesh, K. Sabarish, G. Murugadoss, R. Thangamuthu and P. Sakthivel, *Environ. Sci. Pollut. Res.*, 2020, **27**, 43212.
5 L. Wang, H. Wei, Y. Fan, X. Gu and J. Zhan, *J. Phys. Chem. C*, 2009, **113**, 14119.
6 H. Yu, D. T. Hoang, H. S. Kim and H. Lee, *J. Mater. Chem. A*, 2025, **13**, 6049.
7 H.-A. Ju, D. T. Hoang, W.-S. Jang, Y.-H. Kim, E.-B. Park, S.-H. Yang, K. Ihm, Y.-M. Kim and H. Lee, *Appl. Surf. Sci.*, 2025, **679**, 161197.
8 S. Pal, A. Kumar, S. Kumar, A. K. De, R. Prakash and I. Sinha, *Catal. Lett.*, 2022, **152**, 3259.
9 L. Shi, X. Wang, Y. Hu and Y. He, *Sol. Energy*, 2020, **196**, 505.
10 X. Liu, Z. Fang, X. Zhang, W. Zhang, X. Wei and B. Geng, *Cryst. Growth Des.*, 2009, **9**, 197.
11 A. Samadi-Maybodi, M. R. Shariati and A. H. Colagar, *ChemPlusChem*, 2018, **83**, 769.
12 R. K. Chava, N. Son and M. Kang, *Appl. Surf. Sci.*, 2021, **556**, 149731.
13 Z. Khan, T. R. Chetia, A. K. Vardhaman, D. Barpuzary, C. V. Sastri and M. Qureshi, *RSC Adv.*, 2012, **2**, 12122.
14 C. Shi, Y. An, G. Gao, J. Xue, H. Algadi, Z. Huang and Z. Guo, *ACS Sustainable Chem. Eng.*, 2024, **12**, 2538.
15 C. Li, J. Li, L. Qin, P. Yang and D. G. Vlachos, *ACS Catal.*, 2021, **11**, 11336.
16 L. Sun, G. Li, S. Wan and T. An, *Chemosphere*, 2010, **78**, 313.
17 D. J. Aranha and P. R. Gogate, *Ind. Eng. Chem. Res.*, 2023, **62**, 3053.
18 L. Wang, H. Wei, Y. Fan, X. Gu and J. Zhan, *J. Phys. Chem. C*, 2009, **113**, 14119.
19 Z. Wu, H. Yu, L. Kuai, H. Wang, T. Pei and B. Geng, *J. Colloid Interface Sci.*, 2014, **426**, 83.
20 V. N. Pham, S. Lee, H. Lee and H. S. Kim, *Inorg. Chem.*, 2023, **62**, 3703.
21 L. Wang, H. Wei, Y. Fan, X. Gu and J. Zhan, *J. Phys. Chem. C*, 2009, **113**, 14119.
22 Y. Zhu, J. Chen, L. Shao, X. Xia, Y. Liu and L. Wang, *Appl. Catal., B*, 2020, **268**, 118744.
23 Y. Cheng, Y. Wang, F. Bao and D. Chen, *J. Phys. Chem. B*, 2006, **110**, 9448.
24 X. Bian, K. Hong, L. Liu and M. Xu, *Appl. Surf. Sci.*, 2013, **280**, 349.
25 A. Khalid, R. M. Ahmed, M. Taha and T. S. Soliman, *J. Alloys Compd.*, 2023, **947**, 169639.
26 X. Jin, Y. Fang, T. Salim, M. Feng, Z. Yuan, S. Hadke and L. H. Wong, *Adv. Mater.*, 2021, **33**, 2104346.
27 A. Phuruangrat, T. Thongtem and S. Thongtem, *Mater. Lett.*, 2012, **80**, 114.
28 K. Zhang, Z. Mou, S. Cao, S. Wu, X. Xu and C. Li, *Int. J. Hydrogen Energy*, 2022, **47**, 12605.
29 M. A. Sabur and M. A. Gafur, *J. Nanomater.*, 2024, 9577778.
30 S. Muruganandam, G. Anbalagan and G. Murugadoss, *Appl. Nanosci.*, 2015, **5**, 245.
31 L. Saravanan, A. Pandurangan and R. Jayavel, *J. Nanopart. Res.*, 2011, **13**, 1621.
32 V. N. Pham, H. Jeon, S. Hong and H. Lee, *Inorg. Chem.*, 2022, **61**, 16887.
33 J. Cui, K. K. Niu, R. Z. Zhang, H. Liu, S. Yu and L. B. Xing, *Chem. Commun.*, 2024, **60**, 4310.
34 W. Zhang, H. Qian, Q. Hou and M. Ju, *Green Chem.*, 2023, **25**, 893.
35 W. Wu, G. Zhang, J. Zhang, G. Wang, C. H. Tung and Y. Wang, *Chem. Eng. J.*, 2021, **404**, 126433.
36 Z. Zhang and K. Deng, *ACS Catal.*, 2015, **5**, 6529.
37 R. Li, H. Kobayashi, J. Guo and J. Fan, *J. Phys. Chem. C*, 2011, **115**, 23408.

Published on 16 August 2025. Downloaded by Sookmyung Women's University on 11/28/2025 2:24:59 AM.

38 H. Jeon, D. T. Hoang, J. Baik, S. Hong and H. Lee, *Inorg. Chem.*, 2024, **63**, 12370.
39 D. Sasikumar, Y. Takano, H. Zhao, R. Kohara, M. Hamada, Y. Kobori and V. Biju, *Sci. Rep.*, 2022, **12**, 11371.
40 G. Xu, M. Du, T. Li, Y. Guan and C. Guo, *Sep. Purif. Technol.*, 2021, **275**, 119157.
41 F. Soltani-Nezhad, A. Saljooqi, A. Mostafavi and T. Shamspur, *Ecotoxicol. Environ. Saf.*, 2020, **189**, 109886.
42 H. Zhang, S. Gan, D. Hou, X. Q. Qiao, R. Chi and D. S. Li, *Catal. Sci. Technol.*, 2023, **13**, 1196.
43 J. He, Y. Mei and W. C. Zheng, *Optik*, 2019, **194**, 163087.
44 S. Omar, M. Omar, N. F. Attia, G. M. El-Subruiti and A. Eltaweil, *Surf. Interfaces*, 2024, **45**, 103817.
45 Z. Zheng, F. Han, B. Xing, X. Han and B. Li, *J. Colloid Interface Sci.*, 2022, **624**, 460.
46 Y. Liu, L. Zhou, Y. Hu, C. Guo, H. Qian, F. Zhang and X. W. D. Lou, *J. Mater. Chem.*, 2011, **21**, 18359.
47 T. Li, M. Wang and Y. Hao, *Sci. Total Environ.*, 2023, **857**, 159254.
48 Y. Wang, L. Zhao, X. Cai, Y. Chen, J. Xu, L. Zhang, X. Lin, X. Dai, Y. Jiang, J. Liao and Y. Zhang, *Environ. Sci. Pollut. Res.*, 2024, **31**, 15091.
49 A. Aguinaco, J. M. Mánuel, E. Blanco, M. Domínguez, R. Litrán, J. J. Delgado and M. Ramírez-del-Solar, *Materials*, 2022, **15**, 6718.
50 Q. Li, B. Guo, J. Yu, J. Ran, B. Zhang, H. Yan and J. R. Gong, *J. Am. Chem. Soc.*, 2011, **133**, 10878.
51 A. Raza, H. Shen, A. A. Haidry and S. Cui, *Appl. Surf. Sci.*, 2019, **488**, 887.
52 X. Wei, T. Xie, L. Peng, W. Fu, J. Chen, Q. Gao, G. Hong and D. Wang, *J. Phys. Chem. C*, 2011, **115**, 8637.
53 Z. He, Y. Fang, X. Wang and H. Pang, *Synth. Met.*, 2011, **161**, 420.
54 L. Wang, B. Cheng, L. Zhang and J. Yu, *Small*, 2021, **17**, 2103447.
55 R. Luo, M. Gao, C. Wang, J. Zhu, R. Guzman and W. Zhou, *Adv. Funct. Mater.*, 2024, **34**, 2307625.
56 M. de la Mata and S. I. Molina, *Nanomaterials*, 2022, **12**, 337.
57 S. Sun, X. Li, C. Zhang, X. Wang, J. Wang, C. Wang and Y. Bai, *Adv. Mater.*, 2024, **36**, 2405876.
58 B. Sun, Z. Liang, Y. Qian, X. Xu, Y. Han and J. Tian, *ACS Appl. Mater. Interfaces*, 2020, **12**, 7257.
59 W. S. Jang, Y. Jin, Y. H. Kim, S. H. Yang, S. J. Kim, J. A. Hong, J. Lee, H. Lee and Y. M. Kim, *Appl. Catal., B*, 2022, **305**, 121083.
60 M. H. Aguirre, A. Shkabko and A. Weidenkaff, *Cryst. Growth Des.*, 2010, **10**, 3562.
61 H. S. Kim, Y. J. Kim, Y. R. Son, V. N. Pham, K. J. Kim, C. W. Kim and H. Lee, *Sci. Rep.*, 2022, **12**, 11295.
62 N. Senamart, K. Deekamwong, J. Wittayakun, S. Prayoonpokarach, N. Chanlek, Y. Poo-arporn, S. Wannapaiboon, P. Kidkhunthod and S. Loiha, *RSC Adv.*, 2022, **12**, 25578–25586.
63 A. Samadi-Maybodi, M. R. Shariati and A. H. Colagar, *ChemPlusChem*, 2018, **83**, 769.

 This journal is © The Royal Society of Chemistry 2025

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

Nanoscale

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Nanoscale*, 2025, **17**, 25783

Defect engineering in OH-functionalized carbon nanotubes for visible-light-driven ROS-mediated anticancer therapy and high-performance photocatalysis

Hyungbin Park,†[a,b] Sunyoung Hwang,†[c] Joowon Choi,[d] Kyungtae Kang,*[d] Hangil Lee *[c] and Seungwoo Hong *[a,b]

Carbon nanotube (CNT)-based therapies are emerging as powerful tools in oncology due to their ability to selectively target cancer cells while minimizing damage to healthy tissues. Leveraging the critical role of reactive oxygen species (ROS) in cancer progression, this study explores ROS-mediated photodynamic strategies to enhance therapeutic efficacy. We present a systematic design of five CNT derivatives—CNT-OH, Ti-doped CNT-OH, $NaBH_4$-treated Ti-doped CNT-OH, Cr-doped CNT-OH, and $NaBH_4$-treated Cr-doped CNT-OH—engineered to improve visible light absorption and photocatalytic activity. Our approach introduces two key innovations: transition metal doping to create oxygen vacancies that boost light responsiveness and $NaBH_4$ treatment to induce structural defects and elevate surface charge. These modifications not only fine-tune the physicochemical properties of CNTs but also modulate their biological interactions. Ti-doped CNTs exhibit efficient internalization into renal cancer cells and trigger apoptosis under visible light, offering a controlled and targeted therapeutic mechanism. In contrast, Cr-doped CNTs induce necrosis, likely due to inherent cytotoxicity, highlighting the importance of material composition in cell fate decisions. Furthermore, we demonstrate the dual applicability of these CNTs in sustainable photocatalysis. $NaBH_4$-treated Cr-doped CNT-OH shows superior photocatalytic conversion of furfuraldehyde, 5-hydroxymethylfurfural, and toluene, attributed to its optimized surface properties. This work underscores the potential of defect-engineered CNTs as multifunctional platforms for both advanced cancer therapy and sustainable chemical transformation.

Received 23rd September 2025,
Accepted 11th October 2025

DOI: 10.1039/d5nr04008f

rsc.li/nanoscale

Introduction

Cancer manifests through well-established hallmarks, including accumulation of dysregulated metabolites and aberrant cell proliferation.[1–4] Due to this lethal ailment—one of the leading causes of death worldwide—human health has faced tremendous challenges over recent decades, marked by aggressiveness and substantial barriers to effective diagnosis and treatments.[5,6] In addition, invasive anticancer strategies, including surgical resection of tumors, radiotherapy, and chemotherapy with a combination of drugs, often trigger adverse reactions and show limited efficacy in advanced stages of cancer progression, highlighting the critical need to address therapeutic efficacy, drug resistance and off-target toxicity.[7–9]

Among numerous state-of-the-art emerging anticancer therapies, carbon nanotube (CNT)-based cancer therapy shows groundbreaking potential for improved prognostic outcomes due to its capabilities of transport, near-infrared (NIR) light absorption and selective impact on cancerous tissues with minimal damage to healthy cells.[10,11] The landscape of cancer treatment is undergoing transformation with the help of unique cylindrical nanostructures and functional properties, including mechanical, electrical, and thermal features.[12] For instance, the useful conversion of external strong absorption light in the NIR region into local thermal energy generates sufficient temperature to cause local hyperthermia (a temperature greater than 42 °C), which induces the death of cancer cells.[13–15] Despite extensive investigations on biomedical applications of CNTs in photothermal therapy, this strategy still needs to thoroughly mitigate several adverse effects including its inherent size, structure, shape related cytotoxicity, stability under physiological conditions, cellular uptake, and biodistribution.[16–18]

[a]*Department of Chemistry & Nanoscience, Ewha Womans University, Seoul 03760, Korea. E-mail: hsw@ewha.ac.kr*
[b]*Graduate Program in Innovative Biomaterials Convergence, Ewha Womans University, Seoul 03760, Korea*
[c]*Department of Chemistry, Sookmyung Women's University, Seoul 04310, Republic of Korea. E-mail: easyscan@sookmyung.ac.kr*
[d]*Department of Applied Chemistry, Kyung Hee University, Yongin, Gyeonggi 17104, South Korea*
† These authors contributed equally to this work.

This journal is © The Royal Society of Chemistry 2025

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

Given the critical role of intracellular reactive oxygen species (ROS) concentration in shaping heterogeneous features of cancer cells, the increasing development of cancer therapy – specifically chemodynamic therapy – leans heavily on ROS-mediated cancer treatment mechanisms.[19–21] In particular, cellular redox homeostasis, which denotes the delicate equilibrium between oxidizing and reducing agents, is partially regulated by metalloenzymes such as superoxide dismutase, catalase, and peroxidase.[22,23] However, considering the redox balance from a different perspective, cells exhibit pronounced sensitivity to ROS, encompassing (su)peroxide anions and hydroxyl radical species. Subtle fluctuations in intracellular ROS levels may significantly influence cellular fate – whether it is cell genesis or cell death.[24–26] It is quite clear that extensive endeavors have focused on modulating the redox equilibrium within the tumor microenvironment, thereby inducing oxidative stress as a means to exert anticancer effects. In this context, we fine-tuned OH-functionalized CNTs *via* transition metal doping to enhance visible-light-driven photocatalytic activity through heterojunction-induced charge separation.[27,28] However, a doping level above 5 wt% could generate excessive ROS, leading to cytotoxicity and environmental hazards. A rational strategy is required to mitigate the potential toxicity of transition metal-doped CNTs while preserving visible-light responsiveness.

Herein, we prepared five CNT variants—CNT-OH, Ti-doped CNT-OH (Ti_{ad}-CNT-OH), $NaBH_4$-treated Ti-doped CNT-OH (r-Ti_{ad}-CNT-OH), Cr-doped CNT-OH (Cr_{ad}-CNT-OH), and $NaBH_4$-treated Cr-doped CNT-OH (r-Cr_{ad}-CNT-OH)—to enhance visible light absorption and photocatalytic activity efficiently.[29–33] It has been well documented that transition metal doping induces oxygen vacancies (O_V) on CNT surfaces, improving visible light responsiveness, while $NaBH_4$ treatment generates numerous defect structures.[29] Surprisingly, Ti- and Cr-doped CNTs exhibit stable internalization in renal cancer cells and result in apoptotic cell death upon visible light irradiation, while r-TM_{ad}-CNT-OH (TM = Cr or Ti) shows marginal cellular uptake. Specific generation of ROS such as hydroxyl radicals within renal cancer cells triggers selective cell death. To address the versatility of these enhanced photocatalysts, the photocatalytic conversion reactions (PCRs) of furfuraldehyde (FA), 5-hydroxymethylfurfural (HMF), and toluene were investigated. Among CNT derivatives, r-Cr_{ad}-CNT-OH demonstrated high conversion yields due to the formation of oxygen vacancies, surface lattice contraction, and defect structures that efficiently suppressed the electron–hole recombination. This study provides insights into the optimal surface structural modifications for dual applications: expanding the repertoire of anticancer candidates and advancing visible-light-driven photocatalysts in sustainable energy solutions.

Experimental

Preparation of CNT-OH derivatives

CNT-OH composites were synthesized *via* a modified thermal method.[34] After fabricating and confirming CNT-OH composites,[35] two Cr ions (TM nitrate *n*-hydrate: $Cr(NO_3)_3{\cdot}9H_2O$ and $Cr(NO_3)_2{\cdot}6H_2O$, Sigma-Aldrich, 99%) or Ti ions (TM nitrate *n*-hydrate: $TiCl_4$, Sigma-Aldrich, 99%) were added to the CNT-OH composite at 95 °C under stirring (rpm = 110) until a homogeneous and transparent mixture was formed (approximately 3 h).[36,37] The solution mixture was then heated at 220 °C for 10 h. The resulting Cr or Ti-ion-doped CNT-OH (Cr_{ad}-CNT-OH and Ti_{ad}-CNT-OH) was filtered and washed with double-distilled water to remove any residue. After fabricating TM_{ad}-CNT-OH, we obtained r-TM_{ad}-CNT-OH through the following procedure. TM_{ad}-CNT-OH was mixed with 10 wt% $NaBH_4$ and transferred to a tube furnace (Lab House). The mixed solution was heated at a ramp rate of 12.5 °C min^{-1} to 400 °C, then it was heated for 1 hour under Ar-purging conditions. It was then washed 10 times with double distilled water (DDW) and 10 times with ethanol to remove contaminants such as Na^+ and BO_2^- ions, followed by drying at 95 °C.

Cell culture

Human renal carcinoma cell lines (786-O, Caki-1, and A-498) and human dermal fibroblast cells (CCD-986Sk) were cultured as monolayers. 786-O cells were maintained in Roswell Park Memorial Institute 1640 medium (RPMI 1640, Welgene), Caki-1 and A-498 cells in Dulbecco's Modified Eagle's Medium (DMEM, Welgene), and CCD-986Sk cells in Iscove's Modified Dulbecco's Medium (IMDM, Welgene). All media were supplemented with 10% fetal bovine serum (FBS, Gibco, USA) and 1% penicillin–streptomycin (Welgene). Cells were incubated at 37 °C under a humidified atmosphere containing 5% CO_2.

Cell viability

To determine cell viability, cells were seeded in 96-well plates and allowed to adhere for 24 hours. After incubation, the cells were treated with CNT-OH, Ti_{ad}-CNT-OH, r-Ti_{ad}-CNT-OH, Cr_{ad}-CNT-OH, and r-Cr_{ad}-CNT-OH, and DW was used as a control for 6 hours. Then the cells were irradiated with a 445 nm laser for 20 minutes and incubated at 37 °C with 5% CO_2, then cell viability was evaluated using 2-(2-methoxy-4-nitrophenyl)-3-(4-nitrophenyl)-5-(2,4-disulfophenyl)-2*H*-tetrazolium, monosodium salt (WST-8, Dojindo). Absorbance was measured at 450 nm using a microplate reader (VICTOR Nivo™, Waltham, MA, USA).

Cellular reactive oxygen species assay

To investigate intracellular ROS levels, a DCFDA-based assay was performed using a cellular ROS detection assay kit (Abcam, #ab113851) in accordance with the manufacturer's instructions. 786-O cells were seeded into black 96-well plates at a density of 6000 cells/well and incubated overnight to allow for adherence. Cells were then treated with Ti_{ad}-CNT-OH, Cr_{ad}-CNT-OH, and DW and incubated for 24 hours. After incubation, DCFDA was added to each well to a final concentration of 25 μM and incubated for 20 minutes. The DCFDA-containing medium was then removed, and the cells were washed once with phosphate-buffered saline (PBS). Fluorescence

 This journal is © The Royal Society of Chemistry 2025

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

images were then acquired using the Lionheart FX automated imaging microscope (BioTek).

Mitochondrial membrane potential assay

The mitochondrial membrane potential ($\Delta\Psi_m$) was assessed using the JC-1 staining assay (Abcam #ab113850). Briefly, 786-O cells were seeded into 96-well plates and incubated for 24 hours to allow for attachment. The cells were treated with nanoparticles for 6 hours, followed by irradiation at 445 nm. To evaluate mitochondrial depolarization, 100 μL of JC-1 solution was added to each well containing the nanoparticle-treated medium, resulting in a final JC-1 concentration of 5 μg mL^{-1}. The plate was then incubated at 37 °C for 15 minutes under standard culture conditions. Prior to fluorescence measurement, the JC-1-containing medium was removed, and the cells were washed with PBS. JC-1 aggregates were detected at 586/647 nm (red fluorescence), and JC-1 monomers were measured at 469/525 nm (green fluorescence). Fluorescence images were then acquired using the Lionheart FX automated imaging microscope (BioTek).

PI and Calcein AM staining

Live/dead cell staining was performed using calcein acetoxymethyl ester (Calcein-AM) to identify viable cells and propidium iodide (PI) to detect membrane-compromised dead cells. Briefly, 786-O cells were seeded at a density of 2.5×10^5 cells per well in a 6-well plate and incubated overnight to allow for cell attachment. Cells were then treated with DW, Ti_{ad}-CNT-OH, and Cr_{ad}-CNT-OH for a total of 24 h. For the Ti_{ad}-CNT-OH, cells were irradiated with a 445 nm laser for 20 min after 6 h of incubation, followed by an additional incubation. The DW and Cr_{ad}-CNT-OH groups were incubated for 24 hours without irradiation. Following the incubation, cells were stained with Calcein-AM (Thermo Fisher, USA) and PI (Thermo Fisher, USA) at 37 °C for 30 minutes under light-protected conditions. Fluorescence images were acquired using the Lionheart FX automated imaging microscope (BioTek).

Immunocytochemistry

For immunocytochemistry, cells were seeded and allowed to adhere for 24 h. The cells were then fixed with 4% paraformaldehyde in phosphate-buffered saline (PBS) for 15 minutes at room temperature and permeabilized with 0.20% Triton X-100 in PBS for 10 minutes. After fixation and permeabilization, 4′,6-diamidino-2-phenylindole (DAPI, CTS #4083) was added to each well to a final concentration of 1 μg mL^{-1} and incubated for 5 minutes at room temperature in the dark. The cells were subsequently washed three times with PBS to remove excess stains. Fluorescence images were then acquired using the Lionheart FX automated imaging microscope (BioTek).

Western blot analysis

Renal cancer cells treated with Ti_{ad}-CNT-OH and Cr_{ad}-CNT-OH were harvested for protein analysis. Total proteins were extracted using an RNA/protein extraction kit (MACHEREY-NAGEL) according to the manufacturer's instructions. Protein concentrations were determined using the BCA assay for normalization. Equal amounts of protein lysates were separated on 8–16% SDS-PAGE gels (Bio-Rad, #BR4561106) and transferred onto polyvinylidene difluoride (PVDF) membranes (GenDEPOT, #W7031-540). The membranes were blocked, and primary antibodies were diluted in PBS containing 1% skim milk and 1% Tween-20 (Sigma), followed by overnight incubation at 4 °C. The following primary antibodies were used: cleaved caspase-9 (Cell Signaling Technology, #7237), cleaved PARP (Abcam, #ab32064), Bcl-2 (Abcam, #ab182858), cytochrome c (Abcam, #ab133504), and β-actin (Abcam, #ab8227). After washing, the membranes were incubated for 1 hour at room temperature with horseradish peroxidase (HRP)-conjugated secondary antibodies diluted in PBS containing 2% skim milk and 1% Tween-20. Protein bands were visualized using a chemiluminescent detection reagent (ATTO) and imaged using the LuminoGraph III Lite imaging system (WSE-6370K, ATTO, Japan).

Statistical analysis

The statistical analysis was performed using a two-tailed paired *t* test, one way ANOVA–Dunnett's test and nonlinear regression analysis in GraphPad Prism 10 software (San Diego, CA, USA).

Characterization of CNT-OH derivatives

C K-edge, O K-edge, Cr L-edge, and Ti L-edge spectra were obtained through soft X-ray absorption spectroscopy (XAS) at the Pohang Accelerator Laboratory (PAL, Republic of Korea) beamline 4D. Raman spectral data were collected using an Ar^+-ion laser (λ_{ex} = 514.5 nm, Stabilite 2017, Spectra-Physics, USA) as the excitation source connected to a spectrometer (TRIAX 550, Horiba Jobin Yvon, Japan). Photocurrent measurements were carried out using a 6 W LED lamp as the light source for photocatalytic reactions at 445 nm, and the photocurrent densities were measured using a two-electrode system comprising the fabricated working electrodes and a Pt gauze (0.8 × 0.8 cm) as the counter electrode using a multichannel potentiostat (WonATech). The O_Vs on the surfaces of the five CNTs were detected at 300 K using electron paramagnetic resonance (EPR; JES-X320, JEOL Ltd) spectroscopy in the 345–350 mT range (frequency: 9446 MHz; power: 10 mW; modulation frequency: 100 kHz; width: 0.4 mT).

Photocatalytic measurements

The photocatalytic conversion reaction (PCR) efficiencies of the five samples were evaluated under visible light irradiation. Suspensions of the composites (0.5 g L^{-1}) were prepared in an aqueous solution containing the target substrate (25 mM furfuraldehyde (FA), 5-hydroxymethylfurfural (HMF), or toluene) and stirred in the dark for 3 hours to allow the system to reach the adsorption-desorption equilibrium. The reactor containing the composite solution was placed 10 cm in front of the LEDs (λ = 445 nm, output power = 6 W, illumination intensity = 230 mW cm^{-2}; Thorlab) under magnetic stirring at ~110 rpm. The concentrations of furfuraldehyde, HMF and toluene were

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

measured using a UFLC LC-20AD (Shimadzu, Japan) high-performance liquid chromatography system.

Results and discussion

Macroscale spectroscopic analysis of the electronic structure of TM_{ad}-CNT-OH and r-TM_{ad}-CNT-OH

To explore the morphological and electronic structural changes of OH-functionalized CNTs induced by transition metal (TM) doping and $NaBH_4$ treatment, X-ray absorption spectroscopy (XAS), Raman spectroscopy, photocurrent measurements, and electron paramagnetic resonance (EPR) spectroscopy were conducted. We first investigated the electronic structures of the samples through XAS analysis for C K-edge and O K-edge. Fig. 1a shows the characteristic C K-edge spectra for the five samples. The C 1s–π* transition (C1) at 284.2 eV was associated with the C=C sp^2 bond, whereas the C 1s–π* transition (C2) at 287.4 eV corresponded to C=O or C–OH bonds. The third peak transitions, C3 at 290.7 eV and C4 at 291.8 eV, were attributed to σ* excitons and 1s–σ* transitions for aliphatic C–C bonds, respectively.[38,39] In general, the C2 peak intensity of the C K-edge XAS spectrum of carbon allotropes can explain the amount of defect structures in the samples. Compared to the C2 peak of CNT-OH, the intensity of the C2 transition peak of CNT-OH doped with transition metals increased considerably. In particular, the C2 transition peak of TM_{ad}-CNT-OH calcined with additional $NaBH_4$ was significantly increased compared to that of CNT-OH. This result suggests that $NaBH_4$ calcination following transition metal doping can significantly increase the defect structure in CNT-OH. The doping of transition metal ions onto the CNT-OH surface can result in their adsorption onto or substitution at the –OH group positions, creating defect structures. Interestingly, further defect structures could be amplified by additional $NaBH_4$ calcination, and it was plausible to predict that the reduction reactions at the metal adsorption sites would be further accelerated.

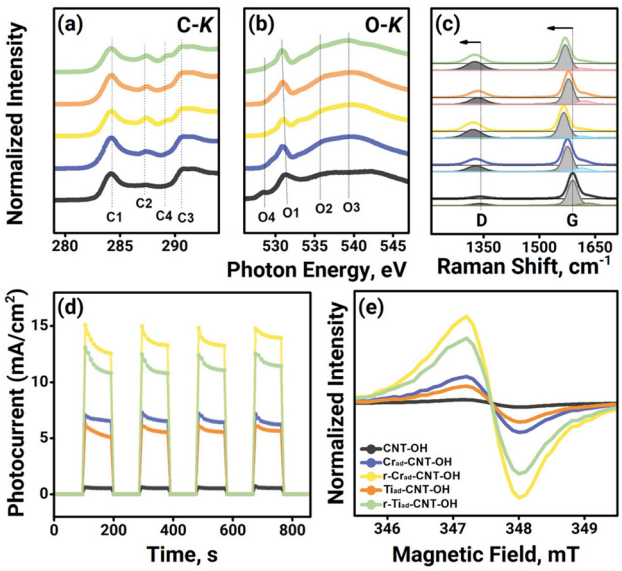


Fig. 1 XAS results of (a) C K-edge and (b) O K-edge XAS spectra, (c) Raman spectra, (d) photocurrent measurements, and (e) EPR spectra of the CNT-OH (black), Cr_{ad}-CNT-OH (blue), r-Cr_{ad}-CNT-OH (yellow), Ti_{ad}-CNT-OH (orange), and r-Ti_{ad}-CNT-OH (green) samples.

The O K-edge XAS spectra shown in Fig. 1b were measured by using an X-ray energy larger than the 530 eV excitation energy of an oxygen 1s core electron to an empty state. Based on the O K-edge of CNT-OH, the absorptions in the range of 530–534 eV (*i.e.*, O1) were attributed to the O 1s–π* transition of C=O and OH groups, while the absorptions in the range of 534–544 eV (*i.e.*, O2 and O3) appeared due to the O 1s–σ* transition of –OH, C–O, and C=O groups.[40] In particular, the absorption edges at 528.5 eV (O4) and 531.1 eV (O1) corresponded to π* resonances associated with the C=O and CNT-OH functional groups. Based on the O K-edge spectra of CNT-OH, the changes in the O K-edge spectra of TM_{ad}-CNT-OH and r-TM_{ad}-CNT-OH were compared. The most notable feature in the XAS spectra was the change in the π* orbital peak associated with the OH group present in the CNTs after doping with TM ions. The resonance peak intensities of the four other samples were markedly lower than the O4 peak in CNT-OH. Cr- or Ti-doped CNT-OH exhibited a diminished OH peak, suggesting that transition metals had adsorbed onto or replaced –OH sites, thereby enhancing defect structures associated with their catalytic activity. The electronic states of the doped Cr and Ti ions in TM_{ad}-CNT-OH were analyzed in detail by measuring the L-edge XAS spectra of Cr and Ti (Fig. S1). The spectra confirmed the electronic states of Cr^{3+} and Ti^{4+} oxides, suggesting that the difference in catalytic properties stemmed from defect density in metal oxides. Even after $NaBH_4$ calcination, Cr and Ti ions likely occupied –OH sites as stable Cr_2O_3 and TiO_2, maintaining molecular-level defects, though their precise form remains indistinct.[41,42]

Therefore, the XAS measurement confirmed that transition metal ion adsorption at –OH sites in transition metal ion-doped CNT-OH could generate high-density defect structures.

Fig. 1c shows the Raman spectra of CNT-OH, TM_{ad}-CNT-OH, and r-TM_{ad}-CNT-OH. The D-peak at 1345.8 cm^{-1} and G-peak at 1577.7 cm^{-1} in the Raman spectrum of CNT-OH were assigned to the E_{2g} phonons at the Brillouin zone centers and the breathing modes of sp^2-carbon atoms, respectively.[43] The Raman shifts of the D and G peaks of TM_{ad}-CNT-OH and r-TM_{ad}-CNT-OH shifted to lower binding energies, indicating that an increase in defect structures on the surface of CNT-OH weakened binding interactions. This further supported defect formation due to TM doping and $NaBH_4$ calcination.

Creating O_Vs on the surfaces of CNT-OH through TM ion doping is generally a simple post-synthetic modification strategy that has the potential to improve the catalytic efficiency of the parent material.[44,45] O_Vs serve as mobile electron donors because they modulate the local valence states of neighboring cations at the atomic scale. In particular, O_Vs enable: (i) additional electron-charge carriers, (ii) metal-ion absorption at the material surface, and (iii) changes in the geometrical com-

 This journal is © The Royal Society of Chemistry 2025

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

position of the oxide.[46–48] In this context, we used sodium borohydride ($NaBH_4$), which is a well-known reducing agent,[49,50] to generate additional O_Vs. Borohydride ions were hydrolyzed to produce hydrogen gas in the aqueous solution with transition-metal ions according to the following equation:

$$TM_{ad}\text{-CNT-OH} + Na^+ + BH_4^- + 2H_2O \rightarrow TM_{ad}\text{-CNT-}(OH)_{1-x} + Na^+ + BO_2^- + 4H_2$$

Here, metal ions were transformed into dark precipitates known as "metallic catalysts" or "metal borides"; therefore, electron transfer from the borohydride to the metal ion is the key step in this process generating several metal ions and O_Vs in the CNT-OH and creating multiple defects on their surfaces.[51,52]

We further investigated the photocurrents of the five test CNT-OH composites to examine the charge-separation efficiencies of the photogenerated electron–hole pairs by irradiating them with visible light for four on–off cycles and measuring their transient photocurrent responses (Fig. 1d). Interestingly, the current density of TM_{ad}-CNT-OH increased by approximately 25 times, and in the two samples of r-TM_{ad}-CNT-OH, a current density more than 50 times higher was observed. This was consistent with the formation of defect structures on the surface, which trapped electrons at specific defect sites, thereby mitigating the recombination of photo-generated charge carriers and generating significantly higher currents than those produced in CNT-OH.

Furthermore, the EPR spectra of CNT-OH, TM_{ad}-CNT-OH, and r-TM_{ad}-CNT-OH display a prominent signal at a *g*-factor of about 2.003, which is indicative of the Zeeman effect induced by a single electron trapped within the defect structures (Fig. 1e). By comparing the integrated radical signals, it was evidenced that TM_{ad}-CNT-OH and r-TM_{ad}-CNT-OH contained more defect structures than CNT-OH. Overall, additional $NaBH_4$ calcination was more effective than TM doping alone, which is a direct result of the higher number of O_Vs formed on the surfaces of the modified CNT-OH.

Generation of ROS and ROS-mediated renal cancer cell death

To assess the ability of CNT-OH derivatives to induce the *in situ* formation of ROS such as hydroxyl radicals (˙OH), we conducted the photocatalytic reactions between CNT-OH derivatives and specific organic radical probes. Terephthalic acid (TA) was employed to detect hydroxyl radical species due to its selective oxidation.[53] The reaction solution containing the five CNT derivatives and TA exhibited negligible fluorescence intensity. Upon 445 nm irradiation, strong fluorescence intensity was observed using fluorescence spectrophotometry (λ_{ex} = 310 nm) within 30 min due to the reaction between hydroxyl radical species and non-fluorescent TA producing fluorescent 2-hydroxyterephthalic acid (TA-OH), which exhibits strong emission bands at 410 nm at room temperature (Fig. 2a). However, CNT-OH did not show any fluorescence intensity. This clearly suggests that the photocatalytic reaction between the CNT derivatives and water resulted in the formation of hydroxyl radical species.[54] Interestingly, the normalized fluorescence intensity order was found to be CNT-OH (none) < r-Cr_{ad}-CNT-OH < Cr_{ad}-CNT-OH < Ti_{ad}-CNT-OH <

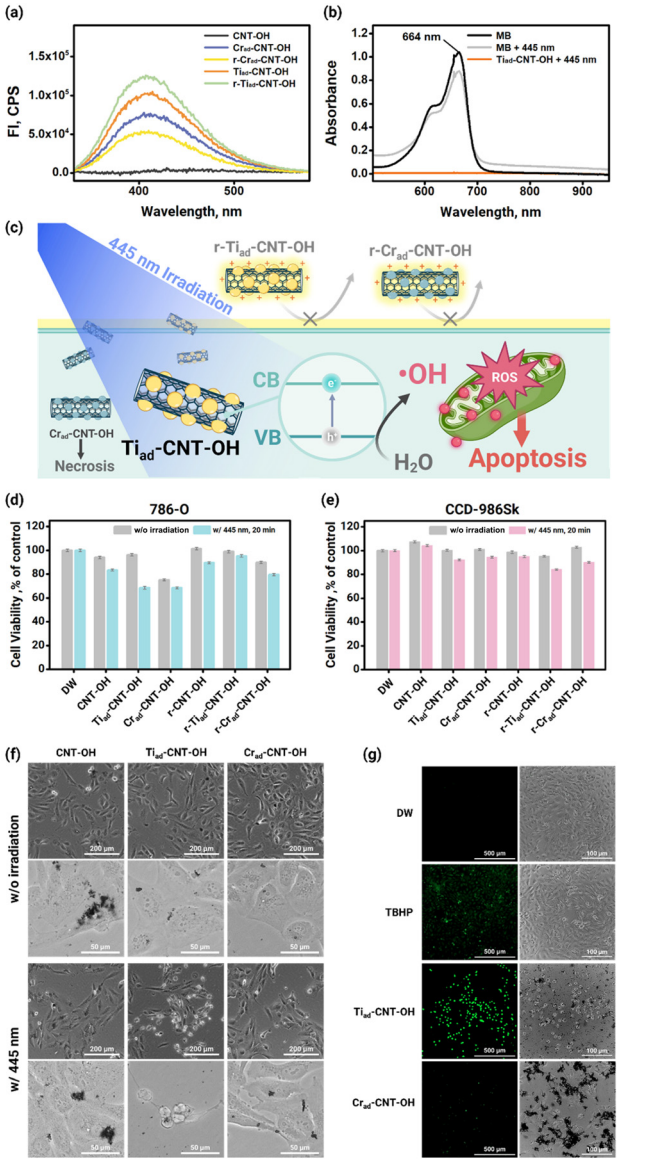


Fig. 2 (a) Fluorescence spectrum of hydroxyl radical generation detected using terephthalic acid (TA, 10 mM) in the presence of CNT-OH (black), Cr_{ad}-CNT-OH (blue), r-Cr_{ad}-CNT-OH (yellow), Ti_{ad}-CNT-OH (orange), and r-Ti_{ad}-CNT-OH (green) samples under 445 nm irradiation. (b) UV-vis spectrum obtained in the photocatalytic degradation of methylene blue (0.010 mM) in the presence of Ti_{ad}-CNT-OH in DW at 20 °C under 445 nm light irradiation. (c) Schematic illustration of the proposed mechanism for selective apoptosis. Photodynamic anticancer effect of CNT-OH (40 μg mL^{-1}), r-CNT-OH (40 μg mL^{-1}), TM_{ad}-CNT-OH (40 μg mL^{-1}), and r-TM_{ad}-CNT-OH (40 μg mL^{-1}) on (d) renal cancer (786-O) and (e) fibroblast (CCD-986Sk) with or without 445 nm irradiation. (f) Microscopy images of 786-O cells treated with 20 μg mL^{-1} of CNT-OH, Ti_{ad}-CNT-OH, and Cr_{ad}-CNT-OH for 24 hours with or without 445 nm irradiation. Scale bars: 200 μm and 50 μm. (g) Intracellular ROS levels in 786-O cells treated with 40 μg mL^{-1} of Ti_{ad}-CNT-OH (irradiated at 445 nm), Cr_{ad}-CNT-OH (without irradiation) and TBHP. Scale bars: 500 μm and 100 μm.

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

r-Ti_{ad}-CNT-OH. Hence, the present result confirmed the potential of transition metal doping and/or $NaBH_4$ treatment of OH-functionalized CNTs for the *in situ* production of hydroxyl radical species in aqueous solutions under aerobic conditions. To confirm the generation of hydroxyl radical species, a bleaching assay was carried out by using methylene blue (Fig. 2b). We assessed hydroxyl radical formation with a methylene-blue bleaching assay. Illumination alone left the 664 nm band essentially intact, whereas the presence of Ti_{ad}-CNT-OH quenched it rapidly, consistent with swift hydroxyl radical-driven degradation of the dye.

Cell viability and intracellular ROS detection

With the deepening of our studies, it was intriguing to understand how the photocatalytic production of hydroxyl radicals by CNT derivatives could contribute to the disruption of intracellular redox homeostasis, thereby inducing renal cancer cell death (Fig. 2c). The potential cytotoxicity of these CNT derivatives with different concentrations (10, 20, 40, 60 μg mL^{-1}) was scrutinized using the WST-8 assay—which measures mitochondrial activity—in both renal cancer cells (786-O, Caki-1, and A498) and normal skin cells (CCD-986Sk) (Fig. S2 and S3). These results suggest that increasing the dose of CNT derivatives progressively suppressed 786-O cell proliferation, with Cr-doped CNT-OH exerting the strongest inhibition—evidence that chromium incorporation heightens CNT cytotoxicity.

In the absence of irradiation, a 24 hour incubation with all CNT derivatives (except for Cr_{ad}-CNT-OH) did not induce any observable cell death in either cell type, strongly demonstrating the minimal damage to healthy cells under the conditions tested (Fig. 2d). Under 442 nm irradiation, Ti_{ad}-CNT-OH demonstrated pronounced photodynamic anticancer activity against renal cancer cells, which was in part consistent with previous ROS detection experiments; r-Ti_{ad}-CNT-OH showed no impact on the cell viability (Fig. 2a). In contrast, Cr_{ad}-CNT-OH exhibited a modest decline in cell viability, likely owing to the inherent cytotoxicity of the chromium ions while r-Cr_{ad}-CNT-OH did not elicit significant photodynamism-mediated cancer cell death (Fig. 2d). In addition, CNT-OH derivatives exhibited minimal interaction with normal tissues, whether irradiated or not, suggesting reduced cytotoxicity, consistent with CNT behavior (Fig. 2e). These findings clearly underscored that the choice of the doping metal plays a key role in modulating the cytotoxicity of CNT-OH derivatives. Unexpectedly, despite their capacity for *in situ* ROS generation under visible light irradiation, $NaBH_4$-treated transition metal doped CNT-OH, such as r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH, exhibited negligible cytotoxicity both with and without light exposure (*vide infra*).

Microscopy images of 786-O cells were taken before and after treatment with CNT derivatives in the presence and absence of irradiation (Fig. 2f). It revealed that after incubation, two CNT-OH derivatives, Ti_{ad}-CNT-OH and Cr_{ad}-CNT-OH, penetrated the cell membrane whereas r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH were immiscible with the cellular membrane and remained in extracellular environments (Fig. 2f and S4). This is a key observation that explains why the additional $NaBH_4$ treatment on transition metal doped CNT-OH could not exhibit anticancer activity while additional $NaBH_4$ treatment enhanced their photocatalytic activity (Fig. 2c). X-ray photoelectron spectroscopy was performed to explore the surface charge of each CNT derivative (Fig. S5). It revealed carbon binding energies at 284.7 eV in CNT-OH, Ti_{ad}-CNT-OH, and Cr_{ad}-CNT-OH, while r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH exhibited a slight shift to ~285.0 eV, indicating increased surface charge in CNT-OH. This finding conclusively demonstrated that increased surface charges of $NaBH_4$-treated CNT derivatives enhanced surface polarity, thereby hindering their cellular uptake across the lipophilic membrane.

To scrutinize the correlation between intracellular ROS production by CNT-OH derivatives and renal cancer cell proliferation, quantitative detection of intracellular ROS concentration in 786-O cells was performed under identical incubation conditions by using the cell permeant reagent, 2′,7′-dichlorofluorescindiacetate (DCFDA) (Fig. 2g and S6). In line with the confirmed results with TA, the fluorescence intensity due to the presence of ROS was remarkably prominent (>2-fold) when Ti_{ad}-CNT-OH was incubated together with 786-O cells while Cr_{ad}-CNT-OH displayed a slight increase of fluorescence intensity (~1.3-fold) as compared to that of the control experiment (Fig. S6). These observations allowed us to speculate that Ti_{ad}-CNT-OH more efficiently generates intracellular ROS *via* a photodynamic process in a heterogeneous milieu (*i.e.*, cancer cells) compared to Cr_{ad}-CNT-OH under identical conditions.

Taken together, we propose that CNT-OH derivatives demonstrate off-target toxicity in both hetero- and homogeneous milieus. While the presence of Cr_{ad}-CNT-OH exhibited slight attenuation of cytotoxicity in both healthy and tumor cells, Ti_{ad}-CNT-OH efficiently increased the intracellular concentration of ROS, thereby inducing cell death by presumably disturbing key regulatory signaling processes meticulously related to intracellular ROS homeostasis.

Immunocytochemistry and western blot analysis

Since mitochondrial membrane potential ($\Delta\Psi_m$) reflects redox homeostasis,[55,56] JC-1 assay was conducted to evaluate its dynamics (Fig. 3a). JC-1 dye transitions from a monomeric green emission (low $\Delta\Psi_m$) to aggregated red fluorescence (high $\Delta\Psi_m$) under polarized conditions. Ti_{ad}-CNT-OH showed predominant monomeric green fluorescence, indicating significant mitochondrial depolarization, while Cr_{ad}-CNT-OH exhibited only a slight green fluorescence. As a reduced red/green fluorescence ratio suggests apoptosis or mitochondrial dysfunction and an increased red/green fluorescence ratio suggests enhanced mitochondrial activity or stability, Ti_{ad}-CNT-OH demonstrated effective photodynamic activity.

To assess the efficacy of the anticancer activity of the CNT-OH derivatives, flow cytometry using calcein AM (detection of live cells) and propidium iodide (PI, detection of damaged and/or dead cells) was performed to distinguish and quantify live and dead cells (Fig. 3b). Without light exposure, Cr_{ad}-CNT-OH exhibited PI-stained dead cells. On the other

 This journal is © The Royal Society of Chemistry 2025

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

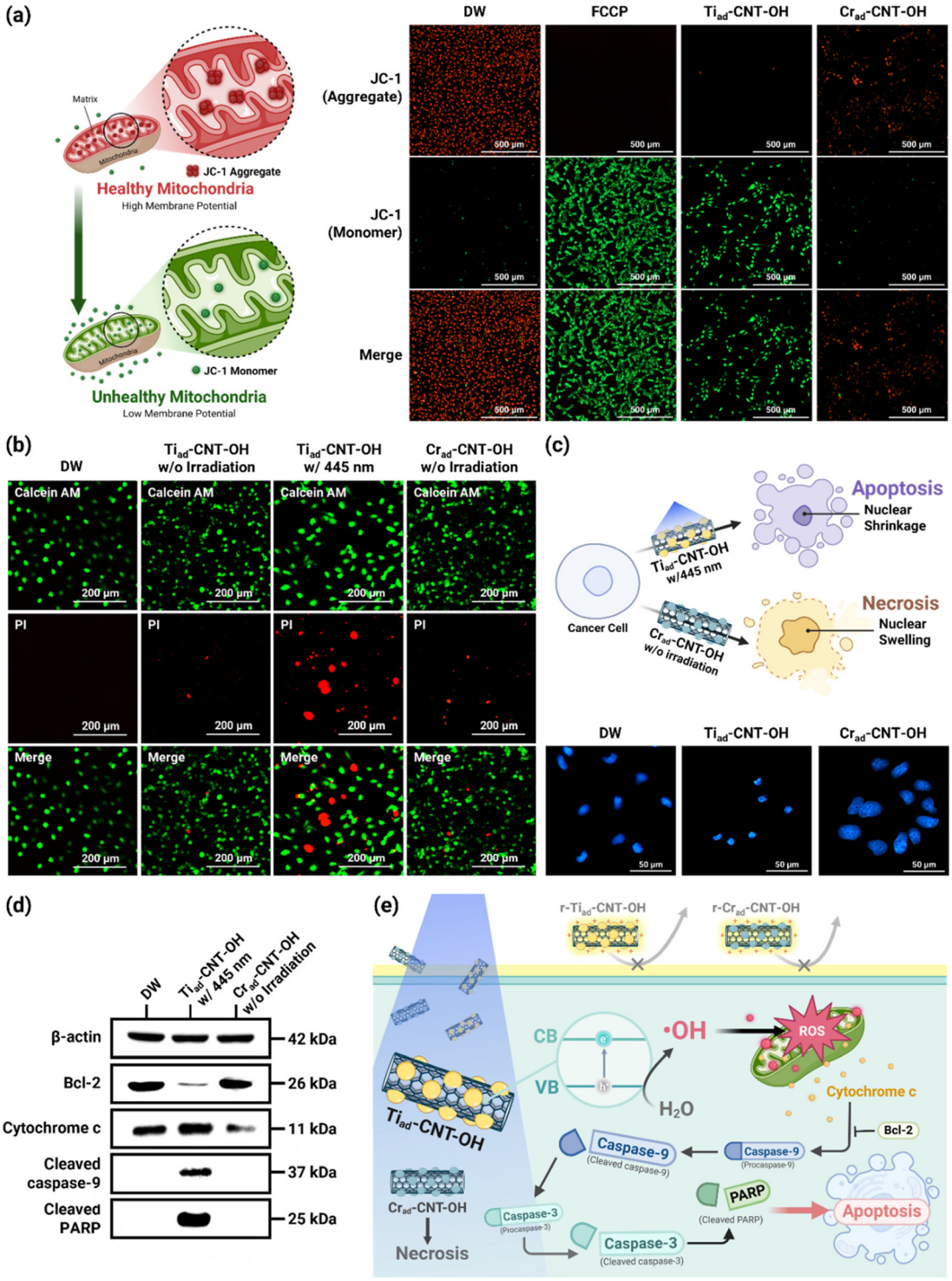


Fig. 3 (a) (Left) Illustration of mitochondrial membrane depolarization. (Right) JC-1 staining of 786-O cells after treatment with 40 μg mL^{-1} CNT-OH, Ti$_{ad}$-CNT-OH (445 nm-irradiation), and Cr$_{ad}$-CNT-OH (without irradiation) for 24 h. Scale bar: 500 μm. (b) Calcein AM and PI staining in 786-O cells treated with 40 μg mL^{-1} CNT-OH, Ti$_{ad}$-CNT-OH (445 nm-irradiation), and Cr$_{ad}$-CNT-OH (without irradiation). Scale bar: 200 μm. (c) DAPI staining of 786-O cells treated with 40 μg mL^{-1} Ti$_{ad}$-CNT-OH (445 nm-irradiation) and Cr$_{ad}$-CNT-OH (without irradiation) for 24 h. Scale bar: 50 μm. (d) Western blot analysis of apoptosis-related protein expression in 786-O cells treated with 40 μg mL^{-1} CNT-OH, Ti$_{ad}$-CNT-OH (445 nm-irradiation) and Cr$_{ad}$-CNT-OH (without irradiation) for 24 hours. (e) Illustration of cell death mechanisms induced by Ti$_{ad}$-CNT-OH (445 nm-irradiation) and Cr$_{ad}$-CNT-OH (without irradiation).

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

hand, Ti_{ad}-CNT-OH barely showed PI active red fluorescence in the absence of irradiation while strong red fluorescence intensity was observed upon irradiation, indicating photodynamically triggered renal cancer cell death.

Immunofluorescence staining with DAPI was used to visualize the nuclear DNA in both living and fixed cells and 786-O cells within 24 h after treatment (Fig. 3c). Very interestingly, it was found that two distinct morphological changes were observed depending on CNT-OH derivatives. When 786-O cells were incubated with Ti_{ad}-CNT-OH, deformation of nuclei including holes and shrinkage accompanied by mitochondrial membrane dismantlement were clearly noticed, indicating the formation of the apoptotic nuclei.[57] However, incubation of 786-O cells with Cr_{ad}-CNT-OH induced necrotic nuclear changes such as DNA degradation and fragmentation into irregular, dispersed pieces, indicative of uncontrolled cell death.[58,59] These nuclear alterations clearly highlighted the progressive breakdown of cellular integrity and a stark contrast between Cr_{ad}-CNT-OH-caused necrotic cell death and Ti_{ad}-CNT-OH-mediated apoptotic cell death.

To understand the antiproliferative effect of CNT derivatives on the renal cancer cell death pathway, the level of genes and proteins closely related to the apoptotic cell death pathway was supervised. First, the expression of Bcl-2, which is known to be an anti-apoptotic member, was verified by using qRT-PCR. A clear decrease of the expression level of Bcl-2 reflected the occurrence of apoptotic cell death in 786-O cells upon incubation of Ti_{ad}-CNT-OH (Fig. 3d). However, the Bcl-2 level was barely affected owing to Cr_{ad}-CNT-OH. The increase in the level of cytochrome c was directly related to the mitochondrial dysfunction and caused the activation of caspase 9 along with PARP indubitably indicating that Ti_{ad}-CNT-OH in the presence of irradiation induced renal cancer cell death (Fig. 3e).[60,61]

Photocatalytic activity analysis of CNT-OH, TM_{ad}-CNT-OH, and r-TM_{ad}-CNT-OH

Furfuraldehyde (FA), an abundant biomass resource, undergoes transformation into low-toxicity derivatives, producing high-value compounds. 5-Hydroxymethylfurfural (HMF), a key biomass-derived platform chemical, oxidizes to 2,5-furandicarboxylic acid (FDCA), a promising sustainable alternative to petroleum-based terephthalic acid.[62,63] Toluene necessitates efficient degradation due to its high volatility and environmental persistence. The selectivity of intermediates in oxidation reactions directly impacts final product yield and efficiency. To track these processes, product concentrations were meticulously monitored over time using high-performance liquid chromatography (HPLC).

To assess the photocatalytic performance of five CNT-OH, we exposed the liquid suspension containing three model molecules (HMF, toluene, and FA) under visible light irradiation ($\lambda \geq 445$ nm) with an illumination intensity of 230 mW cm^{-2}. The photocatalytic degradation (PCD) profiles of three molecules highlighted distinct performance differences among CNT-OH constructs. Fig. 4a–c showed PCD activities of HMF, toluene, and FA, respectively, as functions of

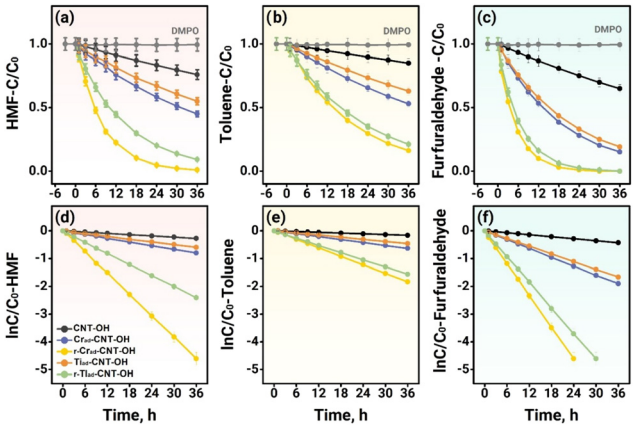


Fig. 4 Photocatalytic conversion activities (PCR) of the various CNT-OH samples toward (a) HMF, (b) toluene, and (c) furfuraldehyde. Experimental conditions: [catalyst] = 0.5 g L^{-1}, $\lambda \geq 445$ nm, $[HMF]_0$ = $[toluene]_0$ = $[furfuraldehyde]_0$ = 25 mM. Plot of C/C_0 against time to calculate the first-order rate constants (k) of (d) HMF, (e) toluene, and (f) furfuraldehyde for the photocatalytic reactions of each CNT sample. [Key: CNT-OH (black), Cr_{ad}-CNT-OH (blue), r-Cr_{ad}-CNT-OH (yellow), Ti_{ad}-CNT-OH (orange), and r-Ti_{ad}-CNT-OH (green) samples].

time. r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH exhibited markedly enhanced PCD properties under visible light, outperforming CNT-OH, Cr_{ad}-CNT-OH, and Ti_{ad}-CNT-OH, which showed negligible activity in the visible-light range. Photocatalytic performance followed the order: r-Cr_{ad}-CNT-OH > r-Ti_{ad}-CNT-OH > Cr_{ad}-CNT-OH > Ti_{ad}-CNT-OH > CNT-OH. These results are summarized in Table 1.

Moreover, the five CNT-OH derivatives exhibited linear correlations between $\ln(C/C_0)$ and reaction time; the photocatalytic degradation reactions over the various CNTs were fitted to pseudo-first-order kinetics to determine their first-order rate constants (k) (Fig. 4d–f). The k values of the CNT-OH, Cr_{ad}-CNT-OH, r-Cr_{ad}-CNT-OH, Ti_{ad}-CNT-OH, and r-Ti_{ad}-CNT-OH in the oxidation reaction of HMF were calculated to be 0.0076, 0.022, 0.13, 0.017, and 0.067 h^{-1}, respectively. Kinetically, r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH accelerated the oxidation reaction by a factor of 16 and 7.8 compared with that recorded for CNT-OH (Fig. 4d). Treatment with r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH significantly enhanced reaction rates, accelerating toluene oxidation by factors of 10 and 8.6 (Fig. 4e) and FA conversion by 15 and 12 (Fig. 4f) compared to CNT-OH.

Organic products were analyzed and tracked by using HPLC analysis and the yields of derivatives were consistent with the observed PCD efficiencies (Fig. S7–S9). The oxidative mechanisms of the three test molecules involved electron and water trapping at NP surface oxygen vacancies, generating hydroxyl radicals ($HO^{\bullet -}$). These radicals, along with photogenerated holes (h^+), facilitated the conversion of HMF to FDCA (Fig. S10a), toluene to benzaldehyde (Fig. S10b), and FA to furoic acid (Fig. S10c and Scheme S1). To confirm the role of $HO^{\bullet -}$ in photocatalysis, 5,5-dimethyl-1-pyrroline *N*-oxide (DMPO) was introduced as a radical scavenger, inhibiting

 This journal is © The Royal Society of Chemistry 2025

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

Table 1 Photocatalytic conversion reaction of HMF, toluene, and FA for 36 h under visible light irradiation ($\lambda \geq 445$ nm) with 230 mW cm^{-2} illumination intensity

Samples	HMF (C/C_0)	Toluene (C/C_0)	FA (C/C_0)
CNT-OH	0.76 ± 0.038	0.85 ± 0.038	0.65 ± 0.033
Cr_{ad}-CNT-OH	0.45 ± 0.023	0.53 ± 0.023	0.15 ± 0.008
r-Cr_{ad}-CNT-OH	0.01 ± 0.001	0.16 ± 0.008	0.00 ± 0.000
Ti_{ad}-CNT-OH	0.55 ± 0.028	0.63 ± 0.032	0.19 ± 0.010
r-Ti_{ad}-CNT-OH	0.09 ± 0.005	0.21 ± 0.011	0.00 ± 0.000

photocatalytic activity (Fig. 4a–c and Table S1). We also confirmed that the CNT-OH derivatives were recovered after photocatalysis (Fig. S11).

These results have significant implications and highlight the dual potential of CNT-OH, both as an anticancer agent and as a photocatalyst, offering new directions for therapeutic and catalytic applications. Among the five CNT-OH derivatives, Cr_{ad}-CNT-OH and Ti_{ad}-CNT-OH exhibited anticancer effects in renal cancer cells, inducing necrosis or apoptosis, while r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH had no impact on cell viability. In contrast, r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH demonstrated strong photo-oxidizing capabilities, whereas Cr_{ad}-CNT-OH and Ti_{ad}-CNT-OH displayed only moderate photocatalytic activity toward organic substrates. To understand this disparity between anticancer and photocatalytic performance, we prioritized the role of surface charge and oxygen vacancies in CNT-OH derivatives. Doping with early transition metals, such as Ti and Cr, enhanced photocatalytic current generation compared to CNT-OH (Fig. 1d). Previous reports suggested that Cr ion doping on a metal oxide significantly increased metal oxide cytotoxicity while Ti ion doping had minimal cytotoxic effect.[64,65] In the present work, Ti_{ad}-CNT-OH showed photodynamic activity in renal cancer cells, inducing apoptosis, a programmed cell death pathway, while Cr_{ad}-CNT-OH displayed negligible photodynamic activity, leading to necrosis, a non-programmed cell death pathway. Interestingly, $NaBH_4$ treatment triggered reduction of the transition metal ion doped CNT-OH surfaces, increasing surface charge and oxygen vacancies, thereby rendering the surface more hydrophilic. Since the phospholipid bilayer of the cell membrane served as a semi-permeable barrier, higher hydrophilicity of the CNT-OH surface owing to increased surface charge reduced cellular permeability (Fig. 2f, S4 and S5). This effect explained why r-Cr_{ad}-CNT-OH and r-Ti_{ad}-CNT-OH exhibited minimal photodynamic activity despite their strong photocatalytic performance toward organic substrates.

Conclusions

We have successfully synthesized transition metal doped and $NaBH_4$ treated OH-functionalized CNTs, resulting in significant morphological and electronic structural changes and enhancing defect density and photocatalytic performance. XAS, Raman spectroscopy, and EPR measurements confirmed that $NaBH_4$ calcination amplified oxygen vacancies and surface defects, improving charge separation and visible light-driven catalytic efficiency. The *in situ* photocatalytic generation of ROS, particularly hydroxyl radicals, was validated through photocatalytic reactions with organic radical probes. Photocatalytic production of reactive oxygen species (ROS) by CNT derivatives disrupted intracellular redox homeostasis, leading to renal cancer cell death. The WST-8 assay confirmed dose-dependent cytotoxic effects in renal cancer cells, particularly in Ti_{ad}-CNT-OH under irradiation, while Cr_{ad}-CNT-OH exhibited slight cytotoxicity due to inherent chromium toxicity. Notably, $NaBH_4$-treated CNT-OHs demonstrated negligible cytotoxicity despite strong photocatalytic performance; despite their strong photocatalytic efficiency, $NaBH_4$-treated CNT-OH derivatives exhibited minimal anticancer activity due to increased hydrophilicity, limiting cellular uptake. Therefore, CNT-OH derivatives exhibited dual applicability as anticancer agents and photocatalysts, with transition metal doping and $NaBH_4$ treatment playing key roles in modulating their activity. These findings highlighted the intricate relationship between structural modifications, photocatalytic efficacy, and biological interactions, providing insights into optimizing CNT-based materials for therapeutic and catalytic applications.

Author contributions

Hyungbin Park, Sunyoung Hwang, and Joowon Choi: formal analysis, investigation, visualization, and writing – review & editing. Kyungtae Kang, Hangil Lee, and Seungwoo Hong: conceptualization, formal analysis, funding acquisition, project administration, writing – original draft, and writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information (SI). Supplementary information is available. See DOI: https://doi.org/10.1039/d5nr04008f.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (RS-2024-00346153, RS-2025-21072982 for H. Lee, RS-2024-00414215 for H. Park, RS-2025-16071328 for K. Kang 2021R1A6A1A10039823 and RS-2025-20942968 for S. Hong) and by the Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Ministry of Education (2020R1A6C101B194 for S. Hong).

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

References

1 N. N. Pavlova and C. B. Thompson, The emerging hallmarks of cancer metabolism, *Cell Metab.*, 2016, **23**, 27–47.
2 K. Huang, Y. Han, Y. Chen, H. Shen, S. Zeng and C. Cai, Tumor metabolic regulators: key drivers of metabolic reprogramming and the promising targets in cancer therapy, *Mol. Cancer*, 2025, **24**, 7.
3 M. You, Z. Xie, N. Zhang, Y. Zhang, D. Xiao, S. Liu, W. Zhuang, L. Li and Y. Tao, Signaling pathways in cancer metabolism: mechanisms and therapeutic targets, *Signal Transduction Targeted Ther.*, 2023, **8**, 196.
4 I. Martínez-Reyes and N. S. Chandel, Cancer metabolism: looking forward, *Nat. Rev. Cancer*, 2021, **21**, 669–680.
5 R. Nussinov, B. R. Yavuz and H. Jang, Molecular principles underlying aggressive cancers, *Signal Transduction Targeted Ther.*, 2025, **10**, 42.
6 B. Liu, H. Zhou, L. Tan, K. T. H. Siu and X. Y. Guan, Exploring treatment options in cancer: tumor treatment strategies, *Signal Transduction Targeted Ther.*, 2024, **9**, 175.
7 U. Anand, A. Dey, A. K. S. Chandel, R. Sanyal, A. Mishra, D. K. Pandey, V. De Falco, A. Upadhyay, R. Kandimalla, A. Chaudhary, J. K. Dhanjal, S. Dewanjee, J. Vallamkondu and J. M. Pérez de la Lastra, Cancer chemotherapy and beyond: Current status, drug candidates, associated risks and progress in targeted therapeutics, *Genes Dis.*, 2023, **10**, 1367–1401.
8 A. Letai and H. de Thé, Conventional chemotherapy: millions of cures, unresolved therapeutic index, *Nat. Rev. Cancer*, 2025, **25**, 209–218.
9 G. Petroni, L. C. Cantley, L. Santambrogio, S. C. Formenti and L. Galluzzi, Radiotherapy as a tool to elicit clinically actionable signalling pathways in cancer, *Nat. Rev. Clin. Oncol.*, 2022, **19**, 114–131.
10 F. Mohajer, B. Mirhosseini-Eshkevari, S. Ahmadi, M. A. Ghasemzadeh, G. M. Ziarani, A. Badiei, N. Farshidfar, R. S. Varma, N. Rabiee, S. Iravani and S. Iravani, Advanced nanosystems for cancer therapeutics: A review, *ACS Appl. Nano Mater.*, 2023, **6**, 7123–7149.
11 L. Tang, Q. Xiao, Y. Mei, S. He, Z. Zhang, R. Wang and W. Wang, Insights on functionalized carbon nanotubes for cancer theranostics, *J. Nanobiotechnol.*, 2021, **19**, 423.
12 K. J. Hughes, K. A. Iyer, R. E. Bird, J. Ivanov, S. Banerjee, G. Georges and Q. A. Zhou, Review of carbon nanotube research and development: materials and emerging applications, *ACS Appl. Nano Mater.*, 2024, **7**, 18695–18713.
13 Z. Zhang, Y. Du, X. Shi, K. Wang, Q. Qu, Q. Liang, X. Ma, K. He, C. Chi, J. Tang, B. Liu, J. Ji, J. Wang, J. Dong, Z. Hu and J. Tian, NIR-II light in clinical oncology: opportunities and challenges, *Nat. Rev. Clin. Oncol.*, 2024, **21**, 449–467.
14 D. Dolmans, D. Fukumura and R. K. Jain, Photodynamic therapy for cancer, *Nat. Rev. Cancer*, 2003, **3**, 380–387.
15 S. S. Lucky, K. C. Soo and Y. Zhang, Nanoparticles in photodynamic therapy, *Chem. Rev.*, 2015, **115**, 1990–2042.
16 B. O. Murjani, P. S. Kadu, M. Bansod, S. S. Vaidya and M. D. Yadav, Carbon nanotubes in biomedical applications: current status, promises, and challenges, *Carbon Lett.*, 2022, **32**, 1207–1226.
17 X. Li, J. F. Lovell, J. Yoon and X. Chen, Clinical development and potential of photothermal and photodynamic therapies for cancer, *Nat. Rev. Clin. Oncol.*, 2020, **17**, 657–674.
18 Y. Cai, T. Chai, W. Nguyen, J. Liu, E. Xiao, X. Ran, Y. Ran, D. Du, W. Chen and X. Chen, Phototherapy in cancer treatment: strategies and challenges, *Signal Transduction Targeted Ther.*, 2025, **10**, 115.
19 C. Cao, X. Wang, N. Yang and X. Song, Recent advances of cancer chemodynamic therapy based on Fenton/Fenton-like chemistry, *Chem. Sci.*, 2022, **13**, 863–889.
20 H. Sies, V. V. Belousov, N. S. Chandel, M. J. Davies, D. P. Jones, G. E. Mann, M. P. Murphy, M. Yamamoto and C. Winterbourn, Defining roles of specific reactive oxygen species (ROS) in cell biology and physiology, *Nat. Rev. Mol. Cell Biol.*, 2022, **23**, 499–515.
21 J. Liu, X. Han, T. Zhang, K. Tian, Z. Li and F. Luo, Reactive oxygen species (ROS) scavenging biomaterials for anti-inflammatory diseases: from mechanism to therapy, *J. Hematol. Oncol.*, 2023, **16**, 116.
22 D. Kračun, L. R. Lopes, E. Cifuentes-Pagano and P. J. Pagano, NADPH oxidases: redox regulation of cell homeostasis and disease, *Physiol. Rev.*, 2025, **105**, 1291–1428.
23 G. S. Shadel and T. L. Horvath, *Cell*, 2015, **163**, 560–569.
24 A. J. P. O. de Almeida, J. C. P. L. de Oliveira, L. V. D. S. Pontes, J. F. D. S. Júnior, T. A. F. Gonçalves, S. H. Dantas, M. S. D. A. Feitosa, A. O. Silva and I. A. de Medeiros, ROS: Basic concepts, sources, cellular signaling, and its implications in aging pathways, *Oxid. Med. Cell. Longevity*, 2022, **23**, 1225578.
25 L. Milkovic, A. Cipak Gasparovic, M. Cindric, P.-A. Mouthuy and N. Zarkovic, Short overview of ROS as cell function regulators and their implications in therapy concepts, *Cells*, 2019, **8**, 793.
26 Y. Yu, S. Liu, L. Yang, P. Song, Z. Liu, X. Liu, X. Yan and Q. Dong, Roles of reactive oxygen species in inflammation and cancer, *Med. Commun.*, 2024, **5**, e519.
27 H. A. Ju, D. T. Hoang, W. S. Jang, Y. H. Kim, E. B. Park, S. H. Yang, K. Ihm, J. H. Jang, Y. M. Kim and H. Lee, Interfacial contact-driven enhanced environmental photocatalysis of CdS-loaded OH-functionalized carbon nanotubes with low biotoxicity, *Appl. Surf. Sci.*, 2025, **679**, 161197.
28 S. Zhang, F. Rong, S. Huang, S. Zhao, M. Wang, L. He, Z. Zhang and M. Du, Atomic Fe sites embedded within carbon nanotubes for the efficient photodegradation of multiple tetracyclines, *Sep. Purif. Technol.*, 2022, **287**, 120530.
29 Z.-H. Zheng, F.-X. Xiao and Y.-J. Xu, Atomically precise alloy nanoclusters steering photocatalysis, *Small*, 2025, **21**, 2505282.

 This journal is © The Royal Society of Chemistry 2025

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

30 S. S. Thind, M. Paul, J. B. Hayden, A. Joshi, D. Goodlett and J. S. McIndoe, A highly efficient photocatalytic system for environmental applications based on TiO_2 nanomaterials, *Ind. Chem. Mater.*, 2023, **1**, 431.

31 S. Xu, H. Chen, C. Li, R. Nie, Y. Yang, M. Zhou, X. Zhang and H. Zhou, Effect of Ti doping on the structural and electrochemical performance of O3-type Na $(Ni_{0.3}Fe_{0.2}Mn_{0.5})_{1-x}Ti_xO_2$ cathode materials for sodium-ion batteries, *J. Alloys Compd.*, 2023, **962**, 171199.

32 Z. Gao, B. Yao, L. Ji and T. Xu, Effect of reducing agent $NaBH_4$ on photocatalytic properties of $Bi/BiOBr/Bi_2WO_6$ composites, *Surf. Interfaces*, 2019, **4**, 10065–10071.

33 M. P. de la Flor, R. Camarillo, F. Martínez, C. Jiménez, R. Quiles and J. Rincón, Synthesis and characterization of TiO_2/CNT/Pd: An effective sunlight photocatalyst for neonicotinoids degradation, *J. Environ. Chem. Eng.*, 2021, **9**, 106278.

34 N. B. Trinh, T. A. Nguyen, S. V. Vu, H. G. T. Vo, T. N. H. Lo, I. Park and K. Q. Vo, Modified hydrothermal method for synthesizing titanium dioxide-decorated multiwalled carbon nanotube nanocomposites for the solar-driven photocatalytic degradation of dyes, *RSC Adv.*, 2024, **14**, 34037–34050.

35 S. Nguyen, C. B. Flores, M. J. Madou, M. Ray, A. Salazar, R. Vargas, I. Aguilar, N. E. Ornelas, A. Torres-Castro and S. O. Martínez, Synthesis and characterization of hierarchical suspended carbon fiber structures decorated with carbon nanotubes, *J. Mater. Sci.*, 2024, **59**, 2893–2906.

36 P. Boonchata, N. Boontanon, V. Jindal, H. K. Z. Z. Aung, C. Visvanathan, S. Fujii and S. K. Boontanon, Photocatalytic degradation of perfluorooctane sulfonic acid and perfluorooctanoic acid using titanium dioxide/graphene oxide nanocomposite immobilized on polyvinyl alcohol film, *Case Stud. Chem. Environ. Eng.*, 2024, **10**, 100862.

37 T. Yui, Y. Tamaki, K. Sekizawa and O. Ishitani, Photocatalytic reduction of CO_2: from molecules to semiconductors, *Top. Curr. Chem.*, 2011, **303**, 151–184.

38 R. P. Gandhiraman, D. Nordlund, C. Javier, J. E. Koehne, B. Chen and M. Meyyappan, X-ray absorption study of graphene oxide and transition metal oxide nanocomposites, *J. Phys. Chem. C*, 2014, **118**, 18706–18712.

39 V. A. Coleman, R. Knut, O. Karis, H. Grennberg, U. Jansson, R. A. Quinlan, B. C. Holloway, B. Sanyal and O. Eriksson, Defect formation in graphene nanosheets by acid treatment: an X-ray absorption spectroscopy and density functional theory study, *J. Phys. D: Appl. Phys.*, 2008, **41**, 062001.

40 F. Frati, M. O. J. Y. Hunault and F. M. F. de Groot, Oxygen K-edge X-ray absorption spectra, *Chem. Rev.*, 2020, **120**, 4056–4110.

41 W. Fang, M. Xing and J. Zhang, A new approach to prepare Ti^{3+} self-doped TiO2 via NaBH4 reduction and hydrochloric acid treatment, *Appl. Catal., B*, 2014, **160**, 240–246.

42 A. V. Nimmy, V. M. Anandakumar and V. Biju, Enhancing the visible-light sensitive photocatalysis of anatase TiO_2 through surface-modification, *Discover Mater.*, 2025, **5**, 45.

43 Y. Stubrov, A. Nikolenko, V. Gubanov and V. Strelchuk, Manifestation of structure of electron bands in double-resonant raman spectra of single-walled carbon nanotubes, *Nanoscale Res. Lett.*, 2016, **11**, 1–5.

44 K. Tang, Y. Li, X. Sun, J. Wang, Z. Sui and X. Xu, Oxygen vacancy induced superior catalytic performance in Nd–Ce confined inside carbon nanotubes for N_2O–assisted dehydrogenation of ethylbenzene, *J. Colloid Interface Sci.*, 2025, **678**, 658–668.

45 A. R. Yan, X. B. Wang, L. Zhu, X. S. Liu and Z. Y. Wang, Oxygen vacancy strategy enhancing the performance of TiO2/CNT supported ultrafine Pt catalyst for the oxygen reduction reaction, *J. Electroanal. Chem.*, 2024, **974**, 118734.

46 X. Yang, Z. Xiang, J. Wu, P. He, Y. Qi, S. Chen, Y. Lu, R. Xie, Z. Liu and C. He, Research progress on the construction of oxygen vacancy defects and application in photocatalytic oxidation of elemental mercury, *Fuel*, 2024, **355**, 129440.

47 F. Gunkel, D. V. Christensen, Y. Z. Chen and N. Pryds, Oxygen vacancies: The (in)visible friend of oxide electronics, *Appl. Phys. Lett.*, 2020, **116**, 120505.

48 A. Marikutsa, M. Rumyantseva, E. A. Konstantinova and A. Gaskov, The key role of active sites in the development of selective metal oxide sensor materials, *Sensors*, 2021, **21**, 2554.

49 U. T. Khatoon, A. Velidandi and G. N. Rao, Sodium borohydride mediated synthesis of nano-sized silver particles: Their characterization, anti-microbial and cytotoxicity studies, *Mater. Chem. Phys.*, 2023, **294**, 126997.

50 I. Lisiecki, The drastic influence of the $NaBH_4$ reducing agent on the growth of cobalt nanoparticles and their 3D assemblies, *Colloids Surf. A Physicochem. Eng. Asp.*, 2023, **676**, 132145.

51 J. Shen, D. Xu, J. Ji, Q. Zhang and X. Fan, In situ evolved defective TiO2 as robust support for CoB-catalyzed hydrolysis of $NaBH_4$, *Int. J. Hydrogen Energy*, 2023, **48**, 1001–1010.

52 A. Paksoy, S. F. Kurtoğlu, A. K. Dizaji, Z. Altıntaş, S. Khoshsima and A. Uzun, Ö. Balcı, Nanocrystalline cobalt–nickel–boron (metal boride) catalysts for efficient hydrogen production from the hydrolysis of sodium borohydride, *Int. J. Hydrogen Energy*, 2021, **46**, 7974–7988.

53 J. C. Barreto, G. S. Smith, N. H. P. Strobel, P. A. McQuillin and T. A. Miller, Terephthalic acid: a dosimeter for the detection of hydroxyl radicals in vitro, *Life Sci.*, 1995, **56**, 89–96.

54 R. Oulton, J. P. Haase, S. Kaalberg, C. T. Redmond, M. J. Nalbandian and D. M. Cwiertny, Hydroxyl radical formation during ozonation of multiwalled carbon nanotubes: performance optimization and demonstration of a reactive CNT filter, *Environ. Sci. Technol.*, 2015, **49**, 3687–3697.

55 X. Xu, Y. Pang and X. Fan, Mitochondria in oxidative stress, inflammation and aging: from mechanisms to therapeutic advances, *Signal Transduction Targeted Ther.*, 2025, **10**, 190.

56 C. L. Kuo, A. P. Babuharisankar, Y. C. Lin, H. W. Lien, Y. K. Lo, H. Y. Chou, V. Tangeda, L. C. Cheng, A. N. Cheng and A. Y. L. Lee, Mitochondrial oxidative stress in the tumor microenvironment and cancer immunoescape: foe or friend?, *J. Biomed. Sci.*, 2022, **29**, 74.

Published on 13 October 2025. Downloaded by Ewha Womans University on 11/14/2025 6:32:15 AM.

57 E. Vringer and S. W. G. Tait, Mitochondria and cell death-associated inflammation, *Cell Death Differ.*, 2023, **30**, 304–312.
58 K. Newton, A. Strasser, N. Kayagaki and V. M. Dixit, Cell death, *Cell*, 2024, **187**, 235–256.
59 G. Zhang, J. Wang, Z. Zhao, T. Xin, X. Fan, Q. Shen, A. Raheem, C. R. Lee, H. Jiang and J. Ding, Regulated necrosis, a proinflammatory cell death, potentially counteracts pathogenic infections, *Cell Death Dis.*, 2022, **13**, 637.
60 S. Wang, S. Guo, J. Guo, Q. Du, C. Wu, Y. Wu and Y. Zhang, Cell death pathways: molecular mechanisms and therapeutic targets for cancer, *MedComm*, 2024, **5**, e693.
61 X. Tong, R. Tang, M. Xiao, J. Xu, W. Wang, B. Zhang, J. Liu, X. Yu and S. Shi, Targeting cell death pathways for cancer therapy: recent developments in necroptosis, pyroptosis, ferroptosis, and cuproptosis research, *J. Hematol. Oncol.*, 2022, **15**, 174.
62 S. S. Ashrafi and B. Maleki, Catalytic conversion of biomass-derived 5-hydroxymethylfurfural to 5-hydroxymethyl-2-furancarboxylic acid using novel cobalt-based MOF in the presence of deep eutectic solvents, *Sci. Rep.*, 2024, **14**, 31384.
63 M. Sajid, X. Zhao and D. Liu, Production of 2,5-furandicarboxylic acid (FDCA) from 5-hydroxymethylfurfural (HMF): recent progress focusing on the chemical-catalytic routes, *Green Chem.*, 2018, **20**, 5427–5453.
64 R. Tamilselvi, S. Nandhini, E. Muniyandi, P. Saravanakumar, A. Venkatesh and V. Prakash, Cytotoxicity of Metals and Metal Oxides Nanoparticles in Dentistry: A Comprehensive Review, *Biomed. Pharmacol. J.*, 2025, **18**, 471–481.
65 M. Iyer, U. Anand, S. Thiruvenkataswamy, H. W. S. Babu, A. Narayanasamy, V. K. Prajapati, C. K. Tiwari, A. V. Gopalakrishnan, E. Bontempi, C. Sonne, D. Barceló and B. Vellingiri, A review of chromium (Cr) epigenetic toxicity and health hazards, *Sci. Total Environ.*, 2023, **882**, 163483.

 This journal is © The Royal Society of Chemistry 2025

Chemical Engineering Journal 523 (2025) 168457

Contents lists available at ScienceDirect

Chemical Engineering Journal

journal homepage: www.elsevier.com/locate/cej

Enhanced visible-light photocatalytic conversion of biomass by calcination of Ir-doped SnO_2 nanoparticles with a narrowed bandgap using $NaBH_4$

Sunyoung Hwang [a,1], Jeonghyeon Kim [b,1], Heewon Kim [a], Jaeyoon Baik [c], Min-Cheol Kim [a,*], Hangil Lee [d,*]

[a] Department of Chemistry, Sookmyung Women's University, Seoul, 04310, Republic of Korea
[b] Hydrogen Fuel Cell Research Center, Korea Institute of Science and Technology (KIST), Seoul, 02792, Republic of Korea
[c] Beamline Research Division, Pohang Accelerator Laboratory (PAL), POSTECH, Pohang, 37673, Republic of Korea
[d] Department of Chemistry and Research Institute of Natural Sciences, Sookmyung Women's University, Seoul, 04310, Republic of Korea

ARTICLE INFO

Keywords:
Surface oxygen vacancy
$NaBH_4$ calcined Ir-doped SnO_2 NPs
Photocatalytic activity
X-ray photoelectron spectroscopy
DFT calculation

ABSTRACT

SnO_2 nanoparticles (NPs) exhibit a photocatalytic activity under UV light. However, their limited response to visible light restricts their applications, highlighting the need for new methodologies. We demonstrate Ir-doping to SnO_2 only show a slight increase to visible light response and photocatalytic activity. However, treating the Ir-doped SnO_2 with $NaBH_4$ calcination dramatically improves the visible light response and photocatalytic activity. The additional $NaBH_4$-assisted calcination further increased the O_V density and narrowed the bandgap of Ir ion-doped SnO_2 NPs, enabling visible-light-driven photocatalytic oxidation of 5-hydroxymethylfurfural and toluene. This study provides a novel approach to optimize the defect engineering of SnO_2-based catalysts through Ir doping and $NaBH_4$ calcination. The resulting increase in O_V density and band structure modification were well demonstrated by spectroscopic analyses and DFT calculations. The optimized catalyst exhibited outstanding performance under visible light irradiation, achieving >90 % HMF and > 75 % toluene photoconversion into 2,5-furandicarboxylic acid and benzaldehyde, respectively, while maintaining excellent recyclability. These findings advance the development of multifunctional catalysts and are expected to inspire the design of next-generation catalysts for sustainable biomass transformation.

1. Introduction

The development of new photocatalytic materials for energy conversion reactions is crucial for ensuring stable and sustainable energy supplies [1–3]. Semiconductor metal oxide nanoparticles (NPs), such as SnO_2, TiO_2, and CeO_2 NPs, are efficient photocatalysts under ultraviolet (UV) light [4,5]. Especially, SnO_2 NPs are widely studied as photocatalysts due to their high chemical stability, non-toxicity, and high electron mobility. However, wide bandgap (~3.6 eV) of SnO_2 NPs, which allows UV light absorption to generate electron-hole pairs, limits their visible light application [6]. To overcome this limitation [7,8], doping SnO_2 NPs with metal ions has emerged as an effective strategy to improve their photocatalytic properties. Metal ion doping can create oxygen vacancies (O_V) on the surface, which modulate the local valence states of neighboring cations at the atomic scale and serve as mobile electron donors to enhance the photocatalytic activity in the visible light region [9].

In this context, SnO_2 NPs were first doped with Ir^{4+} ion to obtain Ir_{dp}-SnO_2 NPs. The ionic radius of Ir^{4+} (0.625 Å) is comparable to that of Sn^{4+} (0.690 Å), both in six-fold coordination with oxygen. Thus, the close match in ionic sizes minimizes lattice distortion, allowing Ir^{4+} ion to be incorporated into the crystal structure without significantly disrupting the host matrix. As a result, the Ir ion doped SnO_2 NPs maintains structural stability while exhibiting enhanced photocatalytic performance. However, despite the doping of Ir ions, photocatalytic properties in the visible light region were not observed due to the wide bandgap. To improve this, the surface of Ir-doped SnO_2 NPs was reduced using sodium borohydride ($NaBH_4$), which has effective reducing properties, resulting in the synthesis of r-Ir_{dp}-SnO_2 NPs [10]. In the presence of Ir ion, borohydride ions were hydrolyzed in aqueous solutions to produce hydrogen and O_V according to Eq. (1):

* Corresponding authors.
E-mail addresses: mincheolkim@sookmyung.ac.kr (M.-C. Kim), easyscan@sookmyung.ac.kr (H. Lee).
[1] These authors contributed equally to this manuscript.

https://doi.org/10.1016/j.cej.2025.168457
Received 3 July 2025; Received in revised form 25 August 2025; Accepted 13 September 2025
Available online 16 September 2025
1385-8947/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

$$Ir_{dp} - SnO_2 + Na^+ + BH_4^- + 2H_2O \rightarrow Ir_{dp} - SnO_{2-x} + Na^+ + BO_2^- + 4H_2 \quad (1)$$

In this reaction, the doped Ir ions are converted into dark precipitates known as metal borides. Thus, the key step in the reduction process is electron transfer from the borohydride to the Ir ions, which leads to the formation of O_V on the surface of SnO_2 NPs [11]. Electronic structure analysis revealed a significant increase in the O_V induced by the redox reaction of the doped Ir ions and $NaBH_4$ reduction within the SnO_2 NPs. Further analysis showed that the bandgap of the SnO_2 NPs was also significantly narrowed by $NaBH_4$ calcination, which facilitated the light absorption under visible light irradiation. The photocurrent density was substantially increased, which was attributed to the effective trapping of photogenerated electrons at defect sites created during the calcination process. The formation of new electron donor states hindered the recombination of photogenerated electron-hole pairs, thereby enhancing the photocatalytic efficiency of the r-Ir_{dp}-SnO_2 NPs.

To address the versatility of the SnO_2-based photocatalysts under visible light, two potential applications in photocatalytic conversion reactions (PCR) involving 5-hydroxymethylfurfural (HMF) and toluene were investigated. HMF is a biomass-derived chemical that can be oxidized to 2,5-furandicarboxylic acid (FDCA), providing a green alternative to petroleum-based terephthalic acid [12]. Toluene requires degradation owing to its high volatility, environmental persistence, and toxic effects on ecosystems [13]. Here, r-Ir_{dp}-SnO_2 NPs demonstrated multifunctionality as visible-light-driven photocatalysts, achieving the superior oxidation of HMF (90 % conversion) and toluene (75 % conversion). This broad and enhanced photocatalytic activity compared to pristine SnO_2 and Ir_{dp}-SnO_2 NPs (0 % and 21 % for HMF degradation and 0 % and 17 % for toluene degradation), is attributed to the formation of O_V and lattice contraction at the NP surface. This study provides insights into the optimal surface structural modifications of photocatalytic materials, highlighting their potential for advancing visible-light-driven photocatalysts for sustainable and efficient energy applications. These modified SnO_2 NPs can serve as photocatalysts for the photochemical conversion of biomass into fuels.

In this study, we focused on modifying the SnO_2 surface by doping it with Ir ions and reducing it using additional $NaBH_4$. We confirmed that the Ov density could be significantly increased through simple Ir^{4+} ion doping and additional $NaBH_4$ reduction. Interestingly, a narrow bandgap required for visible-light photocatalysis was achieved through $NaBH_4$ reduction. These findings contribute to the advancement of NP-based catalysis and may guide the design of sustainable catalysts. We believe that Ir-ion doping and additional $NaBH_4$ calcination provide a new route for optimizing defect-engineered oxide-based visible-light-active photocatalysts, enabling more efficient biomass conversion.

2. Experimental

2.1. Sample preparation

Tin (IV) chloride pentahydrate ($SnCl_4 \bullet 5H_2O$, 98 %) was added to 2-methoxyethanol ($\geq$ 99.9 %) to prepare SnO_2 NPs [14]. Under high-temperature and high-pressure hydrothermal conditions, $Sn(OH)_4$ directly crystallized as follows: $Sn^{4+} + 4OH^- \rightarrow Sn(OH)_4 \rightarrow SnO_2$ crystalline. After SnO_2 NPs were fabricated, a desired amount (5 wt%) of Ir-ion dopant was added in the form of potassium hexachloroiridate (K_2IrCl_6) (Sigma Aldrich, 99 %). The homogeneous and transparent solutions were transferred to a Teflon-lined autoclave, which was then sealed and heated at 220 °C for 10 h in a convection oven. The resulting Ir_{dp}-SnO_2 NPs were filtered and washed with double-distilled water (DDW) to remove any residues. To obtain r-Ir_{dp}-SnO_2 NPs, the Ir_{dp}-SnO_2 NPs were mixed with 10 wt% $NaBH_4$ and transferred to a tube furnace (Lab House). The mixture was heated at a ramp rate of 12.5 °C/min to 400 °C and held at 400 °C for 1 h under Ar-purging conditions. The product was then washed 10 times with DDW and 10 times with ethanol to remove contaminants such as Na^+ and BO_2^- ions, followed by drying at 95 °C.

2.2. Characterizations

2.2.1. Structural characterization

Atomic-scale images of the three SnO_2 NPs were obtained using a JEM-ARM200F transmission electron microscopy (TEM, NEOARM, JEOL Ltd., Japan). The X-ray diffraction (XRD) patterns of the three NPs were acquired using a D8 Advance TRIO/TWIN diffractometer (Bruker, USA) with Ni-filtered Cu Kα radiation.

2.2.2. Surface and chemical analysis

Electron paramagnetic resonance (EPR, JES-X320, JEOL Ltd.) spectra were acquired at 300 K with a frequency of 9446 MHz, a power of 10 mW, a modulation frequency of 100 kHz, and a width of 0.4 mT in the range of 300–400 mT to detect oxygen vacancies. Nitrogen (N_2) adsorption–desorption isotherms were obtained at 77 K using a BELSORP-Max system (BEL) for surface area measurement. X-ray photoelectron spectroscopy (XPS) measurements were performed for Ir 4*f*, Sn 3*d*, and O 1 *s* core-level spectra, and X-ray absorption spectroscopy (XAS) was conducted for Sn *M*-edge and O *K*-edge at Beamline 8A1 of the Pohang Accelerator Laboratory (PAL, Republic of Korea).

2.2.3. Optical and electronic properties

UV–visible diffuse reflectance spectra (UV–Vis DRS) were measured with a Shimadzu UV-2600i spectrophotometer equipped with an integrating sphere. Photocurrent densities were obtained using a multichannel potentiostat (WonATech) in a two-electrode configuration, with the working electrode prepared by drop-casting the photocatalyst dispersion onto fluorine-doped tin oxide (FTO) glass and a Pt gauze (0.8 × 0.8 cm) used as the counter electrode. An LED lamp (6 W, 445 nm, 230 mW• cm^{-2}; Solis High-Power LED, Thorlabs, Inc., Newton, NJ, USA) served as the light source.

2.3. Photocatalytic conversion activities of HMF and toluene

The photocatalytic conversion activities of the three SnO_2 NPs were determined by the oxidation of 5-hydroxymethylfurfural (HMF, Sigma-Aldrich, 99 %) and toluene (Sigma-Aldrich, 99 %). The prepared NPs (0.015 g) were dispersed in distilled water (30 mL) by sonication. This aqueous suspension was stirred for 2 h to allow equilibrium adsorption of HMF or toluene on the NPs. A 6 W LED lamp equipped with a 10 cm IR water filter and a cutoff filter ($\lambda = 445$ nm) was used as the light source with an illumination intensity of 230 mW/cm^2 (Solis High-Power LED, Thorlabs, Inc., Newton, NJ, USA). The PCR products of HMF and toluene were analyzed using high-performance liquid chromatography (HPLC; LC-20AD pump, Shimadzu).

The conversion of HMF (or toluene) and the productivity of FDCA (or benzaldehyde, BAD) were determined using Eqs. (2) and (3), respectively.

$$\text{HMF (or toluene) conversion [\%]} = \frac{\text{moles of HMF (or toluene) consumed}}{\text{moles of initial HMF (or toluene)}} \times 100 \quad (2)$$

$$\text{Product (FDCA or BAD) Yield [\%]} = \frac{\text{moles of Products formed}}{\text{moles of initial HMF (or toluene)}} \quad (3)$$

2.4. Density functional theory simulations

Density functional theory (DFT) calculations were performed using the Vienna Ab Initio Simulation Package (VASP) [15], employing the Perdew–Burke–Ernzerhof (PBE) [16] functional for exchange correlation, along with D3 dispersion correction [17] and the rotationally

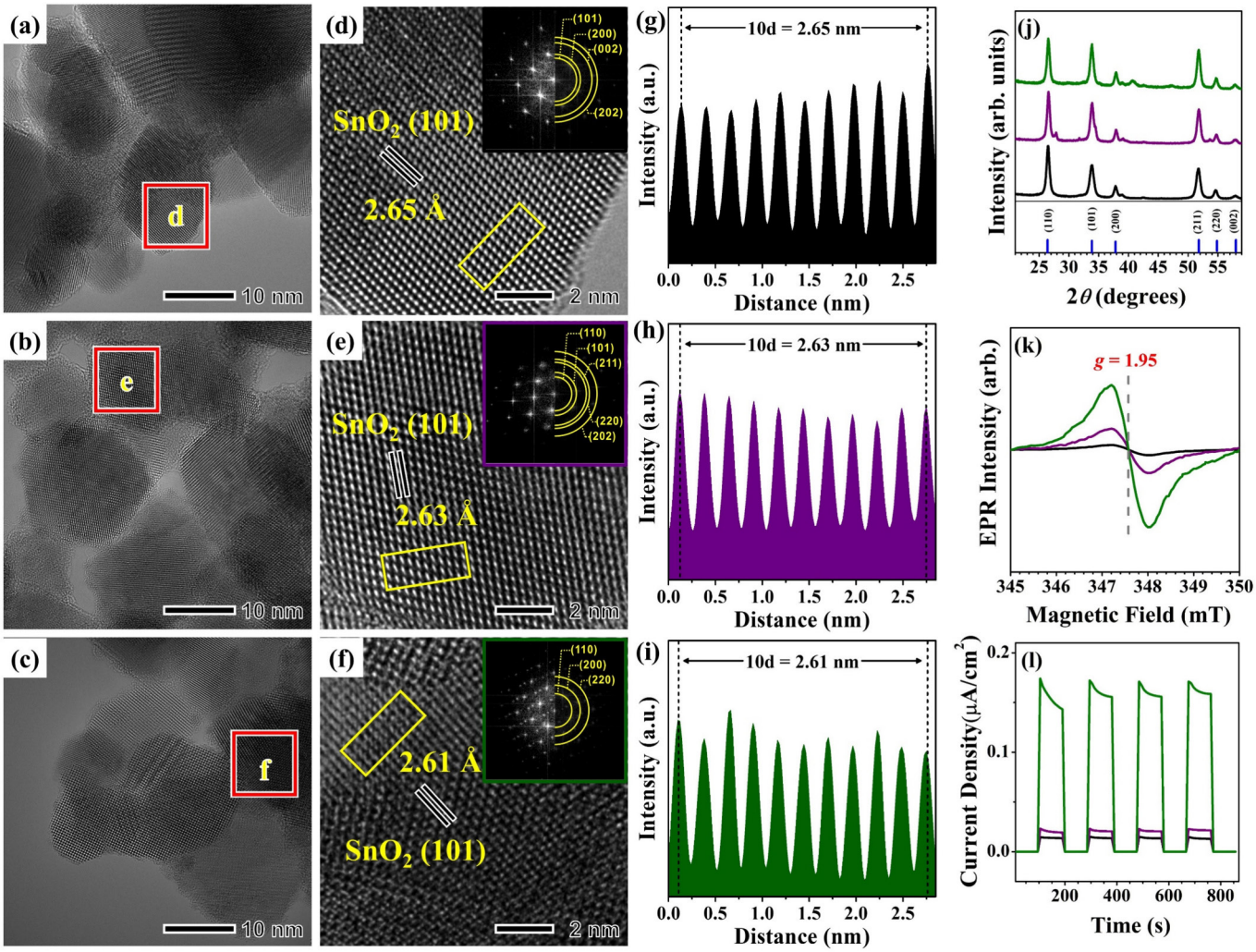


Fig. 1. (a-c) HRTEM images and (d-f) their magnified views taken from the red boxes for (a,d) SnO_2, (b,e) Ir_{dp}-SnO_2, and (c,f) r-Ir_{dp}-SnO_2 NPs. The insets of (d)-(f) show the corresponding fast Fourier transform patterns. (g-i) Intensity profiles measured from the yellow rectangular boxed in HRTEM images of (d) SnO_2, (e) Ir_{dp}-SnO_2, and (f) r-Ir_{dp}-SnO_2 NPs. (j) XRD, (k) EPR, and (l) Photocurrent measurement results of SnO_2 (black), Ir_{dp}-SnO_2 (purple), and r-Ir_{dp}-SnO_2 (green) NPs.

invariant DFT + U scheme proposed by Dudarev et al. [18]. The corresponding values of the U parameter were obtained from existing literature: U = 4.0 eV for the Sn 4*d* electrons and U = 2.0 eV for the Ir 5*d* electrons. The projector-augmented-wave (PAW) method was used to describe the potential from the ionic core, with an energy cutoff of 400 eV. Monkhorst-Pack *k*-point sampling with *k*-point spacing <0.03 $Å^{-1}$ was applied for all systems. Spin polarization was included in all the calculations, and dipole corrections were included in the slab structures. All structures were relaxed until the atomic force of every atom was below 0.05 eV $Å^{-1}$, excluding the lowest atomic layer for slab structures. In addition, the modified Becke-Johnson meta-GGA functional [16] was used to simulate the band structures of the models.

3. Results and discussion

The prepared NPs were analyzed using high-resolution TEM (HRTEM) to determine the effects of Ir ion doping and the additional $NaBH_4$ calcination on their crystal structures. Fig. 1a–c showed the typical HRTEM images of the three SnO_2 NPs. The crystalline SnO_2 NPs exhibited a lattice distance of approximately 0.265 nm, which was assigned to the (101) plane of tetragonal SnO_2 (Fig. 1d, g) [19,20]. The corresponding fast Fourier transform (FFT) pattern, shown in the inset of Fig. 1d, also revealed characteristic diffraction spots corresponding to the (101), (200), (002), and (202) planes of tetragonal SnO_2, indicating the high crystallinity of the SnO_2 NPs [21]. After Ir ion doping and $NaBH_4$ calcination, no distinct structural changes were observed in the Ir_{dp}-SnO_2 and r-Ir_{dp}-SnO_2 NPs (Fig. 1b,c). However, these samples exhibited lattice compression in the intensity profiles obtained from HRTEM. Compared to the pristine SnO_2 NPs, the interlayer *d*-spacings of Ir_{dp}-SnO_2 and r-Ir_{dp}-SnO_2 NPs decreased by −20 and − 40 pm, respectively (Fig. 1e,f,h,i). The shrinkage of the interlayer spacing and the corresponding changes in the lattice might reflect the high density of O_V in the modified SnO_2 NPs, indicating the successful fabrication of highly defective SnO_2 NPs. Furthermore, a larger average reciprocal space distance was observed for r-Ir_{dp}-SnO_2 (8.47 nm^{-1}) compared to SnO_2 NPs (8.40 nm^{-1}), determined from the SnO_2(200) plane in the FFT, which also suggested lattice compression throughout the SnO_2 structure (Fig. 1d,f). Additional diffraction spots were observed in the FFT pattern of r-Ir_{dp}-SnO_2 (The inset of Fig. 1f). These spots were attributed to a further reduced state of SnO_2 by the electron beam, reflecting the reducible nature of r-Ir_{dp}-SnO_2 due to the formation of O_V on the surface [22,23]. The elemental distributions of the synthesized SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs were analyzed by energy-dispersive X-ray spectroscopy (EDX), which revealed that Sn, O, and Ir were homogeneously distributed throughout the NPs (Fig. S1).

Fig. 1j showed the XRD patterns of typical rutile-structured SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs, with the most intense peaks observed at 26.5°, 33.9°, 37.8°, 51.7°, 54.7°, and 58.0° corresponding to the (110),

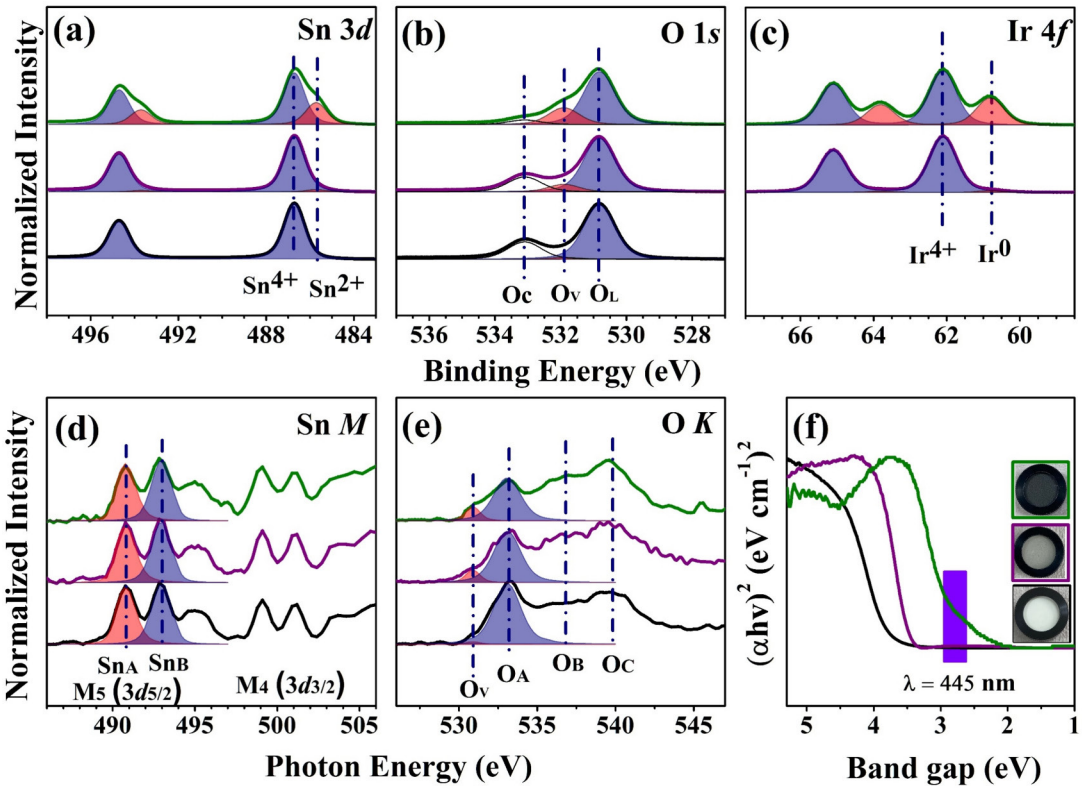


Fig. 2. XPS spectra of SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs at (a) Sn 3*d*, (b) O 1 *s*, and (c) Ir 4*f* core levels. XAS spectra of SnO_2 at (d) Sn *M* and (e) O *K* edges. (f) UV–Vis DRS spectra of SnO_2 (black), Ir_{dp}-SnO_2 (purple), and r-Ir_{dp}-SnO_2 (green) NPs.

(101), (200), (211), (220), and (002) crystal planes, respectively [24]. No significant structural crystal changes were observed for the three SnO_2 NPs, indicating that the main structure of SnO_2 remained after modification. For SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs, the diffraction peak at approximately 26.5° corresponding to the SnO_2 (110) plane exhibited the highest intensity. As shown in Fig. S2a, this peak shifted to higher angles after Ir ion doping and $NaBH_4$ calcination. These results confirm the creation of surface defect sites, such as oxygen vacancy, directly related to the catalytic properties of SnO_2 NPs [25].

The SnO_2(110) peak was also used to determine the crystalline size according to Scherrer's equation [26].

$$D = 0.9\lambda/B\cos\theta \quad (4)$$

where D is the crystal size (nm), λ is the wavelength of the X-ray radiation (0.154 nm), θ is the Bragg angle, and B is the full width at half maximum (FWHM) of the peak at 2θ. As shown in Table S1, the crystalline sizes of SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs showed no significant difference. Therefore, the contribution of crystallite size to the catalytic properties of the SnO_2-based NPs was negligible. The surface area of a catalyst is an important parameter affecting its catalytic efficiency [26]. In this study, N_2 adsorption–desorption isotherm was analyzed using the Brunauer–Emmett–Teller method. The isotherm data, shown in Fig. S3, revealed that the surface areas of SnO_2 and Ir_{dp}-SnO_2 were similar, with values of 20.9 and 22.9 m^2/g, respectively, indicating an absence of significant surface change upon Ir doping. However, the r-Ir_{dp}-SnO_2 NPs showed an increased surface area of 34.1 m^2/g, which was attributed to the formation of local surface defects, including oxygen vacancies or structural disorder, during the $NaBH_4$ calcination process [27,28].

The EPR spectra of the r-Ir_{dp}-SnO_2 NPs showed a prominent signal at a g-factor of approximately 1.95 (Fig. 1 k), indicating the presence of a single electron trapped within the O_V due to the Zeeman effect [29]. In contrast, the signal from SnO_2 was minimal, whereas a slight increase was observed for Ir_{dp}-SnO_2. These results confirmed that $NaBH_4$ calcination treatment was more effective than Ir ion doping in modifying SnO_2 NPs, primarily due to the increased formation of O_V on the surface of NPs.

To further elucidate the charge separation efficiency of the photogenerated electron-hole pairs, we investigated the photocurrents of the SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs under visible-light irradiation (Fig. 1l). To measure the transient photocurrent response, the NPs were irradiated with visible light for four on/off cycles. Upon illumination, all three SnO_2 NPs exhibited a prompt increase in photocurrent, which immediately returned to zero once the visible light was switched off. Interestingly, an almost ten-fold improvement in the current density was observed in r-Ir_{dp}-SnO_2 NPs. This result confirmed that the formation of O_V on the surface could trap electrons at specific defect sites, which inhibited the recombination of photogenerated charge carriers [30]. This enhancement was predominantly attributed to the increased surface density of O_V, rather than to changes in the bulk band structure. To prove this mechanism, similar photocurrent measurements were performed in 0.025 M Na_2SO_3, a well-known hole scavenger (Fig. S2b). All three samples exhibited nearly identical photocurrent densities, indicating that the overall light absorption and charge transport through the bulk band structure were comparable across the samples (under facile hole scavenging at the particle-electrolyte interface). More importantly, the improvement in the photocurrent due to the formation of O_V aligned well with the photocatalytic activity of the SnO_2-based NPs. From the XRD, photocurrent, and EPR results shown in Fig. 1, it is clear that $NaBH_4$ calcination increased the O_V density.

The surface chemical states in SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs were analyzed by XPS [31]. As shown in Fig. 2a, the Sn^{2+} defect peak at 485.75 eV, which was not observed in the pristine SnO_2 NPs, was distinct after Ir ion doping and $NaBH_4$ calcination. Some Sn atoms in SnO_2 were reduced to Sn^{2+} during Ir ion doping, likely due to the extraction of oxygen from SnO_2 to form IrO_2 on the surface [32].

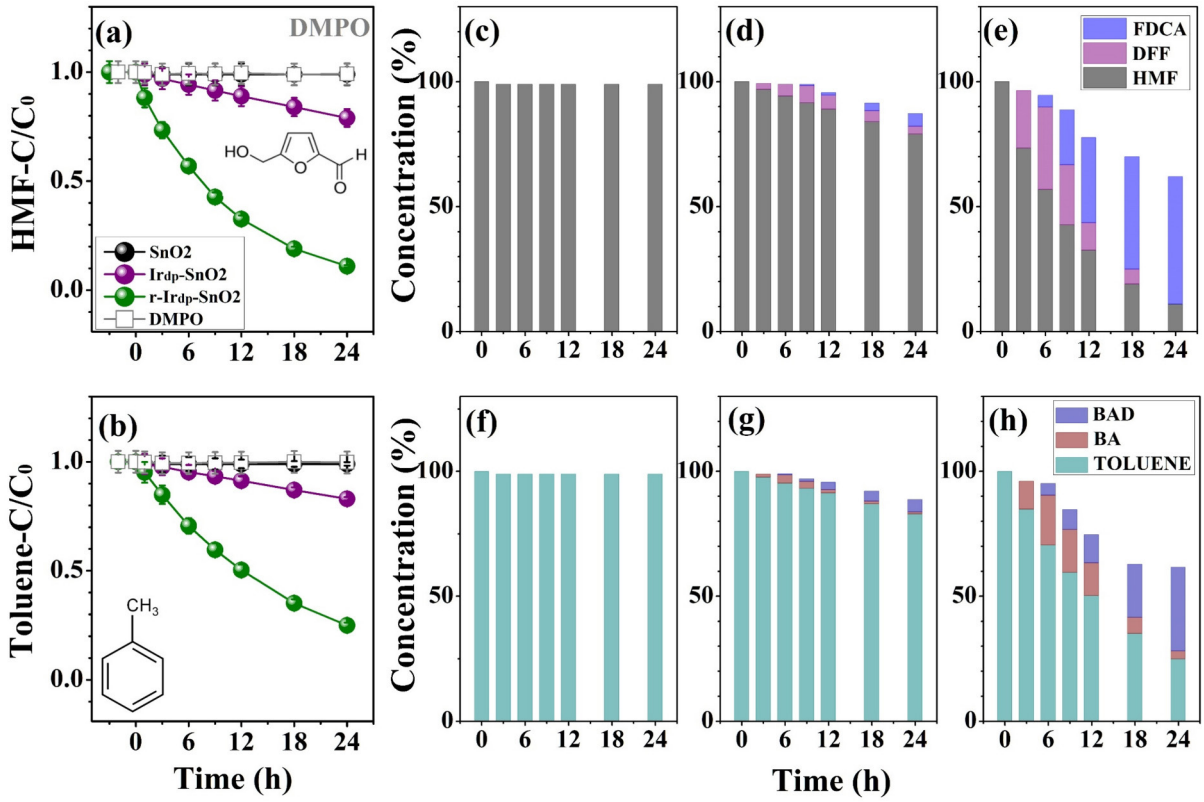


Fig. 3. Photocatalytic properties of the three SnO_2 NPs for the oxidation of (a) HMF and (b) toluene with or without DMPO (radical scavenger; gray); HPLC analysis of the oxidation products of HMF in (c) SnO_2, (d) Ir_{dp}-SnO_2, and (e) r-Ir_{dp}-SnO_2 NPs and HPLC analysis of oxidation products of toluene in (f) SnO_2, (g) Ir_{dp}-SnO_2, and (h) r-Ir_{dp}-SnO_2 NPs. Experimental conditions: [catalyst] = 0.5 g/L, λ = 445 nm, $[HMF]_0 = [toluene]_0 = 25$ mM.

Furthermore, $NaBH_4$ calcination led to the formation of additional defect structures through the removal of oxygen atoms present on the surface, consistent with the results shown in Fig. 1. The deconvoluted O 1 *s* XPS spectra of pristine SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs are shown in Fig. 2b. Three oxygen-containing components were observed, corresponding to lattice oxygen of SnO_2 (O_L, 530.8 eV), oxygen vacancies (O_V, 531.6 eV), and chemisorbed oxygen (O_C, 532.7 eV), respectively [33]. The r-Ir_{dp}-SnO_2 NPs exhibited a significantly increased intensity ratio between the most intense O_V peak and the O_L peak, indicating that the $NaBH_4$ calcination treatment generated abundant O_V on the SnO_2 surface, expected to enhance its photocatalytic activities compared to pristine SnO_2 NPs. Next, the chemical states of Ir before and after $NaBH_4$ calcination were compared. Fig. 2c showed that the IrO_2 present on the surface of the SnO_2 NPs was significantly reduced to metallic Ir by $NaBH_4$ calcination [34]. This result indicated that $NaBH_4$ calcination formed more O_V during the reduction of IrO_2 on the surface, which was also directly related to changes in photocatalytic properties. The above XPS data confirmed the formation of IrO_2 on the surface of the Ir_{dp}-SnO_2 NPs. The IrO_2 contributed to the generation of O_V and subsequently promoted the reduction of Sn^{4+} to Sn^{2+} to maintain charge balance. As a result, the number of O_V increased when Ir ions were doped on the SnO_2 NPs. The IrO_2 formed on the SnO_2 NPs was reduced to metallic Ir after $NaBH_4$ calcination, which led to the formation of additional O_V. The XPS results for the three systems showed that Ir ion doping and $NaBH_4$ reduction promoted electron transfer involving oxygen atoms on the surface of the SnO_2 NPs, induced the formation of defect structures, and affected their photocatalytic properties. Fig. S4 showed that the number of O_V and the amount of metallic Ir both increased with the concentration of $NaBH_4$, verifying the direct relationship between $NaBH_4$ calcination and O_V generation.

Figs. 2d,e showed the XAS results of the SnO_2-based NPs. Three small features between 490 and 497 eV and two small features between 498 and 502 eV were observed in the spectra at the Sn $M_{4,5}$-edges, indicating a transition from the Sn^{4+} 3*d* state to the unoccupied *p* state. The features between 490 and 497 eV and those between 498 and 502 eV were derived from the Sn M_5 $3d_{5/2}$ and M_4 $3d_{3/2}$ transitions, respectively [35]. The Sn *M*-edge measurements of all three SnO_2 samples were similar. Comparing the intensity ratio (S_{nA}/S_{nB}) of the Sn$3d_{5/2}$ S_{nA} and S_{nB} peaks, a slight decrease was observed in the case of r-Ir_{dp}-SnO_2 NP, which can also be interpreted as evidence of increased defect structure. In contrast, the O *K*-edge peaks of Ir_{dp}-SnO_2 and r-Ir_{dp}-SnO_2 NPs were different from those of SnO_2 [36]. In Fig. 2e, O_A is the O 2*p*-Sn 5 *s* transition peak, and O_B and O_C are the O 2*p*-Sn 5*p* transition peaks. In particular, the intensity of the peak due to the O_V near 531 eV was greatly enhanced in Ir_{dp}-SnO_2 and r-Ir_{dp}-SnO_2 NPs.

The surface electronic structures of the SnO_2-based NPs were investigated by measuring their bandgaps in the UV–Vis diffuse reflectance spectroscopy (DRS) spectra [37]. As shown in Fig. 2f, the bandgap energy of the modified SnO_2 NPs differed markedly from that of pristine SnO_2 NPs. Due to its wide bandgap (3.82 eV), pristine SnO_2 could not absorb light in the visible range [38]. Ir ion doping reduced the bandgap of SnO_2 to 3.44 eV. Although this bandgap narrowing, along with the increased Ov concentration, has been reported to improve the photocatalytic activity of SnO_2 in the UV region, it remains insufficient to activate the photocatalytic response in the visible region [39]. Interestingly, in the r-Ir_{dp}-SnO_2 NPs, a surprisingly narrow bandgap (2.03 eV) was observed. The bandgap was sufficiently narrowed to enable visible-light activity, allowing a meaningful photocatalytic response at a wavelength of 445 nm ($E_g = 2.79$ eV). The DRS spectra clearly show that $NaBH_4$ calcination effectively narrows the bandgap, thereby enhancing visible-light-driven photocatalysis.

X-ray fluorescence analysis (XRF) was performed to evaluate whether the doping concentration was maintained after the photocatalytic reaction. The XRF analysis results showed that the Ir content

Table 1
C/C_0 values and the percentages of final products for HMF and toluene degradation in the three SnO_2-photocatalyzed systems after 445 nm irradiation for 24 h.

	SnO_2 NPs	Ir_{dp}-SnO_2 NPs	r-Ir_{dp}-SnO_2 NPs
HMF (C/C_0)	0.99 ± 0.05	0.79 ± 0.04	0.09 ± 0.005
FDCA (%)	0.0	5.0 ± 0.25	51.0 ± 2.55
toluene (C/C_0)	0.99 ± 0.05	0.83 ± 0.01	0.25 ± 0.013
BAD (%)	0.0	4.8 ± 0.24	33.5 ± 1.68

* C and C_0 represent the time-dependent and initial concentrations of HMF and toluene, respectively.

was 4.89 wt% in Ir_{dp}-SnO_2 and 4.86 wt% in r-Ir_{dp}-SnO_2 NPs, which was found to be similar to the levels before the photoreaction experiment (4.92 wt% in Ir_{dp}-SnO_2 and 4.88 wt% in r-Ir_{dp}-SnO_2). This result explains the stability of these samples when reused (Table S2).

The photocatalytic activities of the three SnO_2 NPs were investigated under visible-light irradiation (Fig. 3). The PCR rates of HMF and toluene observed for r-Ir_{dp}-SnO_2 NPs were superior to those of pristine SnO_2 and Ir_{dp}-SnO_2 NPs (Fig. 3a,b). Specifically, the r-Ir_{dp}-SnO_2 NPs exhibited >90 % HMF and > 75 % toluene degradation efficiencies, in accordance with their superior photocatalytic performance compared to the other SnO_2 NPs under 445 nm irradiation. O_V tends to induce positive charges on the surface of SnO_2 NPs, trapping negatively charged electrons upon photogeneration. Consequently, this enhanced electron density accelerates the formation of reactive oxygen species (ROS), thereby promoting the oxidation of HMF and toluene. Therefore, maximizing the O_V concentration through $NaBH_4$ calcination without altering the SnO_2 NP structural characteristics is an effective and straightforward strategy for enhancing photocatalytic efficiency. Given that HMF contains hydroxyl (–OH) and aldehyde (–CHO) functional groups, it is important to determine the yield of the end products of HMF decomposition. Furthermore, a precise mechanistic investigation of HMF photolysis is essential to determine whether photolysis proceeds via aldehyde or alcohol oxidation [40]. Because toluene contains only a methyl ($–CH_3$) functional group, photodegradation proceeds sequentially through alcohol oxidation ($–CH_2OH$) [41]. The conversion mechanism of HMF and toluene via the photocatalytic oxidation by SnO_2 NPs can be explained as follows. The O_V on the SnO_2 surface traps electrons and O_2, thus promoting the formation of $\bullet O_2^-$. Consequently, HMF and toluene are photocatalytically oxidized through the action of $\bullet O_2^-$ and h^+. To confirm the involvement of ROS, such as $\bullet O_2^-$ in the photocatalytic reaction, 5,5-dimethyl-1-pyrroline *N*-oxide (DMPO) was used as an $\bullet O_2^-$ radical scavenger. Fig. 3a, b showed that the presence of DMPO completely inhibited the photocatalytic activity because it preferentially reacted with the radicals formed in the photoreaction, confirming our proposed mechanism.

Fig. 3c-h showed the concentration changes in the end products formed by the photolysis of HMF and toluene in the presence of the three SnO_2 NPs. The final concentrations of FDCA from HMF (Fig. 3c-e) and BAD from toluene (Fig. 3f-h) followed the same trend as the PCR efficiencies of HMF (Fig. 3a) and toluene (Fig. 3b), respectively. In particular, using r-Ir_{dp}-SnO_2 NPs within 24 h, 25 mM of FDCA or BAD was produced, corresponding to conversion yields of approximately 51 % and 33.5 % based on the initial HMF and toluene concentrations, respectively, demonstrating outstanding efficiencies under visible light irradiation. Determining the yield of the final product and the exact mechanistic pathways of the HMF and toluene photolysis are essential for evaluating the photocatalytic performance of SnO_2 NPs. HMF contains an alcohol group (–OH) and an aldehyde group (–CHO). For all three SnO_2 NPs, the $-CH_2OH$ group of HMF was first oxidized to -CHO, resulting in the selective formation of *2,5-diformylfuran* (DFF) as an intermediate and then FDCA as the final product (Fig. 3c-e) [42]. Toluene has a methyl ($–CH_3$). When oxidized, it selectively formed benzyl alcohol (BA) as an intermediate and BAD as the final product (Fig. 3f-h). In conclusion, the selective oxidation of biomass (HMF and toluene) was achieved using SnO_2-based photocatalysts dispersed in an aqueous solution. The reaction proceeded from HMF via DFF to the final product FDCA and from toluene via BA to BAD [43]. Table 1 listed the 24 h PCR profiles of HMF and toluene and their final products (FDCA and BAD) in the three SnO_2-photocatalyzed systems under 445 nm irradiation.

The PCR efficiency also depended on the concentration of $NaBH_4$ used during calcination. As shown in Fig. S5, increasing the amount of $NaBH_4$ during the calcination, improved the PCR efficiency, highlighting the important role of $NaBH_4$ calcination in enhancing photocatalytic performance.

The proposed mechanisms for the photocatalytic oxidation of HMF to FDCA and toluene to BAD using the three SnO_2 systems are shown in Scheme S1. The O_V formed on the surface of the r-Ir_{dp}-SnO_2 NPs can trap electrons and O_2 to promote the formation of $\bullet O_2^-$. As a result, HMF and toluene are converted to FDCA and BAD, respectively, via photocatalytic oxidation under the action of $\bullet O_2^-$ and h^+ [44]. In special, for SnO_2 and Ir_{dp}-SnO_2, the lack of Ov leads to insufficient amounts of generated $\bullet O_2^-$, resulting in the low oxidation rate of HMF and DFF in HMF oxidation as depicted in Scheme S2.

Moreover, the three SnO_2 NPs exhibited linear correlations between $\ln(C/C_0)$ and the reaction time; consequently, their photocatalytic degradation reactions were fitted to pseudo-first-order kinetics to determine their first-order rate constants (k) (Fig. S6). The k values of the pristine SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs for HMF oxidation were calculated to be 0.000, 0.0098, and 0.092 h^{-1}, respectively, suggesting that r-Ir_{dp}-SnO_2 NPs accelerated the reaction compared to SnO_2 and Ir_{dp}-SnO_2 NPs (Fig. S6a). The k values of the pristine SnO_2, Ir_{dp}-SnO_2, and r-Ir_{dp}-SnO_2 NPs for toluene oxidation were calculated to be 0.000, 0.0078, and 0.058 h^{-1}, respectively, confirming the significantly enhanced photocatalytic activity of r-Ir_{dp}-SnO_2 NPs (Fig. S6b). Additionally, the photocatalytic activities of the three SnO_2 NPs were maintained (> 97 %) even after 10 reuse cycles (Fig. S7).

In summary, multiple factors could contribute to the formation of O_V within SnO_2 NPs upon Ir-ion doping and/or $NaBH_4$ treatment. Owing to the lower oxidation state of the Ir ion compared to that of the host Sn ion after $NaBH_4$ calcination, stable O_V was irreversibly created to maintain

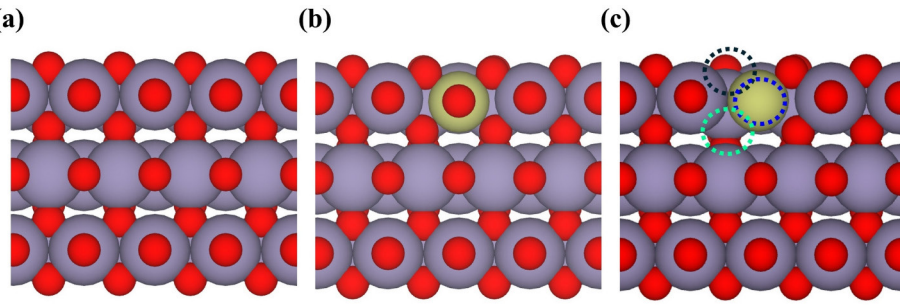
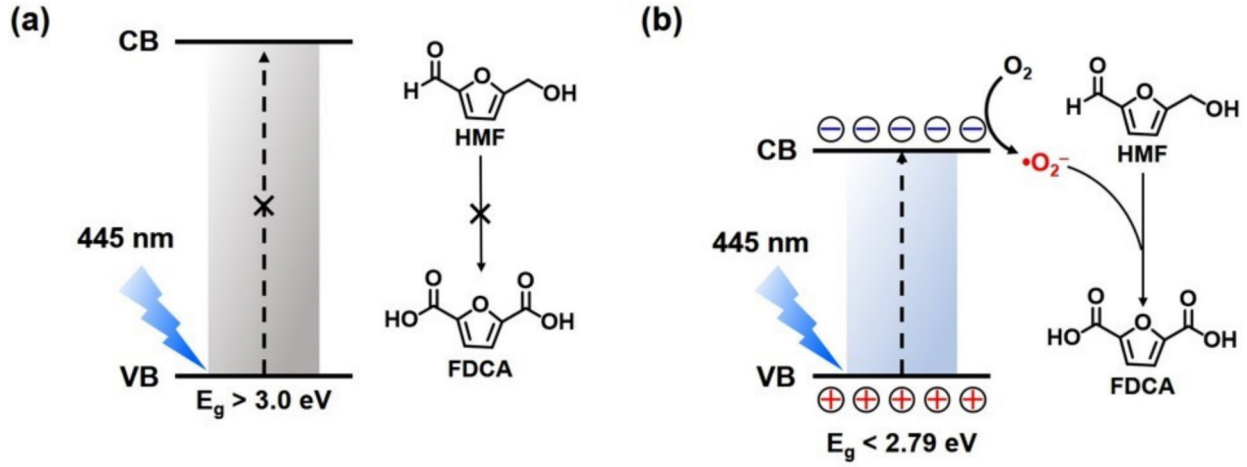


Fig. 4. Molecular structure of (a) SnO_2, (b) Ir_{dp}-SnO_2, and (c) r-Ir_{dp}-SnO_2 models. The color coding for the atoms is: Sn: purple-gray; Ir: olive; O: red. The possible Ov sites in r-Ir_{dp}-SnO_2 are depicted as dotted circles.



Scheme 1. Stages in photocatalysis for (a) SnO_2 and Ir_{dp}-SnO_2 NPs, and (b) r-Ir_{dp}-SnO_2 NPs.

the charge neutrality of the NPs according to the following equation:

$$SnO_2 + Ir^{4+}{}_{dopant} + NaBH_4 \rightarrow Ir_{at\ substituted\ Sn\ site} + O_{at\ regular\ O\ site} + O_V \quad (5)$$

DFT simulations were also performed to gain insight into the O_V formation and bandgap narrowing (Fig. 4). Based on our experimental findings, the SnO_2 structure was modeled as a (110) slab of rutile SnO_2. The Ir-doped (110) SnO_2 slab (denoted as Ir_{dp}-SnO_2) was determined to be the most stable among various relaxed Ir-doped (110) SnO_2 slab structures. The (110) SnO_2 and (110) Ir_{dp}-SnO_2 slabs with various possible O_V sites (depicted as dotted circles in Fig. 4c) were compared to determine the most stable O_V sites for SnO_2 and Ir_{dp}-SnO_2. The O_V formation energy (E_f [O_V]) was calculated as follows:

$$E_f[O_V] = E[O_V - A] - E[A] - E[O] \quad (6)$$

where E[O_V-A] is the VASP energy of the surface A (A = SnO_2 or Ir_{dp}-SnO_2) with O_V, E[A] is the VASP energy of the pristine surface, and E[O] is the DFT reference energy of oxygen (E[O] = −4.523 eV) [45].

The Ov formation energy ($E_f[O_V]$) is significantly lower for Ir_{dp}-SnO_2 (3.34–4.81 eV) than SnO_2 (3.76–5.01 eV), which clearly indicates that Ir doping promotes O_V formation. For both structures, the Ov formed on the surface (black dotted circles in Fig. 4c) showed the highest $E_f[O_V]$. In addition, the large $E_f[O_V]$ of Ir_{dp}-SnO_2 explains why the $NaBH_4$ treatment is essential for O_V formation. The electronegativity difference between Ir and O is smaller than that between Sn and O, which decreases the ionic character of the metal-O bond. Consequently, the neighboring oxygen atoms are relatively electron-deficient, making them more susceptible to reduction. This electron efficiency is further shown in Bader charge analysis, where the Ir dopant-neighboring O has fewer electrons than the Sn-neighboring O with a 0.1e^- charge deficiency. The electron-deficient Ir-neighboring O is visualized by the charge density difference between SnO_2 and Ir_{dp}-SnO_2 (Fig. S8). We also calculated the bandgap (E_g) of SnO_2 (3.88 eV) and r-Ir_{dp}-SnO_2 (2.02 eV), which are consistent with our experimental findings (Experimental bandgap of SnO_2 = 3.82 eV and that of r-Ir_{dp}-SnO_2 = 2.03 eV). For a more detailed analysis, we compare the density of states (DOS) plot of SnO_2 and r-Ir_{dp}-SnO_2 (Fig. S9). It is evident that the Ov in r-Ir_{dp}-SnO_2 introduces mid-gap states near the conduction band, resulting in a narrowed band gap. Such behavior is reported to enhance the carrier concentration under photo-response, [46] enabling visible light-responsive catalytic reactions.

Scheme 1 illustrates the different mechanisms of the visible-light-active photocatalysis of SnO_2-based NPs. The SnO_2 and Ir_{dp}-SnO_2 NPs shown in Scheme 1a exhibit poor photocatalytic properties under visible-light irradiation because of the wide bandgap. On the other hand, for the r-Ir_{dp}-SnO_2 NPs shown in Scheme 1b, its narrow bandgap enables visible-light-driven photocatalytic reactions. Therefore, the photocatalytic performance of the r-Ir_{dp}-SnO_2 NPs is primarily influenced by their optical and electronic structures. Consequently, the formation of high density O_V facilitates the PCR activity of the r-Ir_{dp}-SnO_2 NPs. This, in turn, enhances the overall PCR effect in the visible light range.

4. Conclusion

In this study, SnO_2 NPs with O_V were synthesized by Ir ion doping and subsequent $NaBH_4$ calcination. The $NaBH_4$ calcinated r-Ir_{dp}-SnO_2 NPs show significantly improved photocatalytic activity for 445 nm visible-light irradiated biomass conversion with a 90 % conversion rate for HMF oxidation and 75 % conversion rate for toluene oxidation than SnO_2 (< 1 % conversion) and Ir_{dp}-SnO_2 NPs (~20 % conversion). Through various experimental techniques and theoretical calculations, we rationalize that the improved photocatalytic activity of r-Ir_{dp}-SnO_2 NPs can only be achieved with the conjunction of $NaBH_4$ calcination and Ir dopants. The combination of Ir dopants and $NaBH_4$ calcination enables generation of abundant amounts of Ov that are responsible for the promotion of oxygen radical oxidants and the improved visible light response via band gap narrowing. Our results suggest that Ir-ion doping combined with further $NaBH_4$ calcination treatment leads to O_V modulation and bandgap narrowing, which provides innovative ways to create photocatalysts for efficient biomass conversion under visible light. We expect this strategy to be generally utilized for other widely used metal oxide photocatalysts (RuO_2, WO_3 and etc.) and/or other oxidation reaction of importance (water oxidation, ammonia oxidation, and etc.), improving the visible light photocatalytic efficiency for such systems.

CRediT authorship contribution statement

Sunyoung Hwang: Writing – original draft, Formal analysis, Data curation, Conceptualization. **Jeonghyeon Kim:** Writing – original draft, Formal analysis, Data curation. **Heewon Kim:** Formal analysis, Data curation. **Jaeyoon Baik:** Investigation, Formal analysis. **Min-Cheol Kim:** Supervision, Investigation, Formal analysis, Data curation. **Hangil Lee:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no competing financial interests.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (RS-2024-00346153, RS-2025-21072982).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2025.168457.

Data availability

No data was used for the research described in the article.

References

[1] W.M. Cheng, R. Shang, Transition metal-catalyzed organic reactions under visible light: recent developments and future perspectives, ACS Catal. 10 (2020) 9170–9196.
[2] X. Lang, X. Chen, J. Zhao, Heterogeneous visible light Photocatalysis for selective organic transformations, Chem. Soc. Rev. 43 (2014) 473–486.
[3] L. Tang, X. Meng, D. Deng, X. Bao, Confinement catalysis with 2D materials for energy conversion, Adv. Mater. 31 (2019) 1901996.
[4] L. Ran, D. Zhao, X. Gao, L. Yin, Highly crystalline Ti-doped SnO_2 hollow structured photocatalyst with enhanced photocatalytic activity for degradation of organic dyes, CrystEngComm 17 (2015) 4225–4237.
[5] S. Sharma, N. Kumar, P.R. Makgwane, N.S. Chauhan, K. Kumari, M. Rani, S. Maken, TiO_2/SnO_2 nano-composite: new insights in synthetic, structural, optical and photocatalytic aspects, Inorg. Chim. Acta 529 (2022) 120640.
[6] Z. Liu, D.D. Sun, P. Guo, J.O. Leckie, An efficient bicomponent TiO_2/SnO_2 nanofiber photocatalyst fabricated by electrospinning with a side-by-side dual spinneret method, Nano Lett. 7 (2007) 1081–1085.
[7] C. Sun, J. Yang, M. Xu, Y. Cui, W. Ren, J. Zhang, H. Zhao, B. Liang, Recent intensification strategies of SnO_2-based photocatalysts: a review, Chem. Eng. J. 427 (2022) 131564.
[8] R. Ma, S. Zhang, T. Wen, P. Gu, L. Li, G. Zhao, F. Niu, Q. Huang, Z. Tang, X. Wang, A critical review on visible-light-response CeO_2-based photocatalysts with enhanced photooxidation of organic pollutants, Catal. Today 335 (2019) 20–30.
[9] W. Jiang, H. Loh, B.Q.L. Low, H. Zhu, J. Low, J.Z.X. Heng, K.Y. Tang, Z. Li, X. J. Loh, E. Ye, Y. Xiong, Role of oxygen vacancy in metal oxides for photocatalytic CO_2 reduction, Appl. Catal. Environ. 321 (2023) 122079.
[10] Y. Wen, D. Qu, L. An, X. Gao, W. Jiang, D. Wu, D. Yang, Z. Sun, Defective g-C_3N_4 prepared by the NaBH4 reduction for high-performance H_2 production, ACS Sustain. Chem. Eng. 7 (2018) 2343–2349.
[11] K. Zhang, J.M. Suh, J.W. Choi, H.W. Jang, M. Shokouhimehr, R.S. Varma, Recent advances in the nanocatalyst-assisted $NaBH_4$ reduction of nitroaromatics in water, ACS Omega 4 (2019) 483–495.
[12] C. Li, J. Li, L. Qin, P. Yang, D.G. Vlachos, Recent advances in the photocatalytic conversion of biomass-derived Furanic compounds, ACS Catal. 11 (2021) 11336–11359.
[13] L. Sun, G. Li, S. Wan, T. An, Mechanistic study and mutagenicity assessment of intermediates in photocatalytic degradation of gaseous toluene, Chemosphere 78 (2010) 313–318.
[14] K. Vijayarangamuthu, S. Rath, Modification of the structural and optical properties of tin oxide nanoparticles by co doping and thermal annealing, Appl. Phys. A Mater. Sci. Process. 114 (2014) 1181–1188.
[15] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1996) 15–50.
[16] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865.
[17] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, J. Chem. Phys. 132 (2010) 154104.
[18] S.L. Dudarev, G.A. Botton, S.Y. Savrasov, C.J. Humphreys, A.P. Sutton, Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA1U study, Phys. Rev. B 57 (1998) 1505–1509.
[19] Y. Chen, H. Qin, Y. Cao, H. Zhang, J. Hu, Acetone sensing properties and mechanism of SnO2 thick-films, Sensors 18 (2018) 3425.
[20] J. Li, C. Chen, J. Li, S. Li, C. Dong, Synthesis of tin-glycerate and it conversion into SnO2 spheres for highly sensitive low-ppm-level acetone detection, J. Mater. Sci. Mater. Electron. 31 (2020) 16539–16547.
[21] K.G. Dhinakar, T. Selvalakshmi, S.M. Sundar, A.C. Bose, Structural, optical and impedance properties of SnO2 nanoparticles, J. Mater. Sci. Mater. Electron. 27 (2016) 5818–5824.
[22] Y. Ding, Y. Chen, K.C. Pradel, M. Liu, Z.L. Wang, In-situ transmission electron microscopy study of oxygen vacancy ordering and dislocation annihilation in undoped and Sm-doped CeO_2 ceramics during redox processes, J. Appl. Phys. 120 (2016) 214302.
[23] G.R. Lee, M. Seong, S. Kim, K. Pyeon, R.B. Chung, Conductive SnO_{2-x} thin films deposited by thermal ALD with H_2O reactant, Vacuum 200 (2022) 111018.
[24] M. Shaban, A. Almohammedi, R. Saad, A.M. El Sayed, Design of SnO_2:Ni,Ir nanoparticulate photoelectrodes for efficient photoelectrochemical water splitting, Nanomaterials 12 (2022) 453.
[25] Y. Wang, Z. Liu, R. Wang, $NaBH_4$ surface modification on CeO_2 nanorods supported transition-metal catalysts for low temperature CO oxidation, ChemCatChem 12 (2020) 4304–4316.
[26] H.K. Mallick, Y. Zhang, J. Pradhan, M.P.K. Sahoo, A.K. Pattanaik, Influence of particle size and defects on the optical, magnetic and electronic properties of Al doped SnO_2 nanoparticles, J. Alloys Compd. 854 (2021) 156067.
[27] Y. Wang, C. Qiu, Y. Xie, L. Wang, J. Ding, J. Zhang, H. Wan, G. Guan, Intentionally introducing oxygen vacancies and Ti^{3+} defects on the surface of $Bi_4Ti_3O_{12}$ nanosheets for promoting the photoreduction of CO_2 to CH_3OH, ACS Appl. Nano Mater. 7 (2024) 3012–3023.
[28] J. Zheng, T. Xu, G. Xia, X. Yu, Oxygen defect engineering endows Co3O4 nanosheets with advanced aluminum ion storage, J Mater Chem A 10 (2022) 18322–18332.
[29] D. Manikandan, A.K. Yadav, S.N. Jha, D. Bhattacharyya, D.W. Boukhvalov, R. Murugan, XANES, EXAFS, EPR, and first-principles modeling on electronic structure and ferromagnetism in Mn doped SnO_2 quantum dots, J. Phys. Chem. C 123 (2019) 3067–3075.
[30] D.T. Hoang, V.N. Pham, H. Yu, I.H. Kwak, J. Baik, H.S. Kim, H. Lee, Enhanced photocatalytic activities of sodium borohydride-calcined magnetic manganese ferrite nanoparticles, Inorg. Chem. 63 (2024) 12054–12062.
[31] J. Wang, C. Lu, X. Liu, Y. Wang, Z. Zhu, D. Meng, Synthesis of tin oxide (SnO & SnO_2) micro/nanostructures with novel distribution characteristic and superior photocatalytic performance, Mater. Des. 115 (2017) 103–111.
[32] W.B.H. Othmen, A. Hamdi, A. Addad, B. Sieber, H. Elhouichet, S. Szunerits, R. Boukherroub, Fe-doped SnO_2 decorated reduced graphene oxide nanocomposite with enhanced visible light photocatalytic activity, J. Photochem. Photobiol. A. 367 (2018) 145–155.
[33] J. Wu, R. Zhao, H. Xiang, C. Yang, W. Zhong, C. Zhang, Q. Zhang, X. Li, N. Yang, Exposing highly active (100) facet on a SnS_2/SnO_2 electrocatalyst to boost efficient hydrogen evolution, Appl. Catal. Environ. 292 (2021) 120200.
[34] C. Zong, C. Wang, L. Hu, R. Zhang, P. Jiang, J. Chen, L. Wei, Q. Chen, The enhancement of the catalytic oxidation of CO on Ir/CeO_2 nanojunctions, Inorg. Chem. 58 (2019) 14238–14243.
[35] X. Wang, L. Zhang, L. Miao, M. Kan, L. Kong, H. Zhang, Oxidation and detection of L-cysteine using a modified Au/Nafion/glass carbon electrode, Sci. China Chem. 54 (2011) 521–525.
[36] O.A. Chuvenkova, E.P. Domashevskaya, S.V. Ryabtsev, D.V. Vysotskii, A.E. Popov, Y.A. Yurakov, O.Y. Vilkov, R. Ovsyannikov, N. Appathurai, S.Y. Turishchev, Synchrotron studies of SnO2 wire-like crystals, J. Surf. Investig. X-ray, Synchrotron Neutron Tech. 8 (2014) 111–116.
[37] N. Shanmugam, T. Sathya, G. Viruthagiri, C. Kalyanasundaram, R. Gobi, S. Ragupathy, Photocatalytic degradation of brilliant green using undoped and Zn doped SnO_2 nanoparticles under sunlight irradiation, Appl. Surf. Sci. 360 (2016) 283–290.
[38] X. Ding, F. Fang, J. Jiang, Electrical and optical properties of N-doped SnO_2 thin films prepared by magnetron sputtering, Surf. Coat. Technol. 231 (2013) 67–70.
[39] S.N. Sarangi, G.K. Pradhan, D. Samal, Band gap engineering in SnO_2 by Pb doping, J. Alloys Compd. 762 (2018) 16–20.
[40] V.N. Pham, H. Jeon, S. Hong, H. Lee, Selective oxidation of biomass molecules via ZnO nanoparticles modified using charge mismatch of the doped co ions, Inorg. Chem. 61 (2022) 16887–16894.
[41] J. Cui, K.K. Niu, R.Z. Zhang, H. Liu, S. Yu, L.B. Xing, Photocatalytic selective oxidation of toluene under encapsulated air conditions, Chem. Com. 60 (2024) 4310–4313.
[42] W. Zhang, H. Qian, Q. Hou, M. Ju, The functional and synergetic optimization of the thermal-catalytic system for the selective oxidation of 5-hydroxymethylfurfural to 2,5-diformylfuran: a review, Green Chem. 25 (2023) 893–914.
[43] W. Wu, G. Zhang, J. Zhang, G. Wang, C.H. Tung, Y. Wang, Aerobic oxidation of toluene and benzyl alcohol to benzaldehyde using a visible light-responsive titanium-oxide cluster, Chem. Eng. J. 404 (2021) 126433.
[44] H. Jeon, D.T. Hoang, J. Baik, S. Hong, H. Lee, Reduction-assisted calcination enhances the photocatalytic activity of ZnO and WO_3 nanoparticles in biomass conversion, Inorg. Chem. 63 (2024) 12370–12376.
[45] B. Baldassarri, J. He, X. Qian, E. Mastronardo, S.S. Griesemer, S. Haile, C. Wolverton, Oxygen vacancy formation energy in metal oxides: high throughput computational studies and machine learning predictions, Phys. Rev. Mater. 7 (2023) 065403.
[46] A.E. Mesoudy, D. Machon, A. Ruediger, A. Jaouad, F. Alibart, S. Ecoffey, D. Drouin, Band gap narrowing induced by oxygen vacancies in reactively sputtered TiO2 thin films, Thin Solid Films 769 (2023) 139737.

ACS Sustainable Chemistry & Engineering

pubs.acs.org/journal/ascecg

Research Article

Cr-Promoted Pt@$Fe_{2-x}Cr_xO_3$ Hematite-Type Catalysts for Efficient Base-Free HMF Oxidation under Mild Conditions

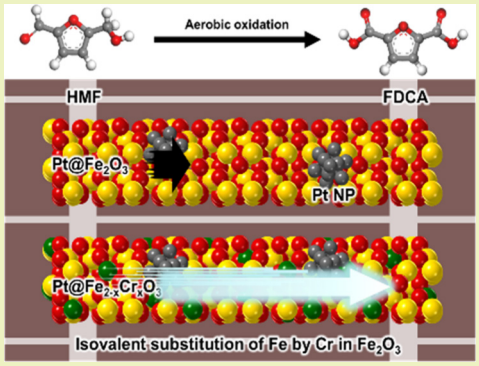
Santhana Krishnan Perumal,[‡] Hyejin Yu,[‡] Hyeonsu Je, Hangil Lee,* Yeonjoon Kim,* and Hyun Sung Kim*

Cite This: *ACS Sustainable Chem. Eng.* 2025, 13, 17470−17482

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The catalytic oxidation of 5-hydroxymethylfurfural (HMF) to 2,5-furandicarboxylic acid (FDCA) offers a sustainable pathway for biomass valorization; however, achieving base-free, efficient, and recyclable catalysis remains challenging. Herein, we report the rational design of chromium-modified Pt@$Fe_{2-x}Cr_xO_3$ nanocatalysts, synthesized via ethanolamine-assisted coprecipitation followed by platinum nanoparticle (≈2.6 nm) deposition, for the aerobic oxidation of HMF under mild conditions. The isovalent substitution of Fe^{3+} with Cr^{3+} enhanced the oxygen vacancy concentration and surface acidity while strengthening the Pt−O interactions. The optimized Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst achieved nearly 100% HMF conversion with a 78.7% FDCA yield at 120 °C and 1 bar O_2 within 12 h, compared to only 41% yield over Pt@Fe_2O_3. Kinetic studies revealed first-order behavior with activation energies of 67.9 kJ mol^{-1} for HMF conversion and 36.3 kJ mol^{-1} for FDCA formation. The catalyst retained >90% of its activity after five consecutive cycles, demonstrating its excellent recyclability. Density functional theory calculations further confirmed that Cr doping increased the HMF adsorption energy (−67.0 vs −56.6 kcal mol^{-1}) and raised the Pt d-band center, thereby promoting the generation of reactive oxygen species. This study establishes a quantitative and generalizable strategy for designing robust and recyclable catalysts for sustainable biomass conversion.



KEYWORDS: *HMF oxidation, heterogeneous catalyst, isovalent substitution, hematite, nanocatalysts*

1. INTRODUCTION

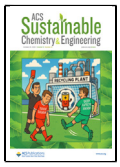
The catalytic oxidation of 5-hydroxymethylfurfural (HMF) to 2,5-furandicarboxylic acid (FDCA) represents a crucial advancement in the valorization of biomass-derived platform molecules due to its potential applications in the production of sustainable polyesters and bioplastics.[1] Among the diverse catalytic strategies explored, catalysts based on noble metals have exhibited exceptional efficacy in facilitating the selective oxidation of HMF under mild conditions.[2] However, key challenges remain, including the necessity for base-free reaction systems, stability of the catalyst, and the development of cost-effective and environmentally benign catalytic materials.[3]

Noble metals such as Au and Pt exhibit high activity and selectivity for HMF oxidation; however, their scarcity, high cost, and frequent dependence on alkaline media hinder their scalability.[4,5] In contrast, transition metal oxides (e.g., MnO_2, Co_3O_4, and Fe_2O_3) are earth-abundant, inexpensive, and environmentally benign.[6,7] However, their relatively low efficiency, selectivity, and poor stability in base-free aqueous systems limit their practical applications. To address these issues, strategies that integrate noble metals with transition metal oxides have gained increasing attention. Such hybrid systems can couple the superior catalytic performance of noble metals with the redox flexibility and durability of oxides, creating sustainable and recyclable catalytic platforms.[8−22]

Metal oxides not only serve as active catalysts but also provide exceptional support for noble metal nanoparticles owing to their tunable redox properties, high surface area, and chemical stability.[8−22] Their ability to generate oxygen vacancies and modulate their acid−base characteristics makes them ideal platforms for enhancing metal−support interactions, stabilizing nanoparticles, and improving recyclability. These features render metal oxides highly attractive as multifunctional supports in biomass oxidation reactions, particularly under base-free aqueous conditions, where catalyst durability and reusability are critical.

Among these oxides, Fe-based oxides, particularly hematite (α-Fe_2O_3), have been widely investigated as catalyst supports owing to their abundance, intrinsic magnetism, and redox versatility.[23] However, pristine Fe_2O_3 often suffers from low

Received: July 17, 2025
Revised: September 30, 2025
Accepted: October 2, 2025
Published: October 8, 2025



Downloaded via SOOKMYUNG WOMENS UNIV on October 20, 2025 at 21:41:28 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications

© 2025 American Chemical Society

conductivity, insufficient oxygen vacancy concentration, and rapid deactivation under liquid-phase oxidation conditions, which limit its catalytic efficiency.[24] To address these shortcomings, various doped Fe_2O_3 systems have been explored; however, challenges in maintaining stability and recyclability persist. In parallel, other transition metal oxides, such as MnO_2 and Co_3O_4, have also been explored as supports owing to their strong oxygen activation ability and surface acidity.[7] Nevertheless, these systems generally exhibit lower selectivity and poor long-term durability in base-free aqueous environments. Collectively, these studies highlight that although Fe- and other metal oxide-based supports are promising, their intrinsic limitations necessitate rational modification strategies, such as isovalent substitution and noble metal integration, to achieve practical applicability in biomass valorization.

Recent efforts have focused on integrating noble metal nanoparticles with metal oxides to harness synergistic interactions that improve redox properties and catalytic stability. Despite these advancements, the development of base-free, cost-effective, and magnetically recoverable catalysts with high activity and selectivity remains an ongoing challenge.[25−27]

In this study, we report the synthesis and characterization of chromium-modified Fe_2O_3 ($Fe_{2-x}Cr_xO_3$) as a support for PtNPs, aiming to develop a highly efficient and recyclable catalyst for the base-free aerobic oxidation of HMF to FDCA. Unlike previously reported Pt-based catalysts modified with transition metals, our approach relies on the isovalent substitution of Fe^{3+} with Cr^{3+} within the hematite lattice, which induces controlled lattice distortion, stabilizes oxygen vacancies, and enhances Pt−support electronic interactions. This rational design enables efficient base-free HMF oxidation under remarkably mild conditions (1 bar O_2, 120 °C), thereby providing a novel and generalizable strategy distinct from conventional surface doping approaches. Compared to previous catalytic systems, the proposed Pt@$Fe_{2-x}Cr_xO_3$ catalysts offer a combination of high surface area, enhanced oxygen vacancies, and optimized acid site distribution, addressing both the catalytic efficiency and stability issues commonly associated with noble metal catalysts. The catalytic performance was evaluated under mild reaction conditions, with a particular focus on understanding the synergistic effects of Pt NPs and Cr-modified Fe_2O_3 on the oxidation efficiency and selectivity. The findings of this study provide new insights into the rational design of multifunctional metal oxide-based catalysts and their potential application in sustainable biomass valorization. The Pt@$Fe_{2-x}Cr_xO_3$ catalysts demonstrated superior activity with enhanced stability, underscoring their promise as efficient catalysts for industrial applications in green chemistry.

2. EXPERIMENTAL SECTION

2.1. Chemicals Used. All chemicals used in this study were commercially obtained and utilized without further purification. Polyvinylpyrrolidone (PVP, >98%, M.W. = 10,000), ethanolamine (99%), ethanol (95%), sodium borohydride, iron(II) sulfate heptahydrate, and chromium chloride were purchased from Alfa Aesar for the purpose of catalyst preparation. 5-Hydroxymethylfurfural (HMF, 99%), 2,5-diformylfuran (DFF, 98%), 5-formyl-2-furancarboxylic acid (FFCA, 98%), and 2,5-furandicarboxylic acid (FDCA, 98%) were also purchased from Sigma-Aldrich for catalytic reactions.

2.2. Synthesis of Fe_2O_3 and $Fe_{2-x}Cr_xO_3$ (x = 0.05−0.25, y = 1−0.75). Iron oxide (Fe_2O_3) was synthesized using an ethanolamine-based complex precipitation method. Initially, 0.005 mol of iron(II) sulfate heptahydrate was dissolved in 10 mL of deionized distilled water (DDW) and subsequently combined with 50 mL of ethanol while stirring in a 250 mL beaker. A 1 M ethanolamine solution was then added dropwise until the pH reached a range of 9.9−10. The appearance of a black-green solution signified the formation of the Fe−OH−N−O−Et complex. The solution was stirred for a duration of 4 h, during which a color transition from black-green to brick-red occurred, indicating the oxidation of the iron complex in the presence of air. The resultant material was subjected to multiple washings to eliminate excess ethanolamine, followed by drying and calcination at 500 °C for 5 h. For the synthesis of chromium-substituted iron oxide complexes, the same methodology was employed, with varying molar ratios of chromium and iron salts. The resulting materials were designated as $Fe_{2-x}Cr_xO_3$, where x represents the values of 0.1, 0.2, 0.3, 0.4, or 0.5. The iron oxide support synthesized without chromium was designated as Fe_2O_3.

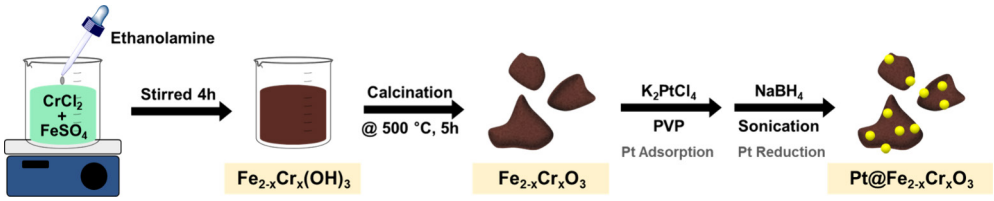
2.3. Synthesis of Pt@Fe_2O_3 and $Fe_{2-x}Cr_xO_3$ (x = 0.1−0.5) Catalysts. Catalysts supported by platinum (Pt) with desired weight percentages were synthesized utilizing a polyvinylpyrrolidone (PVP)-assisted adsorption technique. Initially, a solution of 0.05 M potassium tetrachloroplatinate and 25 mg of PVP was prepared in 300 mL of deionized distilled water (DDW) and subjected to sonication for 5 min at ambient temperature to enhance the interactions between Pt and PVP. Following this, 0.425 g of the support material was added to the solution, and the mixture underwent an additional hour of sonication. Subsequently, a freshly prepared 0.1 M sodium borohydride ($NaBH_4$) solution (50 mL) was introduced dropwise under ice-cold conditions, followed by another hour of sonication. The resultant solid material was isolated through filtration and extensively washed with hot water and ethanol to eliminate residual PVP, sodium ions (Na^+), and chloride ions (Cl^-) until a neutral pH was attained. Ultimately, the catalysts were dried under vacuum at 80 °C overnight, resulting in the formation of Pt@Fe_2O_3 and Pt@$Fe_{2-x}Cr_xO_3$.

2.4. Catalyst Characterization. X-ray diffraction (XRD) was conducted using a PANalytical instrument equipped with Cu Kα radiation (λ = 1.5418 Å) at a beam voltage of 40 kV. Patterns were recorded in the 2θ range of 20−80° with a step size of 0.02° per 0.2 s. Nitrogen adsorption isotherms and pore characteristics were evaluated using a high-precision gas−vapor adsorption analyzer (BELSORP MAX X, Microtrac MRB). Samples were degassed under nitrogen at 150 °C for 15 h before measurement. Pore size distributions were calculated using the Barrett−Joyner−Halenda method. Solid-state FT-IR spectroscopy was conducted on a Frontier FT-IR spectrometer (PerkinElmer). X-ray photoelectron spectroscopy (XPS) was carried out under ultrahigh vacuum on a Thermo Scientific K-Alpha+ apparatus equipped with Al Kα radiation. The binding energy (BE) shift due to the surface charging was adjusted based on a reference to the C 1s line at 284.8 eV. Transmission electron microscopy (TEM) was performed on a Philips-FEI TECNAI F30 field-emission electron microscope operated at an acceleration voltage of 300 kV. The samples were previously suspended in ethanol and dispersed ultrasonically. Drops of the suspension were applied on a lacey support film. Energy-dispersive spectroscopy (EDS) was conducted on an X-MaxN 80 T IE250 (Oxford).

Temperature-programmed desorption of ammonia (NH_3-TPD) measurements were performed on an AutoChem TP5080 chem-isorption analyzer (Xianquan Tianjin, China) equipped with a thermal conductivity detector (TCD). 50 mg of each catalyst was loaded in a quartz reactor and preheated at 200 °C by highly purified He (30 mL min^{-1}) for 1 h and subsequently cooled down to room temperature. After that, the catalyst was saturated with 25 mL min^{-1} anhydrous NH_3 for 1 h at the same temperature and then flushed with He for 1 h to remove the physically absorbed NH_3. Finally, the TPD curve was recorded from room temperature to 800 °C at a rate of 10 °C min^{-1} in He (50 mL min^{-1}). O_2-TPD and temperature-programmed reduction with hydrogen (H_2-TPR) experiments were also conducted

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

Scheme 1. Schematic Illustration of Preparation of the Pt@$Fe_{2-x}Cr_xO_3$ Nanocomposites



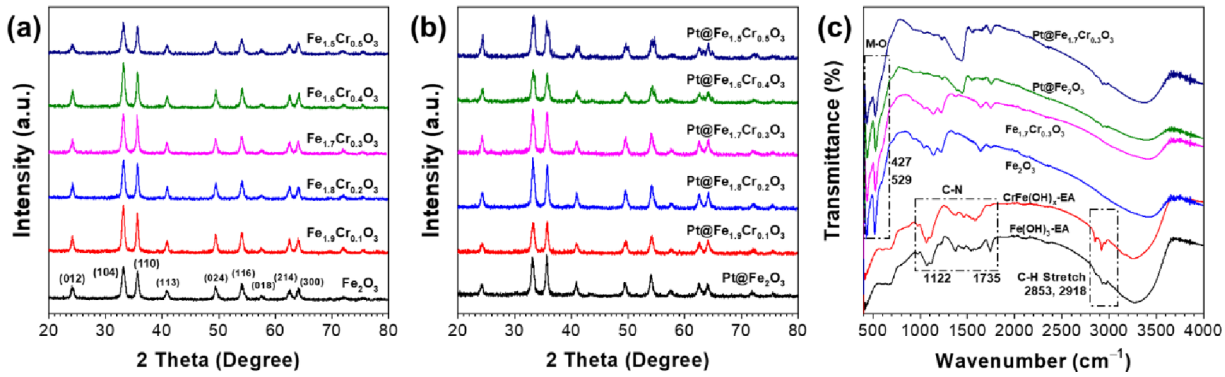


Figure 1. (a, b) XRD patterns and (c) FT-IR spectra of the Pt@$Fe_{2-x}Cr_xO_3$ nanocomposites.

on an AutoChem TP5080 chemisorption analyzer (Xianquan Tianjin, China) with a TCD. Prior to reduction, 50 mg of the catalyst was loaded in the quartz reactor and preheated at 300 °C under a pure O_2 stream (25 mL min^{-1}) for 1 h with a ramp rate of 10 °C min^{-1}. After the sample was cooled down to ambient temperature in the same stream, the catalyst was then exposed to 5% H_2/N_2 (30 mL min^{-1}), and the temperature was subsequently raised from room temperature to 800 °C with a ramp rate of 10 °C min^{-1}.

2.5. Catalyst-Driven Aerobic Oxidation of HMF. A Teflon-lined stainless-steel high-pressure reactor (Berghof, BR-25) with a PTFE liner capacity of 25 mL was used for the liquid-phase, base-free aerobic oxidation of HMF. In each reaction, 0.5 mmol of HMF was dissolved in 5 mL of deionized distilled water (DDW), and the required amount of the catalyst was added to the mixture. A photograph of the reactor is provided in Figure S1. The system was purged with molecular oxygen for 3 min to create an oxygen-rich atmosphere. After achieving the necessary oxygen pressure, the reactor was positioned in a preheated sand bath maintained at a temperature range of 100−130 °C and stirred magnetically at 600 rpm for a specified duration. Upon completion of the reaction, the reaction mixture containing the catalyst and products was transferred to a 15 mL conical tube. As FDCA exhibits low solubility in DDW, the solvent was completely removed using a sophisticated evaporator (BioChromato) prior to product analysis. To ensure complete recovery of the residual catalyst in the reactor, the vessel was rinsed with methanol, and the washings were combined in a conical tube. Subsequently, 10 mL of methanol was added to the dried mixture to facilitate the separation of the catalyst and products, and high-speed centrifugation at 9500 rpm for 6 min was performed. The heterogeneous catalyst settled completely, and the supernatant was used to analyze the HMF conversion and product yields.

For catalyst recovery, the spent catalyst was washed three times with DDW and ethanol, separated by high-speed centrifugation (9500 rpm, 6 min), and then vacuum-dried at 80 °C overnight for reuse. To compensate for the inevitable catalyst mass loss during recycling and maintain the initial catalyst loading, a small quantity of the fresh catalyst was added.

For high-performance liquid chromatography (HPLC) analysis, 0.1 mL of the product mixture was diluted with 0.9 mL of acetonitrile. The reaction products were analyzed using gas chromatography (Agilent 8860 GC) and HPLC (Atlantis dC18 column, Waters). The conversion of HMF and the yield percentages of 2,5-diformylfuran (DFF) were quantified using a gas chromatograph equipped with a flame ionization detector, while the concentrations of FFCA and FDCA were determined using HPLC coupled with a dual-absorbance UV−vis detector (Waters 2487). A 5 mM ammonium formate solution was employed as the mobile phase for the HPLC analysis.

Quantitative analysis was conducted using an external standard method, with calibration curves established from standard solutions at varying concentrations for HMF, DFF, FFCA, and FDCA, achieving R^2 values greater than 0.999. The accuracy of the data was maintained within a ±5% error margin. The carbon balance of the catalyst was calculated, with the most effective catalyst demonstrating a carbon balance of 98−100%, thereby affirming the reliability of the experimental findings. The conversion of HMF, product yield, initial reaction rate, and FDCA productivity were determined using standard equations, as delineated in eqs 1 and 2.

$$\text{HMF conversion}[\%] = \frac{\text{moles of HMF consumed}}{\text{moles of initial HMF}} \times 100 \quad (1)$$

$$\text{product(DFF, FFCA, or FDCA) yield}[\%] = \frac{\text{moles of products formed}}{\text{moles of initial HMF}} \quad (2)$$

3. RESULTS AND DISCUSSION

3.1. Characterization of our Catalysts. As illustrated in Scheme 1, ethanolamine served as an additive that increased the pH to ∼9.9−10, thereby inducing the precipitation of $Fe(OH)_x{\cdot}Cr(OH)_x$ intermediates, which were subsequently transformed into crystalline α-Fe_2O_3 and $Fe_{2-x}Cr_xO_3$ upon calcination. The resulting $Fe_{2-x}Cr_xO_3$ supports were then used to synthesize Pt@$Fe_{2-x}Cr_xO_3$ composites, where platinum species were adsorbed from K_2PtCl_4 in the presence of polyvinylpyrrolidone (PVP) and reduced with $NaBH_4$ under sonication, yielding uniformly dispersed Pt nanoparticles on the oxide surface (Scheme 1).

X-ray diffraction (XRD) analysis was used to examine the phase compositions of the synthesized Fe_2O_3 and series of $Fe_{2-x}Cr_xO_3$ supports, as well as the Pt@$Fe_{2-x}Cr_xO_3$ catalysts. The XRD patterns (Figure 1a) reveal that pristine Fe_2O_3 displays characteristic peaks at 24.2, 33.1, 35.6, 40.4, 54.1, 57.5,

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

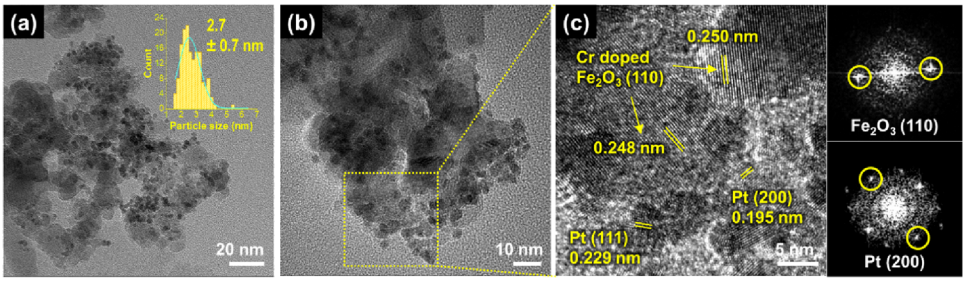


Figure 2. (a, b) TEM images with a size distribution graph and (c) HR-TEM image with lattice parameters and SAED patterns of Pt@$Fe_{1.7}Cr_{0.3}O_3$ nanocomposites as indicated.

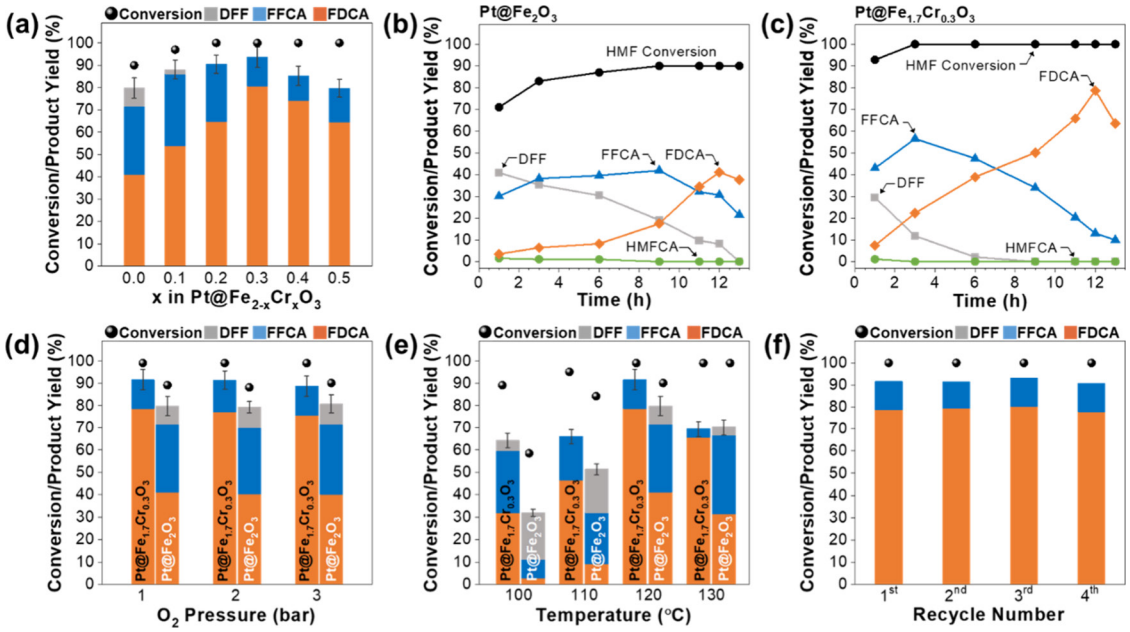


Figure 3. (a) Catalytic performance of a series of Pt@$Fe_{2-x}Cr_xO_3$ catalysts; (b, c) influence of reaction time; (d) impact of pressure; (e) effect of temperature; (f) influence of catalyst dosage, with comparisons made to the Pt@Fe_2O_3 and Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalysts.

62.4, 64, 71.8, and 75.4°, which correspond to the pure α-Fe_2O_3 phase (hematite, JCPDS 33-0664).[28] The $Fe_{2-x}Cr_xO_3$ supports predominantly exhibited the α-Fe_2O_3 phase, indicating the preservation of a rhombohedral crystal structure in samples with varying chromium molar percentages. In particular, owing to the smaller ionic radius of Cr^{3+} compared to that of Fe^{3+}, the incorporation of Cr into the $Fe_{2-x}Cr_xO_3$ lattice resulted in a gradual contraction of the crystal structure. This effect was evidenced by a slight positive shift in the XRD peak positions, which became more pronounced with increasing Cr loading, indicating successful Cr substitution and structural modification. The retention of α-Fe_2O_3 peaks following PtNP deposition suggests a stable crystal structure, potentially contributing to the observed catalytic activity discussed in the subsequent section on catalytic performance (Figure 1b).

Fourier-transform infrared (FT-IR) spectroscopy results are shown in Figure 1c. The Fe and Fe−Cr hydroxyl ethanolamine complexes exhibited characteristic peaks for C−N and C−H stretching at 1122 and 1735 cm^{-1}, as well as at 2853 and 2918 cm^{-1}, respectively. Following calcination, the formation of metal oxides was confirmed by the appearance of two distinct vibrational bands at approximately 427 and 529 cm^{-1}, which correspond to the M−O−M (Fe−O−Fe or Fe−O−Cr) stretching modes.

Field-emission scanning electron microscopy (FE-SEM) images of the Fe_2O_3 support (Figure S2), acquired at a scale of 1 μm under various magnifications, revealed irregularly shaped agglomerates. The irregular morphology can be attributed to the ethanolamine-assisted coprecipitation and subsequent calcination process, where rapid nucleation and particle growth often lead to partial aggregation rather than well-defined structures. Such agglomeration is typical for metal oxide nanoparticles prepared via precipitation, particularly after drying and high-temperature treatment. Upon chromium incorporation, the average aggregate size decreased from 14.7 to 11.9 μm, suggesting improved lattice stabilization and a more homogeneous metal ion distribution. Moreover, the EDX elemental mapping images demonstrate the representative testing of Fe_2O_3 and $Fe_{1.7}Cr_{0.3}O_3$ supports, confirming the homogeneous distribution of Fe and Cr in $Fe_{1.7}Cr_{0.3}O_3$ and the uniform dispersion of Pt on both supports in Pt@Fe_2O_3 and Pt@$Fe_{1.7}Cr_{0.3}O$ (Figure S3). Quantitative elemental analysis of the Pt-decorated Fe_2O_3-based supports was further carried out using X-ray fluorescence (XRF), and the results are summarized in Table S1. The measured Fe, Cr, and Pt loadings were in close agreement with the theoretical values, confirming that the catalysts were synthesized with the intended stoichiometry.

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

Field-emission transmission electron microscopy (FE-TEM) analysis provided detailed insights into the structural characteristics of Pt@$Fe_{1.7}Cr_{0.3}O_3$. Bright PtNPs were observed as distinct dots on the support material (Figure 2a,b). A histogram derived from the measurements of 150 nanoparticles indicates an average platinum particle size of 2.6 nm (inset of Figure 2a). Lattice spacings of 0.25, 0.19, and 0.23 nm, corresponding to the (110) plane of α-Fe_2O_3 and the (220) and (111) planes of platinum, respectively, were observed in the selected-area electron diffraction (SAED) pattern and were consistent with the XRD results, as shown in Figure 2c.[29,30] These observations demonstrate the fine dispersion of platinum on the support material.

A comprehensive array of sorption characterization techniques, including nitrogen sorption (BET), was employed to evaluate the surface area of both the support and catalysts, which is a crucial factor affecting catalytic performance. A summary of the specific surface areas of the samples derived from nitrogen sorption isotherms is presented in Table S2. The BET surface areas of the synthesized Fe_2O_3 and $Fe_{2-x}Cr_xO_3$ supports were approximately 35 m^2/g, indicating comparable values. Upon the incorporation of PtNPs into the synthesized Pt@Fe_2O_3 and Pt@$Fe_{2-x}Cr_xO_3$ compositions, the supports exhibited a modest increase in surface area to approximately 42 m^2/g, which can be attributed to the presence of PtNPs. The retention of α-Fe_2O_3 peaks following PtNP deposition suggests a stable crystal structure, potentially contributing to the observed catalytic activity discussed in the subsequent section on catalytic performance.

3.2. Investigation of Catalytic Performance. The synthesized Pt@$Fe_{2-x}Cr_xO_3$ catalyst was evaluated for the oxidation of HMF to formic dicarboxylic acid (FDCA) under the optimized reaction conditions. A systematic analysis of the fundamental parameters, including the reaction temperature, duration, oxygen pressure, catalyst dosage, and varying concentrations of HMF, was conducted to maximize the HMF conversion and FDCA yield while minimizing the formation of reaction intermediates such as diformylfuran (DFF) and 5-formyl-2-furancarboxylic acid (FFCA).

To elucidate the contributions of the support material ($Fe_{2-x}Cr_xO_3$) and the catalyst (Pt@$Fe_{2-x}Cr_xO_3$), a series of HMF oxidation tests were performed on all the supported catalysts to identify the most effective catalyst. Under identical conditions, no detectable conversion of HMF or yield of FDCA was observed in the absence of a catalyst (Figure 3a). The introduction of Cr into Fe_2O_3 significantly enhanced its catalytic activity compared to that of unmodified Fe_2O_3. Specifically, the use of PtNPs supported on unmodified Fe_2O_3 yielded a maximum FDCA yield of 41% at complete HMF conversion, underscoring the pivotal role of platinum sites in facilitating the HMF oxidation reaction. Further investigations (Figure 3a) into the varying chromium loadings (x = 0.1, 0.2, 0.3, 0.4, and 0.5) within the $Fe_{2-x}Cr_xO_3$ supports resulted in complete HMF conversion, with the optimal FDCA yield of 78.7% occurring at a chromium loading of 0.3, beyond which the yield diminished. This indicates that $Fe_{1.7}Cr_{0.3}O_3$ serves as the most effective support for the Pt@$Fe_{2-x}Cr_xO_3$ catalyst system in the oxidation of HMF. The performance of Pt@$Fe_{2-x}Cr_xO_3$ was benchmarked against that of Pt@Fe_2O_3 as the reference catalyst.

Reaction time is a critical factor in optimizing FDCA production (Figure 3b,c). As the reaction progressed, the FDCA yield increased proportionally, whereas the concentrations of the intermediate products, such as DFF and FFCA, decreased. The Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst achieved 100% conversion of HMF after 3 h under the specified conditions, resulting in yields of 11.8% DFF, 56.5% FFCA, and 22.3% FDCA. In comparison to Pt@Fe_2O_3, the Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst consistently demonstrated superior FDCA yields at all reaction times. However, extending the reaction time to 12 h resulted in increased yields of FDCA and FFCA, which reached 78.7 and 13%, respectively. Prolonging the reaction duration beyond 12 h did not yield further improvements in FDCA yield, establishing 12 h as the optimal reaction time.

The influence of the oxygen pressure was assessed by varying the O_2 pressure between 1 and 3 bar at a constant temperature of 120 °C (Figure 3d). The Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst exhibited yields of 78.7% FDCA and 13% FFCA at 1 bar O_2, resulting in a total yield of 91.7%. An increase in O_2 pressure from 1 to 3 bar led to a decrease in the FDCA yield from 78.7 to 78.1%, with similar yields. Further increasing the pressure from 3 to 5 bar resulted in a significant decline in the FDCA yield to 51.3%. The Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst consistently outperformed Pt@Fe_2O_3 in terms of FDCA and FFCA yields across various O_2pressures. These results confirmed that an O_2 pressure of 1 bar was optimal for maximizing the FDCA yield. Although higher pressures typically enhance the oxidation products of HMF (FDCA and FFCA), excessive pressure results in overoxidation of reactants and the generation of overoxidized byproducts. Additional studies conducted in both open and closed systems under different atmospheres (O_2, air, and argon) demonstrated that the autogenous pressure developed within a closed reactor, along with the reactive oxygen species (ROS), was sufficient to achieve high yields of FDCA and FFCA (Table S3). These findings underscore the necessity of optimizing oxygen pressure for efficient FDCA production, with subsequent studies conducted under 1 bar O_2 pressure.

The effect of temperature on HMF oxidation was investigated by varying the reaction temperature (100−130 °C) under 1 bar O_2 pressure (Figure 3e). At 100 °C, limited HMF conversion and low FDCA yields were observed for Pt@Fe_2O_3, whereas Pt@$Fe_{1.7}Cr_{0.3}O_3$ achieved excellent conversion and product yields at all temperatures. Increasing the temperature to 120 °C resulted in complete HMF conversion and an exceptional FDCA yield. However, at 130 °C, while HMF conversion remained complete, the FDCA yield decreased to 20.6%, likely due to the formation of overoxidized byproducts. These results highlight the importance of maintaining moderate temperatures to prevent undesirable side reactions while maximizing the FDCA yield. Consequently, subsequent oxidation reactions were conducted at 120 °C under 1 bar O_2 pressure.

To evaluate the recoverability and potential for industrial application, the most effective Pt@ Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst was subjected to multiple catalytic cycles under the optimized reaction conditions. The catalyst was tested at 120 °C for a reduced reaction time of 12 h under 1 bar O_2 pressure. Even after four consecutive cycles, HMF conversion remained at 100%, with consistent yields of DFF, FFCA, and FDCA (Figure 3f). This indicated that the catalyst exhibited no significant deactivation during the HMF oxidation reaction, demonstrating its excellent recyclability and stability. XPS analysis of the spent Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst after repeated oxidation cycles revealed that both the Pt oxidation states and the O 1s vacancy-related signals remained nearly identical to

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

those of the fresh catalyst, as shown in Figure S4. This stability indicates that the electronic structure and oxygen vacancy concentration of the catalyst were preserved during the reaction, which is consistent with its excellent recyclability.

To benchmark the performance of our catalyst, we have summarized representative studies employing Au- and Pt-based catalysts for HMF oxidation in Table S4. As can be seen, most of these systems required alkaline additives, elevated oxygen pressures, or reaction temperatures above 120 °C to achieve high FDCA yields. In contrast, the $Pt@Fe_{1.7}Cr_{0.3}O_3$ catalyst reported in this work achieved nearly complete HMF conversion and high FDCA yield under neutral aqueous conditions at 1 bar O_2 and 120 °C, highlighting its superior efficiency under mild and sustainable conditions.

3.3. Implications of the Design of Efficient Biomass Vaporization Catalysts. To gain a deeper insight into the catalyst performance, surface characterizations were conducted using TPD, TPR, and HR-XPS analyses. These techniques provide valuable information on the surface properties of the catalysts, including their adsorption behavior, redox characteristics, and elemental composition.

A thorough investigation into the effects of Cr doping on the enhancement of catalytic performance was conducted using NH_3 temperature-programmed desorption (NH_3-TPD) and H_2 temperature-programmed reduction (H_2-TPR) analyses. These methodologies were used to assess the surface characteristics, including reducibility and presence of active sites on both the support and catalysts. The NH_3-TPD profiles of the supports and catalysts are shown in Figure 4a. The

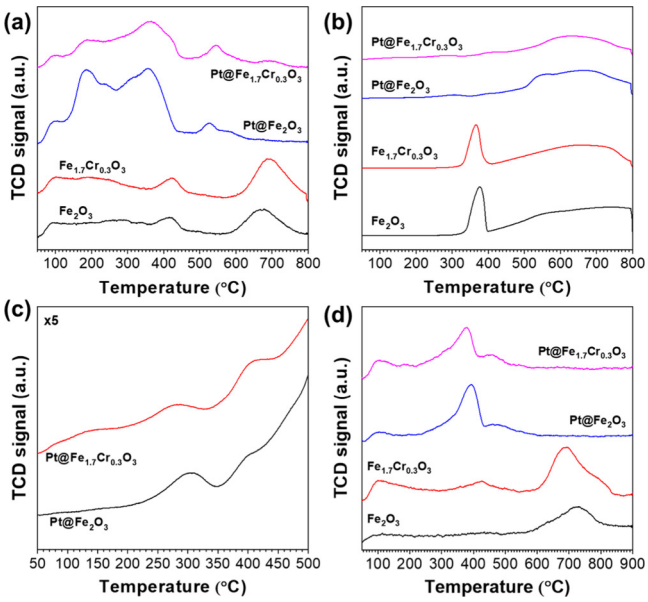


Figure 4. NH_3-TPD profiles (a), H_2-TPR profiles (b, c), and O_2-TPD profiles (d) of Fe_2O_3, $Fe_{1.7}Cr_{0.3}O_3$, $Pt@Fe_2O_3$, and $Pt@Fe_{1.7}Cr_{0.3}O_3$ catalysts.

desorption peaks were classified into three categories: weakly acidic sites (<250 °C), moderately acidic sites (250−400 °C), and strongly acidic sites (>400 °C).[31] As shown in Figure 4a and summarized in Table 1, both Fe_2O_3 and $Fe_{1.7}Cr_{0.3}O_3$ exhibited a combination of weak, moderate, and strong acidic sites. Notably, $Fe_{1.7}Cr_{0.3}O_3$ demonstrated a total acidity of 0.98 mmol NH_3/g catalyst, surpassing the 0.73 mmol NH_3/g catalyst observed for Fe_2O_3. Following the loading of platinum

Table 1. Acidity Profile of Supports and Catalysts

	acidity (mmol NH_3/g cat)[a]			
catalysts	weak acid sites	moderate acid sites	strong acid sites	total acidity (mmol NH_3/g cat)
Fe_2O_3	0.24		0.49	0.73
$Fe_{1.7}Cr_{0.3}O_3$	0.12	0.33	0.53	0.98
$Pt@Fe_2O_3$	0.74	0.81	0.20	1.75
$Pt@Fe_{1.7}Cr_{0.3}O_3$	0.47	1.06	0.44	1.97

[a]Obtained from NH_3-TPD (weak acid sites at >250 °C, moderate acid sites at 250−400 °C, and strong acid sites at >400 °C).

(Pt), the distribution of acidic sites underwent a transformation, resulting in a higher proportion of weak acid sites. Specifically, $Pt@Fe_{1.7}Cr_{0.3}O_3$ exhibited a total acidity of 1.55 mmol NH_3/g catalyst, with weak and moderate acidic sites predominating. This alteration is attributed to the synergistic interaction between the PtNPs and the support, which enhances the catalytic efficiency for the selective oxidation of HMF. Desorption peaks occurring below 370 °C were identified as Brønsted acid sites, whereas those above 370 °C were associated with Lewis acid sites.[32] $Pt@Fe_{1.7}Cr_{0.3}O_3$ exhibited both types of acid sites, including coordinatively unsaturated sites of Fe, Cr, and Pt. Brønsted acid sites facilitate the formation of carboxyl groups, whereas Lewis acid sites enhance the interactions with carbonyl groups in the FFCA intermediate, thereby improving product selectivity.[7,33] The synergistic interaction between the Brønsted and Lewis acid sites positions $Pt@Fe_{1.7}Cr_{0.3}O_3$ as a highly effective catalyst for HMF oxidation.

The H_2-TPR profiles of Fe_2O_3, $Fe_{1.7}Cr_{0.3}O_3$, and their Pt-functionalized derivatives are shown in Figure 4b. Fe_2O_3 exhibited two distinct reduction peaks: one at approximately 377 °C, corresponding to the reduction of Fe_2O_3 to Fe_3O_4, and another above 450 °C, indicative of the reduction of Fe_3O_4 to Fe.[31,32] In contrast, $Fe_{1.7}Cr_{0.3}O_3$ displayed its first reduction peak at a temperature approximately 10 °C lower than that of Fe_2O_3, reflecting the synergistic effect between Cr and Fe. Notably, the Pt-based catalysts ($Pt@Fe_2O_3$ and $Pt@Fe_{1.7}Cr_{0.3}O_3$) did not exhibit any reduction peaks below 400 °C, as the PtNPs were prereduced during catalyst preparation using sodium borohydride ($NaBH_4$), which resulted in enhanced low-temperature redox properties. Detailed analysis revealed additional reduction bands at lower temperatures, which were attributed to the simultaneous reduction of high-valence Cr, Fe, and Pt species on the catalyst surface (Figure 4c). Collectively, these findings indicate that $Pt@Fe_{1.7}Cr_{0.3}O_3$ possesses superior catalytic properties, owing to its increased surface area, optimized distribution of acidic sites, and enhanced reducibility, thereby establishing it as an ideal candidate for HMF oxidation.

In addition to NH_3-TPD and H_2-TPR, O_2 temperature-programmed desorption (O_2-TPD) was conducted to directly probe the formation of oxygen vacancies and oxygen mobility in the Cr-modified catalysts (Figure 4d). The desorption profiles exhibited two characteristic regions: a low-temperature peak (~300−400 °C) corresponding to surface-adsorbed oxygen species associated with oxygen vacancies and a high-temperature peak (700−900 °C) ascribed to lattice oxygen release. Notably, $Pt@Fe_{1.7}Cr_{0.3}O_3$ displayed its main low-temperature desorption peak at 378 °C, which is lower than that of $Pt@Fe_2O_3$ (395 °C). The downshift in desorption

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

temperature for Pt@$Fe_{1.7}Cr_{0.3}O_3$ suggests weaker M−O bond strength and higher oxygen mobility, reflecting an increased density of labile oxygen vacancies. In contrast, the higher desorption temperature of Pt@Fe_2O_3 indicates that excessive Cr incorporation leads to stronger oxygen binding and reduced vacancy mobility, which is consistent with its lower catalytic efficiency. These O_2-TPD results are in excellent agreement with the subsequent O 1s XPS analysis, further confirming that an optimal Cr content (x = 0.3) maximizes oxygen vacancy concentration and mobility, thereby facilitating oxygen activation and superior catalytic activity in HMF oxidation.

Figure 5 presents the X-ray photoelectron (XPS) spectra for the Pt 4f, Cr 2p, Fe 2p, and O 1s regions of Pt@Fe_2O_3 and

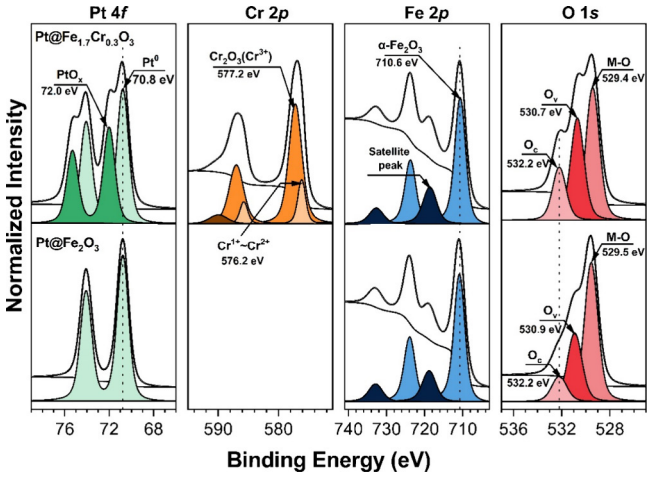


Figure 5. High-resolution XPS spectra of Pt 4f, Cr 2p, Fe 2p, and O 1s for Pt@Fe_2O_3 and Pt@$Fe_{1.7}Cr_{0.3}O_3$ (left to right) as indicated.

Pt@$Fe_{1.7}Cr_{0.3}O_3$ nanoparticles. Analysis of the binding energy values and peak deconvolutions provides significant insights into the oxidation states and electronic interactions of the constituent elements. The Pt 4f spectrum displays characteristic peaks associated with the Pt^0 and PtO_x species. In the Pt@Fe_2O_3 sample, the Pt^0 peak is predominantly observed at approximately 70.8 eV. However, for the Pt@$Fe_{1.7}Cr_{0.3}O_3$ sample, the Cr incorporation results in a shift of the binding energy values to approximately 72.0 eV, which corresponds to PtO_x, indicating a slight elevation.[12,34,35] This shift is accompanied by increased surface concentrations of Pt^0 (75.8%) and Pt^{2+} (24.2%), suggesting an electronic interaction between the PtNPs and the $Fe_{1.7}Cr_{0.3}O_3$ matrix. These XPS results indicate that Cr incorporation increases the proportion of PtO_x species; however, both Pt^0 and PtO_x contribute synergistically to the oxidation of HMF. While Pt^0 sites facilitate substrate adsorption and dehydrogenation, PtO_x species promote oxygen activation, and their coexistence at the Pt^0/PtO_x heterointerface provides the most active sites for efficient production of FDCA.

The Cr 2p spectrum reveals the presence of both Cr^{3+} and lower oxidation states (Cr^{1+}−Cr^{2+}), with the peak corresponding to Cr^{3+} identified at 577.2 eV, while the Cr^{1+}−Cr^{2+} peak is observed at 576.2 eV.[33,36,37] These findings imply that the reduction of chromium may influence the electronic structure and surface properties of the catalyst. The Fe 2p spectrum exhibits a prominent peak at 710.6 eV, characteristic of α-Fe_2O_3, along with a peak that corroborates the presence of Fe^{3+} species.[38] The spectral features remained consistent between Pt@Fe_2O_3 and Pt@$Fe_{1.7}Cr_{0.3}O_3$ samples, indicating that chromium does not significantly alter the iron oxidation state, although it modifies its electronic environment.

The O 1s spectra of Pt@Fe_2O_3 and Pt@$Fe_{1.7}Cr_{0.3}O_3$ reveal three principal peaks corresponding to metal−oxygen (M−O) bonding, surface oxygen vacancies (O_v), and chemisorbed oxygen species (O_c).[12,33,34,36,37] The M−O peak, indicative of lattice oxygen, was observed at 529.5 eV for Pt@Fe_2O_3. However, with the incorporation of Cr, this peak exhibits a slight shift to a lower binding energy of 529.4 eV. Similarly, the O_v peak, which is associated with surface oxygen vacancies, shifted from 530.9 eV for Pt@Fe_2O_3 to 530.7 eV for Pt@$Fe_{1.7}Cr_{0.3}O_3$. These observed shifts in the binding energy can be attributed to the introduction of Cr, which has a lower electronegativity than Fe. Cr within the Fe_2O_3 lattice results in diminished binding strength between oxygen and the surrounding metal cations owing to its lower electronegativity. Consequently, M−O bonding requires less energy for electron ejection, which results in a decrease in the binding energy. Furthermore, the formation of O_v is influenced by the presence of Cr because it tends to facilitate defect formation within the oxide lattice, thereby further reducing the binding energy of the O_v peak. However, the O_c peak remained relatively stable, indicating that the primary effect of Cr doping was localized to the lattice oxygen and vacancy sites rather than affecting the adsorbed oxygen species.

Compared to Pt@Fe_2O_3, the Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst exhibited a significantly higher concentration of O_v, indicating that the incorporation of chromium enhances the generation of oxygen vacancies and defects. This increase in O_v is closely associated with the enhancement of acidic sites on the catalyst surface, which play a crucial role in catalytic reactions involving acidic functional groups. This observation is consistent with the results of temperature-programmed reduction (TPR) analysis, which also indicates a greater number of reducible oxygen species in the Cr-doped system, further supporting the enhanced catalytic potential of Pt@$Fe_{1.7}Cr_{0.3}O_3$.

XPS analysis indicated that the incorporation of Cr into Fe_2O_3 influenced the electronic environment of lattice oxygen. This alteration resulted in a decrease in the binding energies of both the metal−oxygen bonds and the surface oxygen vacancies. Such shifts imply an enhancement in the oxygen mobility and defect density within the Pt@$Fe_{1.7}Cr_{0.3}O_3$ composite, potentially improving the catalytic activity by facilitating oxygen-related redox reactions. The insights from the XPS analysis, indicating modified electronic environments due to Cr incorporation, set the stage for understanding the enhanced catalytic kinetics observed, as discussed below. Consequently, chromium was selected as the dopant because its smaller ionic radius compared to that of Fe^{3+} allows isovalent substitution within the hematite lattice, inducing lattice distortion and stabilizing oxygen vacancies. Moreover, the Cr^{6+}/Cr^{3+} redox couple promotes reversible electron transfer, facilitating the formation and regeneration of vacancy sites. These unique features make Cr particularly effective in enhancing the oxygen vacancy density and catalytic performance compared to other nontoxic metal dopants. In addition, we examined other transition metals such as Co, Ni, and Mn under the same preparation conditions, but the increase in oxygen vacancy concentration and catalytic activity was not as pronounced as in the case of Cr incorporation.

Kinetic analysis of base-free oxidation of HMF utilizing Pt@Fe_2O_3 and Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalysts was performed to

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

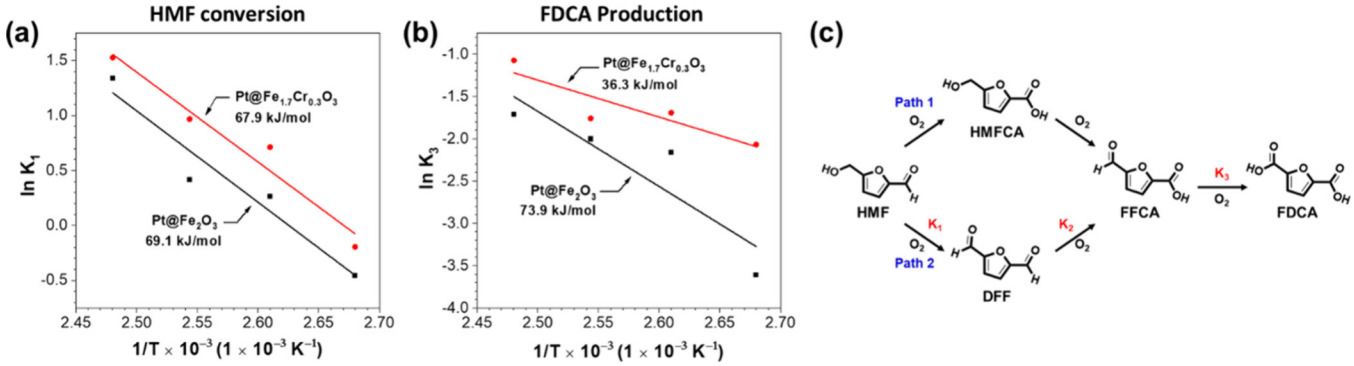


Figure 6. Kinetic analysis of HMF oxidation: (a) conversion and (b) FDCA production over Pt@$Fe_{1.7}Cr_{0.3}O_3$; (c) reaction and kinetic pathway of HMF oxidation to FDCA.

elucidate the various kinetic parameters influencing the conversion of HMF to FDCA. The experiments were conducted in an aqueous medium at different temperatures for short reaction times (1 h). The kinetic parameters were derived from the formation of the oxidation intermediates (Tables S5−S7 and Figure 6). These findings indicate that the conversion of HMF follows first-order kinetics, which is consistent with our experimental observations and the existing literature.[12,39,40] As expected, an increase in the reactant concentration and reaction temperature resulted in an enhanced conversion of HMF and improved yields of 5-formylfuran-2-carboxylic acid (FFCA) and FDCA. The oxidation pathway was elucidated by product analysis, which revealed the formation of DFF and FFCA as key intermediates, whereas HMFCA was not detected under the reaction conditions used in this study (Figure 6c). This observation suggests that path 1 is unlikely to be dominant and that the reaction predominantly proceeds via path 2.

The temperature dependence of the rate constants (K_1, K_2, and K_3) indicates that elevated reaction temperatures facilitate both the conversion of HMF and the formation of FDCA. Importantly, the introduction of chromium atoms did not modify the reaction pathway but significantly enhanced the efficiency of the conversion of HMF to FDCA. The rate constants for the conversion of HMF, FFCA, and FDCA over the catalysts at various temperatures are listed in Tables S5, S6, and S7, respectively. Rate constants for Pt@$Fe_{1.7}Cr_{0.3}O_3$ were consistently higher and increased with temperature, confirming its superior catalytic activity. The activation energies for the conversion of HMF and the formation of FDCA over Pt@$Fe_{1.7}Cr_{0.3}O_3$ were determined to be 67.9 and 36.3 kJ/mol, respectively, both of which are lower than those recorded for Pt@Fe_2O_3. These results suggest that the enhanced kinetics associated with Pt@$Fe_{1.7}Cr_{0.3}O_3$ are fundamental to its superior catalytic performance.

$$dC_{HMF}/dt = -K_1C_{DFF} \quad (R1)$$

$$dC_{DFF}/dt = K_1C_{HMF} - K_2C_{FFCA} \quad (R2)$$

$$dC_{FDCA}/dt = K_3C_{FFCA} \quad (R3)$$

To further explore the role of reactive oxygen species (ROS) in the oxidation of HMF over Pt@$Fe_{1.7}Cr_{0.3}O_3$, control experiments were conducted using 1,4-benzoquinone (PBQ) as a superoxide radical ($\cdot O_2^-$) scavenger and isopropanol (IPA) as a hydroxyl radical ($\cdot OH^-$) scavenger over a brief reaction period (3 h).[41,42] As illustrated in Figure S5, IPA exhibited a negligible effect on the catalytic oxidation of HMF, whereas the yield of FDCA was significantly diminished in the presence of PBQ. This observation indicates that O_2^- species, rather than hydroxyl radicals, play a predominant role in Pt@$Fe_{1.7}Cr_{0.3}O_3$-catalyzed oxidation of HMF to FDCA. Electron paramagnetic resonance (EPR) spectroscopy with DMPO as a spin-trapping agent further confirmed the generation of superoxide radicals ($\cdot O_2^-$) under the reaction conditions, as evidenced by the characteristic quartet signal of the DMPO−$\cdot O_2^-$ adduct (Figure S6). This result provides direct experimental evidence for the involvement of superoxide in the HMF oxidation pathway. It is well-established that the oxidation of HMF on noble metal-based catalysts is initiated through substrate adsorption onto noble metal sites, followed by dehydrogenation. This half-reaction generates reactive oxygen species that subsequently react with the abstracted hydrogen, resulting in the formation of water.[12,33,43,44]

3.4. Computational Characterization of Catalytic Reaction. We carried out computational mechanistic studies on HMF oxidation using Pt@Fe_2O_3 and Pt@$Fe_{1.7}Cr_{0.3}O_3$ using density functional theory (DFT). This theoretical investigation further explains the higher HMF oxidation activity of Pt@$Fe_{1.7}Cr_{0.3}O_3$ compared to that of Pt@Fe_2O_3. The computational results are illustrated in Figure 7, and details of the

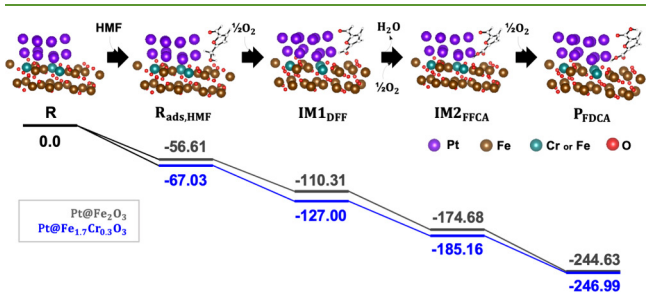


Figure 7. Calculated energy diagram of HMF oxidation to FDCA via DFF and FFCA catalyzed by Pt@Fe_2O_3 and Pt@$Fe_{1.7}Cr_{0.3}O_3$. The total electronic energies of all reactants, intermediates, and products are given in kcal/mol and were calculated using the PBE functional and the corresponding pseudopotentials.

computational methods are provided in Section S1 for details.[45−51] To model the catalyst structures, we built a two-layered Fe_2O_3(110) slab model with periodic boundary conditions and deposited a two-layered Pt_{12} cluster with a (111) facet above the Fe_2O_3 surface. The unit cell stoichiometry for modeling Pt@Fe_2O_3 was $Pt_{12}Fe_{24}O_{36}$. To

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

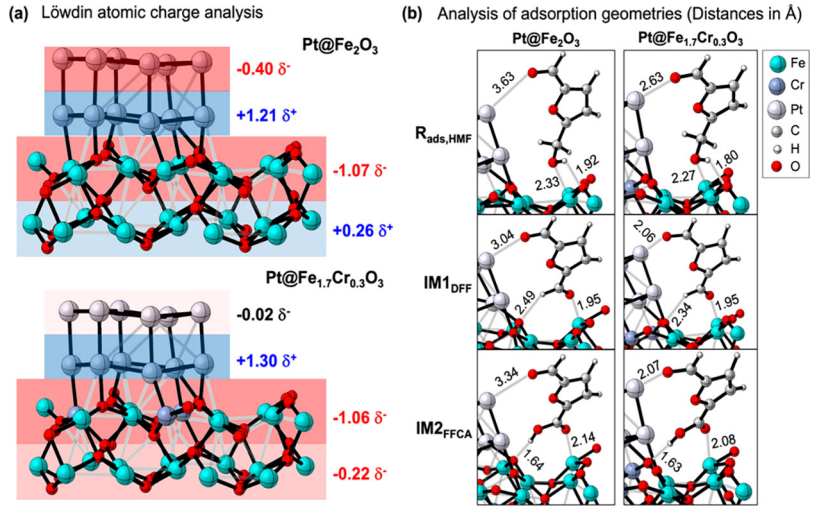


Figure 8. (a) Calculated layer-wise Löwdin atomic charges of $Pt@Fe_2O_3$ and $Pt@Fe_{1.7}Cr_{0.3}O_3$. (b) Adsorption geometries of HMF, DFF, and FFCA on $Pt@Fe_2O_3$ and $Pt@Fe_{1.7}Cr_{0.3}O_3$ with some key bond distances.

simulate $Pt@Fe_{1.7}Cr_{0.3}O_3$, we chose three surface Fe atoms that were near Pt but were the second nearest to each other, and they were replaced with Cr. The resulting unit cell had a stoichiometry of $Pt_{12}Fe_{21}Cr_3O_{36}$, which was close to the experimental molar ratio.

It should be emphasized that the optimized geometries of these two catalysts are consistent with the observations from the XPS experiments. The average Pt−O bond distances were 3.23 and 3.32 Å in the presence and absence of Cr, respectively, whereas the average Pt−Fe distances were 3.46 and 3.26 Å. The presence of Cr reduces the Pt−O bond length while elongating the Pt−Fe distance. This trend is in line with the XPS results, where PtO_x peaks were observed only for $Pt@Fe_{1.7}Cr_{0.3}O_3$, demonstrating that mechanistic studies using these structures would reasonably explain the experimental results. Moreover, we analyzed the projected density of states (PDOS) of Pt and O atoms at the active sites, compared their d-band and p-band center levels, and correlated them with the experimental surface acidity. The calculated d-band/p-band centers of Pt and O in $Pt@Fe_{1.7}Cr_{0.3}O_3$ are 0.02−0.08 eV higher than those in $Pt@Fe_2O_3$ (Figure S7). Higher band centers indicate higher surface acidity and, thus, higher HMF oxidation reactivity. These calculation results are also consistent with the higher acidity of $Pt@Fe_{1.7}Cr_{0.3}O_3$ than $Pt@Fe_2O_3$ measured experimentally (Table 1).

The above results indicate that incorporating Cr into the catalyst leads to synergistic interactions among Cr, Fe, O, and Pt and increases the catalytic activity for HMF oxidation. To further explain the catalytic activities, the reaction energetics were calculated for $Pt@Fe_2O_3$ and $Pt@Fe_{1.7}Cr_{0.3}O_3$. The HMF substrate was first adsorbed on the surface ($\mathbf{R} \rightarrow \mathbf{R_{ads,HMF}}$) with adsorption energies of −56.61 and −67.03 kcal/mol for $Pt@Fe_2O_3$ and $Pt@Fe_{1.7}Cr_{0.3}O_3$, respectively. After HMF adsorption, throughout the reaction via $\mathbf{IM1_{DFF}}$ and $\mathbf{IM2_{FFCA}}$, $Pt@Fe_{1.7}Cr_{0.3}O_3$ continues to undergo a pathway with intermediate energies 10−17 kcal/mol lower than that of $Pt@Fe_2O_3$. Notably, when calculating the reaction energetics, we assumed that O_2 molecules that were completely free from the catalyst surface were introduced to the substrate, after which the reactions took place. This assumption leads to the overestimation of stabilization energies (up to 100−200 kcal/mol); however, the relative comparison between $Pt@Fe_{1.7}Cr_{0.3}O_3$ and $Pt@Fe_2O_3$ remains feasible.[51,52] Figure 7 clearly shows that the increased acidity due to Cr doping leads to higher stabilization of intermediates and higher reactivity of $Pt@Fe_{1.7}Cr_{0.3}O_3$.

Additionally, we analyzed the electron distributions between Pt and $Fe_2O_3/Fe_{1.7}Cr_{0.3}O_3$ (Figure 8a) and the effect of electron distributions on the structures of the adsorbed substrates (Figure 8b). This analysis was conducted to provide a further chemical explanation for the different reactivities of HMF oxidation on $Pt@Fe_2O_3$ and $Pt@Fe_{1.7}Cr_{0.3}O_3$, as shown in Figure 7. As illustrated in Figure 8a, we performed a Löwdin population analysis of the DFT-optimized structures of $Pt@Fe_2O_3$ and $Pt@Fe_{1.7}Cr_{0.3}O_3$. The catalyst structure was partitioned into four layers: the lowest Fe_2O_3 layer, the second-lowest $Fe_2O_3/Fe_{1.7}Cr_{0.3}O_3$ layer interfacing with the Pt cluster, and the top two Pt layers. We then calculated the sum of the Löwdin charges of all atoms in each layer to quantify the extent of electron transfer among the layers. The charges of each layer of $Pt@Fe_2O_3$, from lowest to highest, were +0.26, −1.07, +1.21, and −0.40. Alternating positive and negative charges were observed, and the total atomic charges for Fe_2O_3 and Pt were −0.81 and +0.81, respectively. Meanwhile, both bottom layers of $Pt@Fe_{1.7}Cr_{0.3}O_3$ were negatively charged (−0.22 and −1.26 for the first and second layers, respectively), with a total charge of −1.28. These negative charges resulted in Pt layers with a total positive charge of +1.28, which was more electropositive than the Pt layers of $Pt@Fe_2O_3$ (total charge of +0.81).

It should be emphasized that the electron transfer trends shown in Figure 8a are consistent with our experiments. The calculation results demonstrate that the PtNPs in $Pt@Fe_{1.7}Cr_{0.3}O_3$ lose more electrons from the metal oxide support than those in $Pt@Fe_2O_3$. As discussed above, Cr incorporation reduces Pt−O distances, so oxygens can take more electrons from Pt, leading to more electropositive PtNPs, even though Cr has slightly less electronegativity than Fe. Such electron deficiency (i.e., higher Lewis acidity) resulted in a higher reactivity of HMF oxidation on $Pt@Fe_{1.7}Cr_{0.3}O_3$ than $Pt@Fe_2O_3$. The electron-deficient PtNPs in $Pt@Fe_{1.7}Cr_{0.3}O_3$ can take more electrons and hydrogens from HMF, thus facilitating the addition of oxygen to form FDCA. These results are also in line with the previous study regarding the HMF oxidation on

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

Ru@Nb-grafted SBA-15, where a higher Nb content resulted in more electron-deficient Ru and a higher HMF oxidation reactivity.[53]

The different electron density distributions shown in Figure 8a vary the adsorption geometry (Figure 8b), consistently explaining the higher catalytic activity of Pt@$Fe_{1.7}Cr_{0.3}O_3$ than Pt@Fe_2O_3. For $\mathbf{R_{ads,HMF}}$, the aldehyde oxygen of HMF interacts more closely with a Pt atom in Pt@$Fe_{1.7}Cr_{0.3}O_3$ at a distance of 2.63 Å, compared to that in Pt@Fe_2O_3 (3.63 Å). This shorter distance is attributed to the more electropositive Pt atoms of Pt@$Fe_{1.7}Cr_{0.3}O_3$ compared to Pt@Fe_2O_3. The O−H···O hydrogen bond lengths between HMF and the support were 1.80 and 1.92 Å for Pt@$Fe_{1.7}Cr_{0.3}O_3$ and Pt@Fe_2O_3, respectively. Pt@$Fe_{1.7}Cr_{0.3}O_3$ shows shorter bond distances for the substrate oxygen−Pt interaction and the O−H···O hydrogen bond, leading to a more stable $\mathbf{R_{ads,HMF}}$ than Pt@Fe_2O_3 (−67.03 vs −56.61 kcal/mol, Figure 7) and a higher oxidation reactivity. Notably, these $\mathbf{R_{ads,HMF}}$ structures were obtained from comprehensive conformational searches (details in the Supporting Information) and can be regarded as the most stable conformers closely related to the subsequent oxidation steps.

The other intermediates ($\mathbf{IM1_{DFF}}$ and $\mathbf{IM2_{FFCA}}$) exhibited similar trends in adsorption configurations. The substrate aldehyde oxygen−Pt bond distances in $\mathbf{IM1_{DFF}}$ are 2.04 and 3.06 Å for Pt@$Fe_{1.7}Cr_{0.3}O_3$ and Pt@Fe_2O_3, respectively, and those in $\mathbf{IM2_{FFCA}}$ are 2.07 and 3.34 Å, respectively. Pt@$Fe_{1.7}Cr_{0.3}O_3$ exhibited a closer interaction with the substrate for these two intermediates, leading to lower energies than the corresponding intermediates in Pt@Fe_2O_3 (Figure 7) and a higher catalytic activity than Pt@Fe_2O_3. All the above computational results are in line with our experimental observations and provide insights into the effects of electron densities and adsorption configurations on the tuning of catalytic activities by incorporating Cr.

The aerobic oxidation pathway of HMF to FDCA via 2,5-diformylfuran (DFF) and 5-formylfuran-2-carboxylic acid (FFCA) over the catalyst Pt@$Fe_{1.7}Cr_{0.3}O_3$ was elucidated by examining the oxidation intermediates generated during the reaction. Initially, the aldehyde group of HMF interacts with the Pt^0 active sites present in the asymmetric Pt−O−Fe structure formed during the synthesis of the PtNPs. Acidic centers of Pt@$Fe_{1.7}Cr_{0.3}O_3$ facilitate hydrolysis, leading to the formation of a geminal diol intermediate. Hydrogen temperature-programmed reduction (H_2-TPR) analysis (Figure 3e) indicated a low-temperature hydrogen consumption signal for Pt@$Fe_{1.7}Cr_{0.3}O_3$, which is attributed to the highly active and well-dispersed Pt−O−Fe sites. These sites demonstrate enhanced oxygen mobility and reactivity owing to their lower reducibility.[12,21,54]

The adsorbed oxygen is converted into active lattice oxygen at the oxygen vacancies and Pt^0 sites. The subsequent cleavage of the O−H bond in the geminal diol releases hydrogen, which reacts with the oxygen atoms to produce water and DFF. Further elimination of hydroxyl hydrogen from DFF results in the formation of a metal-alcohol salt intermediate. This intermediate then underwent hydrogen abstraction from the alcohol group to yield an aldehyde intermediate (FFCA). The final hydrolysis of the FFCA aldehyde side chain generated a second geminal diol intermediate, followed by hydrogen removal and interaction with oxygen to produce FDCA. During the reaction, the surface electrons consumed as Fe^{3+} and Cr^{6+} were reduced to Fe^{2+} and Cr^{3+}, respectively. Concurrently, gaseous O_2 dissociated into O^- and O_2^{2-} species, which are subsequently integrated into the active lattice oxygen (O_2^-), thereby regenerating Fe^{2+} and Cr^{3+} back to their original oxidation states of Fe^{3+} and Cr^{6+}, thus ensuring a reversible redox cycle.

The Pt@$Fe_{1.7}Cr_{0.3}O_3$ catalyst exhibits a distinctive Pt(111)/PtO_2(020) heterostructure, which is instrumental in reorienting the local electrons at the interface and optimizing the adsorption and activation of O_2, HMF, and its intermediates.[55,56] This synergistic interaction between the Cr−Fe−O support and PtNPs significantly enhances the selective oxidation of HMF to FDCA, underscoring its potential as an effective catalyst for sustainable oxidation processes.

3.5. Sustainability Considerations. Our catalytic system, Pt@$Fe_{1.7}Cr_{0.3}O_3$, is evaluated through multiple sustainability parameters. First, the reaction proceeds in aqueous medium under base-free conditions with molecular oxygen as the sole oxidant, thereby eliminating the need for corrosive additives or hazardous oxidants. Second, the reaction is conducted under mild conditions (120 °C, 1 bar O_2), which notably reduces energy consumption compared to conventional HMF oxidation approaches that often demand high pressure, strong bases, or organic solvents. Third, the catalyst demonstrates excellent durability and recyclability, retaining over 90% of its catalytic activity after at least five consecutive cycles, supporting long-term applicability. Fourth, the process achieves a nearly complete carbon balance (98−100%), underscoring atom efficiency and minimal byproduct formation. Together, these features underscore environmental compatibility, resource efficiency, and energy economy, thereby aligning strongly with the journal's sustainability framework and supporting the potential of our system for scalable, industrially relevant biomass valorization.

4. CONCLUSIONS

In this study, we successfully synthesized and characterized Pt@$Fe_{2-x}Cr_xO$ catalysts for the base-free aerobic oxidation of HMF to FDCA under mild conditions. The incorporation of chromium (Cr) into Fe_2O_3 significantly enhanced the catalytic performance by modulating the surface acidity, electronic interactions, and oxygen vacancy density, leading to improved oxidation efficiency and selectivity. Compared to conventional Fe_2O_3-based catalysts, Cr-modified Pt@$Fe_{2-x}Cr_xO$ supports exhibited higher surface areas, greater reducibility, and optimized acid site distributions, which collectively contributed to their superior catalytic activity.

Through comprehensive physicochemical characterization (XRD, BET, XPS, NH_3-TPD, and H_2-TPR), we elucidated the structural and electronic modifications induced by Cr substitution. Notably, NH_3-TPD analysis confirmed an increase in the number of weak and moderate acid sites, which facilitated the selective oxidation of HMF, whereas XPS data revealed an increase in the oxygen vacancy concentration, enhancing redox cycling and active oxygen species generation. These features directly contributed to the enhanced oxidation kinetics observed for Pt@$Fe_{2-x}Cr_xO_3$, which outperformed conventional Pt-based catalysts in terms of FDCA yield, stability, and recyclability.

Catalytic performance evaluations demonstrated that Pt@$Fe_{1.7}Cr_{0.3}O_3$ achieved nearly 100% HMF conversion with a 78.7% yield of FDCA, surpassing most state-of-the-art noble metal and transition metal oxide catalysts. Furthermore, the recoverability of Pt@$Fe_{2-x}Cr_xO_3$ enabled facile separation and

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

reusability, retaining over 90% of its original catalytic activity after five consecutive cycles. Mechanistic investigations and kinetic studies further validated that the Cr-enhanced oxygen vacancies and synergistic Pt−support interactions played a pivotal role in optimizing the HMF oxidation pathways, minimizing overoxidation side reactions, and enhancing selectivity.

The insights gained from this study provide a rational design strategy for next-generation metal oxide-supported catalysts with tunable acid−base properties and electronic structures for biomass valorization. The Pt@$Fe_{2-x}Cr_xO_3$ system represents a promising platform for the sustainable conversion of biobased feedstocks into high-value chemicals, offering an economically viable and environmentally friendly alternative to conventional catalytic methods. Future studies will focus on further tuning metal−support interactions and expanding the applicability of these catalysts to other biomass-derived transformations, paving the way for industrial-scale deployment in green chemistry and renewable energy sectors.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssuschemeng.5c07366.

Photograph of the high-pressure reactor; FE-SEM images and elemental mapping of supports and catalysts; high-resolution XPS spectra after repeated cycles; elemental composition from XRF analysis; N_2 sorption, reduction temperature, and acidity profiles; catalytic performance under different atmospheres; comparative literature data of noble metal catalysts; kinetic parameters of HMF, DFF, and FFCA conversions; quenching experiments; EPR spectra of superoxide radicals; computational methods and parameters; projected density of states analysis (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Hangil Lee − *Department of Chemistry and Research Institute of Natural Sciences, Sookmyung Women's University, Seoul 04310, Republic of Korea;* orcid.org/0000-0002-0657-5660; Email: easyscan@sookmyung.ac.kr

Yeonjoon Kim − *BB21 Plus Program, Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea*; Email: yeonjoonkim@pknu.ac.kr

Hyun Sung Kim − *BB21 Plus Program, Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea;* orcid.org/0000-0002-9132-3985; Email: kimhs75@pknu.ac.kr

Authors

Santhana Krishnan Perumal − *BB21 Plus Program, Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea*

Hyejin Yu − *BB21 Plus Program, Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea*

Hyeonsu Je − *BB21 Plus Program, Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssuschemeng.5c07366

Author Contributions

‡S.K.P. and H.Y. contributed equally. S.K.P.: conceptualization, investigation, formal analysis, and writing of the original draft; H.Y.: investigation and formal analysis; H.J.: theoretical calculation; Y.K.: theoretical calculation and writing of the original draft; H.L.: formal analysis and writing of the original draft, H.S.K.: conceptualization, supervision, funding acquisition, and review and editing. All authors contributed to discussion and manuscript preparation.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean government (MSIP) (NRF-2023R1A2C1007484) and Core Research Institute (CRI) program and Basic Science Research program through the NRF under program number 2022R1A6A1A03051158. Y.K. acknowledges that this work was supported by the National Supercomputing Center with supercomputing resources including technical support (KSC-2024-CRE-0557). H.L. acknowledges the National Research Foundation of Korea (NRF) grant, funded by the Korean government (MSIT) (RS-2024-00346153 and RS-2025-21072982).

■ REFERENCES

(1) Rosenfeld, C.; Konnerth, J.; Sailer-Kronlachner, W.; Solt, P.; Rosenau, T.; van Herwijnen, H. W. G. Current Situation of the Challenging Scale-Up Development of Hydroxymethylfurfural Production. *ChemSusChem* **2020**, *13* (14), 3544−3564.

(2) Sajid, M.; Zhao, X.; Liu, D. Production of 2,5-Furandicarboxylic Acid (FDCA) from 5-Hydroxymethylfurfural (HMF): Recent Progress Focusing on the Chemical-Catalytic Routes. *Green Chem.* **2018**, *20* (24), 5427−5453.

(3) Das, S.; Cibin, G.; Walton, R. I. Selective Oxidation of Biomass-Derived 5-Hydroxymethylfurfural Catalyzed by an Iron-Grafted Metal−Organic Framework with a Sustainably Sourced Ligand. *ACS Sustain. Chem. Eng.* **2024**, *12* (14), 5575−5585.

(4) Liao, Y.-T.; Nguyen, V. C.; Ishiguro, N.; Young, A. P.; Tsung, C.-K.; Wu, K. C.-W. Engineering a Homogeneous Alloy-Oxide Interface Derived from Metal-Organic Frameworks for Selective Oxidation of 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid. *Appl. Catal. B Environ.* **2020**, *270*, No. 118805.

(5) Chen, C.; Li, X.; Wang, L.; Liang, T.; Wang, L.; Zhang, Y.; Zhang, J. Highly Porous Nitrogen- and Phosphorus-Codoped Graphene: An Outstanding Support for Pd Catalysts to Oxidize 5-Hydroxymethylfurfural into 2,5-Furandicarboxylic Acid. *ACS Sustain. Chem. Eng.* **2017**, *5* (12), 11300−11306.

(6) Zhang, Z.; Zhen, J.; Liu, B.; Lv, K.; Deng, K. Selective Aerobic Oxidation of the Biomass-Derived Precursor 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid under Mild Conditions over a Magnetic Palladium Nanocatalyst. *Green Chem.* **2015**, *17* (2), 1308−1317.

(7) Gao, T.; Yin, Y.; Zhu, G.; Cao, Q.; Fang, W. Co_3O_4 NPs Decorated Mn-Co-O Solid Solution as Highly Selective Catalyst for Aerobic Base-Free Oxidation of 5-HMF to 2,5-FDCA in Water. *Catal. Today* **2020**, *355*, 252−262.

(8) Han, X.; Geng, L.; Guo, Y.; Jia, R.; Liu, X.; Zhang, Y.; Wang, Y. Base-Free Aerobic Oxidation of 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid over a Pt/C−O−Mg Catalyst. *Green Chem.* **2016**, *18* (6), 1597−1604.

(9) Hou, Q.; Gao, T.; Zhang, H.; Tang, Q.; Fang, W. Pt Nanoparticles in Cooperation with Mn-P Composite Drive Base-Free Selective Oxidation of 5-Hydroxymethylfurfural. *Catal. Today* **2025**, *443*, No. 114990.

(10) Chen, H.; Shen, J.; Chen, K.; Qin, Y.; Lu, X.; Ouyang, P.; Fu, J. Atomic Layer Deposition of Pt Nanoparticles on Low Surface Area Zirconium Oxide for the Efficient Base-Free Oxidation of 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid. *Appl. Catal. Gen.* **2018**, *555*, 98−107.
(11) Ke, C.; Li, M.; Fan, G.; Yang, L.; Li, F. Pt Nanoparticles Supported on Nitrogen-Doped-Carbon-Decorated CeO_2 for Base-Free Aerobic Oxidation of 5-Hydroxymethylfurfural. *Chem. − Asian J.* **2018**, *13* (18), 2714−2722.
(12) Zhang, H.; Gao, T.; Cao, Q.; Fang, W. Tailoring the Reactive Oxygen Species in Mesoporous NiO for Selectivity-Controlled Aerobic Oxidation of 5-Hydroxymethylfurfural on a Loaded Pt Catalyst. *ACS Sustain. Chem. Eng.* **2021**, *9* (17), 6056−6067.
(13) Zhang, H.; Gao, T.; Zhang, Q.; Gu, B.; Tang, Q.; Cao, Q.; Fang, W. Synergistic Catalysis in Loaded PtRu Alloy Nanoparticles to Boost Base-Free Aerobic Oxidation of 5-Hydroxymethylfurfural. *Mater. Today Catal.* **2023**, *3*, No. 100013.
(14) Zhu, W.; Meng, Y.; Yang, C.; Zhao, J.; Wang, H.; Hu, W.; Lv, G.; Wang, Y.; Deng, T.; Hou, X. Effect of Coordination Environment Surrounding a Single Pt Site on the Liquid-Phase Aerobic Oxidation of 5-Hydroxymethylfurfural. *ACS Appl. Mater. Interfaces* **2021**, *13* (41), 48582−48594.
(15) Cheng, X.; Li, S.; Liu, S.; Xin, Y.; Yang, J.; Chen, B.; Liu, H. Highly Efficient Catalytic Oxidation of 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid Using Bimetallic Pt−Cu Alloy Nanoparticles as Catalysts. *Chem. Commun.* **2022**, *58* (8), 1183−1186.
(16) Zhang, H.; Zhang, R.; Zhang, W.; Gu, B.; Tang, Q.; Cao, Q.; Fang, W. Base-Free Selective Oxidation of 5-Hydroxymethylfurfural over Pt Nanoparticles on Surface Nb-Enriched Co-Nb Oxide. *Appl. Catal. B Environ.* **2023**, *330*, No. 122670.
(17) Zhou, C.; Shi, W.; Wan, X.; Meng, Y.; Yao, Y.; Guo, Z.; Dai, Y.; Wang, C.; Yang, Y. Oxidation of 5-Hydroxymethylfurfural over a Magnetic Iron Oxide Decorated rGO Supporting Pt Nanocatalyst. *Catal. Today* **2019**, *330*, 92−100.
(18) Wu, J.; Xie, W.; Zhang, Y.; Ke, X.; Li, T.; Fang, H.; Sun, Y.; Zeng, X.; Lin, L.; Tang, X. Oxygen-Vacancy-Rich MnO_x Supported RuO_x for Efficient Base-Free Oxidation of 5-Hydroxymethylfurfural and 5-Methoxymethylfurfural to 2,5-Furandicarboxylic Acid. *J. Energy Chem.* **2024**, *95*, 670−683.
(19) Antonyraj, C. A.; Huynh, N. T. T.; Lee, K. W.; Kim, Y. J.; Shin, S.; Shin, J. S.; Cho, J. K. Base-Free Oxidation of 5-Hydroxymethyl-2-Furfural to 2,5-Furan Dicarboxylic Acid over Basic Metal Oxide-Supported Ruthenium Catalysts under Aqueous Conditions. *J. Chem. Sci.* **2018**, *130*, 156.
(20) Mishra, D. K.; Lee, H. J.; Kim, J.; Lee, H.-S.; Cho, J. K.; Suh, Y.-W.; Yi, Y.; Kim, Y. J. $MnCo_2O_4$ Spinel Supported Ruthenium Catalyst for Air-Oxidation of HMF to FDCA under Aqueous Phase and Base-Free Conditions. *Green Chem.* **2017**, *19* (7), 1619−1623.
(21) Gao, T.; Chen, J.; Fang, W.; Cao, Q.; Su, W.; Dumeignil, F. $Ru/Mn_XCe_1O_Y$ Catalysts with Enhanced Oxygen Mobility and Strong Metal-Support Interaction: Exceptional Performances in 5-Hydroxymethylfurfural Base-Free Aerobic Oxidation. *J. Catal.* **2018**, *368*, 53−68.
(22) Megías-Sayago, C.; Lolli, A.; Ivanova, S.; Albonetti, S.; Cavani, F.; Odriozola, J. A. Au/Al_2O_3 − Efficient Catalyst for 5-Hydroxymethylfurfural Oxidation to 2,5-Furandicarboxylic Acid. *Catal. Today* **2019**, *333*, 169−175.
(23) Zhang, R.; He, H.; Tang, Y.; Zhang, Z.; Zhou, H.; Yu, J.; Zhang, L.; Dai, B. A Review on Fe_2O_3-Based Catalysts for Toluene Oxidation: Catalysts Design and Optimization with the Formation of Abundant Oxygen Vacancies. *ChemCatChem.* **2024**, *16* (18), No. e202400396.
(24) Popov, N.; Krehula, S.; Ristić, M.; Kuzmann, E.; Homonnay, Z.; Bošković, M.; Stanković, D.; Kubuki, S.; Musić, S. Influence of Cr Doping on the Structural, Magnetic, Optical and Photocatalytic Properties of α-Fe_2O_3 Nanorods. *J. Phys. Chem. Solids* **2021**, *148*, No. 109699.
(25) Chen, C.; Wang, L.; Zhu, B.; Zhou, Z.; El-Hout, S. I.; Yang, J.; Zhang, J. 2,5-Furandicarboxylic Acid Production via Catalytic Oxidation of 5-Hydroxymethylfurfural: Catalysts, Processes and Reaction Mechanism. *J. Energy Chem.* **2021**, *54*, 528−554.
(26) Shen, J.; Chen, H.; Chen, K.; Qin, Y.; Lu, X.; Ouyang, P.; Fu, J. Atomic Layer Deposition of a Pt-Skin Catalyst for Base-Free Aerobic Oxidation of 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid. *Ind. Eng. Chem. Res.* **2018**, *57* (8), 2811−2818.
(27) Aranha, D. J.; Gogate, P. R. A Review on Green and Efficient Synthesis of 5-Hydroxymethylfurfural (HMF) and 2,5-Furandicarboxylic Acid (FDCA) from Sustainable Biomass. *Ind. Eng. Chem. Res.* **2023**, *62* (7), 3053−3078.
(28) Wu, W.; Hao, R.; Liu, F.; Su, X.; Hou, Y. Single-Crystalline α-Fe_2O_3 Nanostructures: Controlled Synthesis and High-Index Plane-Enhanced Photodegradation by Visible Light. *J. Mater. Chem. A* **2013**, *1* (23), 6888−6894.
(29) Atram, R. R.; Bhuse, D. V.; Bhuse, V. M.; Atram, R. G.; Kondawar, S. B. Evolution of Waste Iron Rust into α-Fe_2O_3CNF and α-Fe_2O_3/PANI Composites as an Efficient Positive Electrode for Sustainable Hybrid Supercapacitor. *J. Mater. Sci.: Mater. Electron.* **2021**, *32* (17), 13787−13802.
(30) Murphin Kumar, P. S.; Thiripuranthagan, S.; Imai, T.; Kumar, G.; Pugazhendhi, A.; Vijayan, S. R.; Esparza, R.; Abe, H.; Krishnan, S. K. Pt Nanoparticles Supported on Mesoporous CeO2 Nanostructures Obtained through Green Approach for Efficient Catalytic Performance toward Ethanol Electro-Oxidation. *ACS Sustain. Chem. Eng.* **2017**, *5* (12), 11290−11299.
(31) Zhang, N.; Wang, J.; Li, Q.; Xin, Y.; Zheng, L.; Wang, Y.; Zhang, Z. Enhanced Selective Catalytic Reduction of NO_x with NH_3 over Homoatomic Dinuclear Sites in Defective α-Fe_2O_3. *Chem. Eng. J.* **2021**, *426*, No. 131845.
(32) Fan, Y.; Zhang, J.; Yang, L.; Lu, M.; Ying, T.; Deng, B.; Dai, W.; Luo, X.; Zou, J.; Luo, S. Enhancing SO2-Shielding Effect and Lewis Acid Sites for High Efficiency in Low-Temperature SCR of NO with NH3: Reinforced Electron-Deficient Extent of Fe3+ Enabled by Ti4+ in Fe2O3. *Sep. Purif. Technol.* **2023**, *311*, No. 123272.
(33) Zhang, S.; Yang, Y.; Chu, G.; Wang, S.; Wang, C.; Zhang, Y.; Zhang, L.; Mei, J. Surface Chemistry Regulation on Particle-Support Interaction of Ruthenium and Cr-Fe Oxides for Selective Oxidation of 5-Hydroxymethylfurfural. *Chem. Eng. J.* **2023**, *474*, No. 145670.
(34) Wang, Y.; Song, G.; Xu, Z.; Rosei, F.; Ma, D.; Chen, G. Interfacial Reaction-Directed Synthesis of a Ceria Nanotube-Embedded Ultra-Small Pt Nanoparticle Catalyst with High Catalytic Activity and Thermal Stability. *J. Mater. Chem. A* **2016**, *4* (37), 14148−14154.
(35) Mazzotta, E.; Rella, S.; Turco, A.; Malitesta, C. XPS in Development of Chemical Sensors. *RSC Adv.* **2015**, *5* (101), 83164−83186.
(36) Wang, Z.; Xi, L.; Yang, Y.; Li, Y.; Han, X.; Zuo, Y.; Wang, J. Spin-Dependent Transport Properties of CrO_2 Micro Rod. *Nano-Micro Lett.* **2014**, *6* (4), 365−371.
(37) Luo, J.-W.; Song, J.-D.; Jia, W.-Z.; Pu, Z.-Y.; Lu, J.-Q.; Luo, M.-F. Catalytic Dehydrofluorination of 1,1,1,3,3-Pentafluoropropane to 1,3,3,3-Tetrafluoropropene over Fluorinated NiO/Cr_2O_3 Catalysts. *Appl. Surf. Sci.* **2018**, *433*, 904−913.
(38) Wang, J.-C.; Ren, J.; Yao, H.-C.; Zhang, L.; Wang, J.-S.; Zang, S.-Q.; Han, L.-F.; Li, Z.-J. Synergistic Photocatalysis of Cr(VI) Reduction and 4-Chlorophenol Degradation over Hydroxylated α-Fe_2O_3 under Visible Light Irradiation. *J. Hazard. Mater.* **2016**, *311*, 11−19.
(39) Davis, S. E.; Benavidez, A. D.; Gosselink, R. W.; Bitter, J. H.; de Jong, K. P.; Datye, A. K.; Davis, R. J. Kinetics and Mechanism of 5-Hydroxymethylfurfural Oxidation and Their Implications for Catalyst Development. *J. Mol. Catal. Chem.* **2014**, *388*−*389*, 123−132.
(40) Donoeva, B.; Masoud, N.; de Jongh, P. E. Carbon Support Surface Effects in the Gold-Catalyzed Oxidation of 5-Hydroxymethylfurfural. *ACS Catal.* **2017**, *7* (7), 4581−4591.
(41) Li, W.; Cheng, C.; Zhao, J.; Song, Y.; Xue, C. Enhanced Azo Dye Removal through Sequential Ultrasound-Assisted-Treatment and Photocatalysis Using CdZnS. *Angew. Chem.* **2025**, *137* (11), No. e202425508.

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

(42) Song, R.; Yao, J.; Yang, M.; Ye, Z.; Xie, Z.; Zeng, X. Active Site Regulated Z-Scheme MIL-101(Fe)/Bi_2WO_6/Fe(III) with the Synergy of Hydrogen Peroxide and Visible-Light-Driven Photo-Fenton Degradation of Organic Contaminants. *Nanoscale* **2022**, *14* (18), 7055−7074.
(43) Wei, Y.; Zhang, Y.; Chen, Y.; Wang, F.; Cao, Y.; Guan, W.; Li, X. Crystal Faces-Tailored Oxygen Vacancy in Au/CeO_2 Catalysts for Efficient Oxidation of HMF to FDCA. *ChemSusChem* **2022**, *15* (13), No. e202101983.
(44) Liu, Y.; Ma, H.-Y.; Lei, D.; Lou, L.-L.; Liu, S.; Zhou, W.; Wang, G.-C.; Yu, K. Active Oxygen Species Promoted Catalytic Oxidation of 5-Hydroxymethyl-2-Furfural on Facet-Specific Pt Nanocrystals. *ACS Catal.* **2019**, *9* (9), 8306−8315.
(45) Giannozzi, P.; Baseggio, O.; Bonfà, P.; Brunato, D.; Car, R.; Carnimeo, I.; Cavazzoni, C.; de Gironcoli, S.; Delugas, P.; Ferrari Ruffino, F.; Ferretti, A.; Marzari, N.; Timrov, I.; Urru, A.; Baroni, S. Quantum ESPRESSO toward the Exascale. *J. Chem. Phys.* **2020**, *152* (15), No. 154105.
(46) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865−3868.
(47) Dal Corso, A. Pseudopotentials Periodic Table: From H to Pu. *Comput. Mater. Sci.* **2014**, *95*, 337−350.
(48) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456−1465.
(49) Sohier, T.; Calandra, M.; Mauri, F. Density Functional Perturbation Theory for Gated Two-Dimensional Heterostructures: Theoretical Developments and Application to Flexural Phonons in Graphene. *Phys. Rev. B* **2017**, *96* (7), No. 075448.
(50) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. CREST—A Program for the Exploration of Low-Energy Molecular Chemical Space. *J. Chem. Phys.* **2024**, *160* (11), No. 114110.
(51) Yizengaw, K. W.; Abay, T. A.; Ayele, D. W.; Jiang, J.-C. The Remarkable Performance of a Single Iridium Atom Supported on Hematite for Methane Activation: A Density Functional Theory Study. *RSC Adv.* **2022**, *12*, 23736−23746.
(52) Zhang, X.; Klaver, P.; van Santen, R.; van de Sanden, M. C. M.; Bieberle-Hütter, A. Oxygen Evolution at Hematite Surfaces: The Impact of Structure and Oxygen Vacancies on Lowering the Overpotential. *J. Phys. Chem. C* **2016**, *120* (32), 18201−18208.
(53) Perumal, S. K.; Lee, S.; Yu, H.; Heo, J.; Kang, M. J.; Kim, Y.; Park, M.; Lee, H.; Kim, H. S. Synergistic Interaction between Ruthenium Catalysts and Grafted Niobium on SBA-15 for 2,5-Furandicarboxylic Acid Production Using 5-Hydroxymethylfurfural. *ACS Appl. Mater. Interfaces* **2024**, *16* (6), 7353−7363.
(54) Mo, S.; Zhang, Q.; Zhang, M.; Zhang, Q.; Li, J.; Fu, M.; Wu, J.; Chen, P.; Ye, D. Elucidating the Special Role of Strong Metal−Support Interactions in Pt/MnO2 Catalysts for Total Toluene Oxidation. *Nanoscale Horiz.* **2019**, *4* (6), 1425−1433.
(55) Ren, Z.; Yang, Y.; Wang, S.; Niu, Y.; Wang, L.; Yu, J.; Song, Y.; Zhang, X.; Zhang, B.; Wei, M. Heterogeneous Interface Catalysts with Electron Local Exchange toward Highly Selective Oxidation of Biomass Platform Compounds. *ACS Catal.* **2023**, *13* (8), 5665−5677.
(56) Zubair, M.; Usov, P. M.; Ohtsu, H.; Yuwono, J. A.; Gerke, C. S.; Foley, G. D. Y.; Hackbarth, H.; Webster, R. F.; Yang, Y.; Lie, W. H.; Ma, Z.; Thomsen, L.; Kawano, M.; Bedford, N. M. Vacancy Mediated Electrooxidation of 5-Hydroxymethyl Furfuryl Using Defect Engineered Layered Double Hydroxide Electrocatalysts. *Adv. Energy Mater.* **2024**, *14* (35), 2400676.

https://doi.org/10.1021/acssuschemeng.5c07366
ACS Sustainable Chem. Eng. 2025, 13, 17470−17482

small structures

www.small-structures.com

Enhancing the Multifunctional Photocatalytic Activity of Sustainable Magnetic Nanoparticles

Young-Hoon Kim, Minjeong Shin, Eun-Byeol Park, Sunyoung Hwang, Jung A. Hong, Jae Hyuck Jang, Young-Min Kim, and Hangil Lee**

The development of efficient photocatalytic materials has intensified in response to increasing emphasis on sustainable energy conversion and environmental restoration. However, the excessive use and indiscriminate release of heavy metals from photocatalytic nanoparticles pose potential environmental risks. This study provides insights into optimizing visible-light-active photocatalysts to enhance their photocatalytic properties and stability. Specifically, cobalt doping and controlled pH modulation are employed on the modified Fe_3O_4, forming two distinct samples: Co@Fe_3O_4-B (base-treated) and Co@Fe_3O_4-A (acid-treated) nanoparticles. Microscopic characterization reveals that Co@Fe_3O_4-A undergoes a phase transition to hematite, whereas Co@Fe_3O_4-B retains its mixed-phase configuration. Spectroscopic analyses confirm that the cobalt dopants decorate the outskirts of each nanoparticle, forming a core–shell structure. However, Co@Fe_3O_4-B exhibit Co^{2+} states with a high oxygen vacancy content, whereas Co@Fe_3O_4-A contain mixed $Co^{2+/3+}$ states. The high density of defect states in the Co@Fe_3O_4-B results in superior photocatalytic efficiency, achieving near-complete oxidation of furfuraldehyde (97% conversion) and 5-hydroxymethyl-furfural (94% conversion), as well as effective degradation of toluene (78% conversion). This study demonstrates that the combined approach of doping and pH treatment is promising for the surface-defect engineering of photocatalysts, enhancing their multifunctional performance and reusability for sustainable energy conversion.

1. Introduction

The increasing demand for renewable energy has propelled the development of sustainable and efficient strategies for highly efficient photocatalytic applications.[1] Semiconductor photocatalysts, notably TiO_2 and ZnO, offer a promising solution owing to their favorable bandgaps, extended charge carrier lifetimes, and photocatalytic efficiencies.[2] However, the widespread application of semiconductor photocatalysts is often constrained by inherent limitations, including low activity under visible light and secondary environmental pollution problems due to difficulties in their recovery rate and reusability. To overcome these issues, it is imperative to develop photocatalysts with facile separation and recycling capabilities without additional energy input.[3]

Magnetite (Fe_3O_4) nanoparticles (NPs) are potential candidates for visible-light active photocatalysts due to their intrinsic bandgap (≈1.8 eV). In addition, their ferrimagnetic properties facilitate recovery by magnetic separation. Recent reviews have highlighted the potential of magnetic iron oxide NPs for environmental

Y.-H. Kim
Center for Nanophase Materials Sciences
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA

M. Shin
School of Chemistry and Energy
Sungshin Women's University
Seoul 01133, Republic of Korea

E.-B. Park, J. A. Hong, Y.-M. Kim
Department of Energy Science
Sungkyunkwan University (SKKU)
Suwon 16419, Republic of Korea
E-mail: youngmk@skku.edu

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/sstr.202500123.

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/sstr.202500123

S. Hwang, H. Lee
Department of Chemistry and Research Institute of Natural Sciences
Sookmyung Women's University
Seoul 04310, Republic of Korea
E-mail: easyscan@sookmyung.ac.kr

J. H. Jang
AI Enhanced Electron Microscopy Research Group
Korea Basic Science Institute (KBSI)
Daejeon 34133, Republic of Korea

Y.-M. Kim
Center for 2D Quantum Heterostructures
Institute for Basic Science (IBS)
Suwon 16419, Republic of Korea

remediation, citing their high catalytic activity, magnetic recoverability, and compatibility with green synthesis methods.[4] Furthermore, Fe-based magnetic photocatalysts can enhance visible-light responsiveness through bandgap modulation and intra-gap state formation, improving charge separation and reactive oxygen species generation.[5] These unique characteristics, combined with the application of external magnetic fields, reported to synergistically enhance photocatalytic activity through photomagnetic coupling, further underscore their promise in water treatment systems.[6] However, the inherent photocatalytic activity of Fe_3O_4 NPs remains suboptimal, necessitating the development of strategies to enhance their performance. Approaches such as surface modification, heterojunction formation, and defect engineering have been investigated to improve catalytic performance. Notably, metal ion doping, and pH modulation have shown considerable promise in tailoring the electronic structure and performance of Fe_3O_4 NPs,[7] which introduce structural alterations and generate oxygen vacancy defects that serve as active catalytic sites.[8]

This study introduces modified Fe_3O_4 NPs that have been doped with Co ions and then treated with pH control to enhance their photocatalytic and magnetic properties. Cobalt ions, which provide a variety of oxidation states, can introduce active sites and increase magnetic reactivity compared to iron.[9] By systematically adjusting the pH conditions of Co ion-doped Fe_3O_4, we propose two distinct samples (Co@Fe_3O_4-A (acid-treated: pH = 1.0) and Co@ Fe_3O_4-B (base-treated: pH = 13.0) NPs) and elucidate how these treatments influence the surface structure and phase transition, thereby affecting their properties. Structural characterizations revealed that Co@Fe_3O_4-A NPs promoted a nearly complete phase transition to hematite (α-Fe_2O_3), whereas Co@Fe_3O_4-B NPs consisted of a mixed-phase composition of Fe_3O_4 and α-Fe_2O_3. Multiscale spectroscopies based on X-ray photoelectron spectroscopy (XPS), X-ray absorption spectroscopy (XAS), and electron energy loss spectroscopy (EELS) clearly confirmed that Co doping was prominently localized on the surface, forming a chemically modified core–shell structure. In addition, Co@Fe_3O_4-B NPs exhibited uniform Co^{2+} oxidation states accompanied by oxygen vacancies, whereas Co@Fe_3O_4-A NPs displayed mixed oxidation states of Co^{3+} and Co^{2+}.

To address the versatility of these enhanced photocatalysts, three potential applications in the photocatalytic conversion reactions (PCR) of furfuraldehyde, 5-hydroxymethylfurfural, and toluene were investigated. Furfuraldehyde, a widely available biomass compound, undergoes transformation into low-toxic derivates, producing high-value compounds. 5-hydroxymethylfurfural is a biomass-derived platform chemical with its oxidation to 2,5-furandicarboxylic acid, providing a green alternative to petroleum-based terephthalic acid.[10] Toluene requires degradation due to its high volatility, environmental persistence, and toxic effects on ecosystems. Here, Co@Fe_3O_4-B NPs demonstrated multifunctionality as photocatalysts, achieving the superior oxidation of furfuraldehyde (97% conversion rate) and 5-hydroxymethylfurfural (94% conversion rate), as well as the degradation of toluene (78% conversion rate). This broad range of activity, coupled with the enhancement in performance compared to the pristine Fe_3O_4 NPs (41% improvement for furfuraldehyde oxidation, 45% for 5-hydroxymethylfurfural oxidation, and 73% for toluene degradation), are likely attributed to the formation of oxygen vacancies and lattice contraction at the NP surface. Furthermore, these NPs displayed consistent reusability (≈99%) over 10 cycles. This study provides insights into the optimal surface structural modifications of photocatalytic materials, highlighting the potential for advancing visible-light-driven photocatalysts in sustainable and efficient energy applications.

2. Results and Discussion

2.1. Crystal Structure Characterization

Among the several strategies for metal doping, Co was selected as the dopant because of its multivalent nature and potential to enhance magnetic responsiveness, with a nominal doping level of ≈3.5 wt%. Subsequently, pH treatment was introduced to further refine the crystal structure of the Co-doped Fe_3O_4 NPs. First, to assess the impact of pH modulation (**Figure 1**), a comparative analysis with pristine Fe_3O_4 NPs was undertaken via powder X-ray diffraction (XRD). As illustrated in Figure 1a, the diffraction peaks of the Fe_3O_4 NPs align with the characteristic reflection values (30.16°, 35.52°, 37.13°, 43.14°, 53.50°, 57.04°, and 62.64°), consistent with the Fe_3O_4 phase (PDF 01-1111).[11] Meanwhile, Co@Fe_3O_4-A NPs exhibited the diffraction peaks (24.15°, 33.14°, 35.63°, 40.87°, 49.45°, 54.03°, 57.55°, and 62.43°) of the α-Fe_2O_3 phase (PDF 33-0664).[12] Interestingly, Co@Fe_3O_4-B NPs displayed mixed-phase characteristics with peaks corresponding to both Fe_3O_4 and α-Fe_2O_3, suggesting a partial conversion of Fe_3O_4 to α-Fe_2O_3. The varying intensity of the α-Fe_2O_3 peaks across the pH-treated Fe_3O_4 samples indicates that the conversion of Fe_3O_4 to α-Fe_2O_3 is affected by the pH level of the solution. The relevant chemical reactions observed during the phase transition under acidic conditions are represented by Equation (1)

$$2Fe_3O_4 + 2H_3O^+ \rightarrow 3Fe_2O_3 + 3H_2 + 1/2O_2 \quad (1)$$

These observations are consistent with the Pourbaix diagram for iron, which illustrates the thermodynamic stability of various iron oxides as a function of pH and potential.[13] According to this diagram, Fe_3O_4 tends to oxidize to Fe_2O_3 under acidic conditions due to its increased stability at lower pH levels. In other words, the Co@Fe_3O_4-A NPs underwent complete conversion to α-Fe_2O_3. At higher pH values, the stability region of Fe_3O_4 overlaps with those of $Fe(OH)_2$ and FeOOH, indicating possible phase transformations under basic conditions. The presence of OH^- ions not only facilitate these transformations but may also induces surface defects and lattice distortions,[14] leading to peak shifts toward larger angles in the magnified XRD pattern (Figure 1b,c). This implies that basic conditions promote partial oxidation and restructuring of NPs, resulting in nanoparticles with mixed phases of α-Fe_2O_3 and Fe_3O_4 with structural modification. The chemical reaction for the mixed phase observed under basic conditions is expected to be Equation (2)

$$4Fe_3O_4 + 2OH^- \rightarrow 3Fe_2O_3 + 2Fe_3O_4 + H_2 + 1/2O_2 \quad (2)$$

To further examine the effect of the pH treatment on the morphological changes and phase configurations, bright-field

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

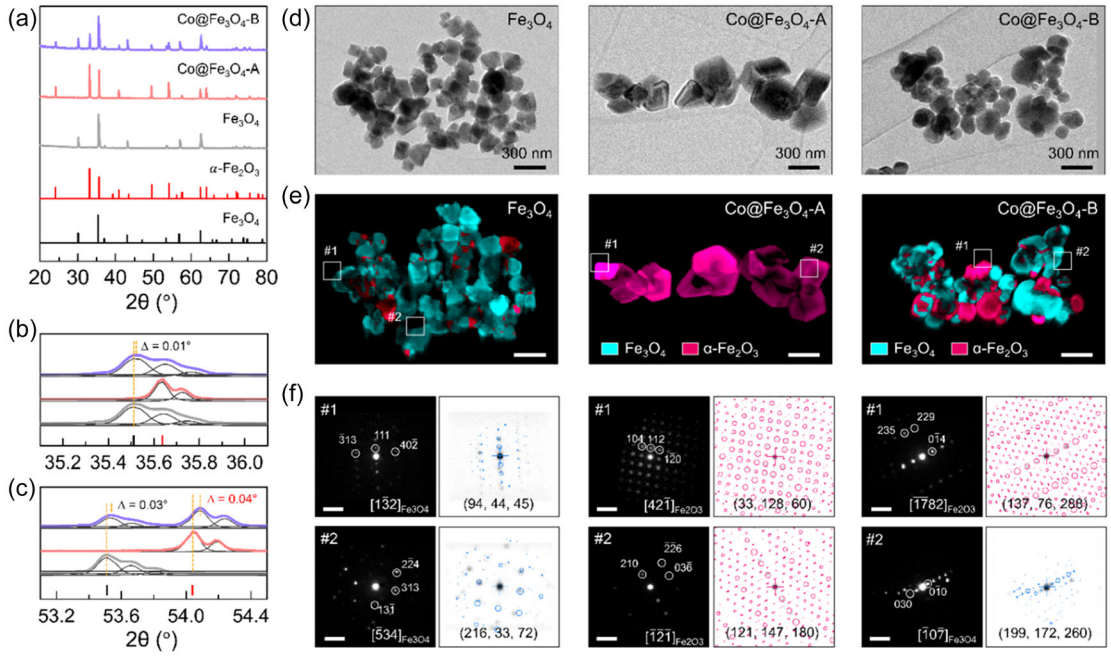


Figure 1. Effect of pH modulation on the crystal structure and morphology of Co-doped Fe_3O_4 NPs. a) XRD of the Fe_3O_4 (gray), Co@Fe_3O_4-A (red), and Co@Fe_3O_4-B (violet) NPs (Fe_3O_4: PDF 01-1111, α-Fe_2O_3: PDF 33-0664). b,c) Magnified XRD analysis showing peak shifts. d,e) BFTEM images and corresponding crystal-phase distribution obtained from SEND mapping. False-color mapping indicates Fe_3O_4 ($Fd\bar{3}m$, cyan) and Fe_2O_3 ($R\bar{3}c$, red). Notably, Co@Fe_3O_4-A has fully transformed into α-Fe_2O_3 (red), while Co@Fe_3O_4-B exhibits a mixed-phase configuration containing both Fe_3O_4 (cyan) and α-Fe_2O_3 (red). f) Validation of the matching results; (left) experimental diffraction patterns and (right) simulated patterns for two randomly selected regions of each sample, confirming the crystal structure, orientation, and tilt geometry. The reciprocal scale bars are 1/5 nm.

transmission electron microscopy (BFTEM) imaging and scanning electron nanobeam diffraction (SEND) mapping were performed on all three samples. BFTEM imaging (Figure 1d) showed that the pristine Fe_3O_4 NPs displayed a consistent distribution, with an average size of 135 nm. In contrast, the Co@Fe_3O_4-A NPs were subjected to a substantial morphological alteration to larger particles (≈460 nm on average) with well-developed facets and the phase transition to α-Fe_2O_3.[15] Meanwhile, Co@Fe_3O_4-B NPs exhibited a morphological change primarily to a round shape, while their sizes slightly increased (≈200 nm on average). To understand the resulting phase distribution, SEND crystal-phase mapping with a finely converged probe (≈1.68 mrad) was performed (Figure 1e). The results revealed that Co@Fe_3O_4-A NPs was composed of more than 90% α-Fe_2O_3 (marked in red), while Co@Fe_3O_4-B NPs exhibited a mixed-phase composition, with a ratio of 60% Fe_3O_4 (cyan) to 40% α-Fe_2O_3 (red). The veracity of the automated mapping results was confirmed by randomly selected diffraction patterns from the mapped regions and verifying the agreement with the simulated patterns of Fe_3O_4 and α-Fe_2O_3 (Figure 1f). Because all NPs were arbitrarily oriented, the orientation and tilt geometry of the identified NPs were annotated by three-index Euler angles in the diffraction patterns to reproduce the simulation patterns.

Atomic-scale high-angle annular dark-field (HAADF) imaging confirmed that Co@Fe_3O_4-B NPs retained a combination of inverse spinel cubic (space group $Fd\bar{3}m$) and trigonal (space group $R\bar{3}c$) structures with random distributions (Figure S1, Supporting Information). Compared with the pristine structure, Co@Fe_3O_4-B NPs exhibited a slight lattice contraction when measured along the [220] direction. The contraction reached ≈30 pm for the extended-distance measurements of the 20 surface layers (Figure S2, Supporting Information). The lattice contraction observed in Co@Fe_3O_4-B NPs is a structural signature of the surface modification induced by the base treatment, aligning with the peak shifts in the XRD analysis (Figure 1b,c). These results suggest that the base treatment resulted in a mixture of Fe_3O_4 and α-Fe_2O_3 (Co@Fe_3O_4-B), but the magnetic Fe_3O_4 with surface contraction was kept as a dominant phase. In contrast, the acid treatment induced a nearly complete transition to α-Fe_2O_3 from Fe_3O_4 with a substantial size increase (Co@Fe_3O_4-A).

2.2. Electronic Structure Characterization

To assess whether doping concentration was maintained after pH treatment, X-ray fluorescence (XRF) analysis was performed. The XRF results showed that the cobalt content was 3.46 wt% (Co@Fe_3O_4-B) and 3.51 wt% (Co@Fe_3O_4-A), confirming that the actual doping levels were in close agreement with the nominal value. To assess the impact of Co doping on the electronic structure, multiscale spectroscopy was employed on the three samples with different morphological sizes (**Figure 2**a-c). First, we performed XPS to obtain comprehensive information into the changes of chemical states for the modified samples compared with that of the pristine sample. The Fe 2*p* XPS spectra

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

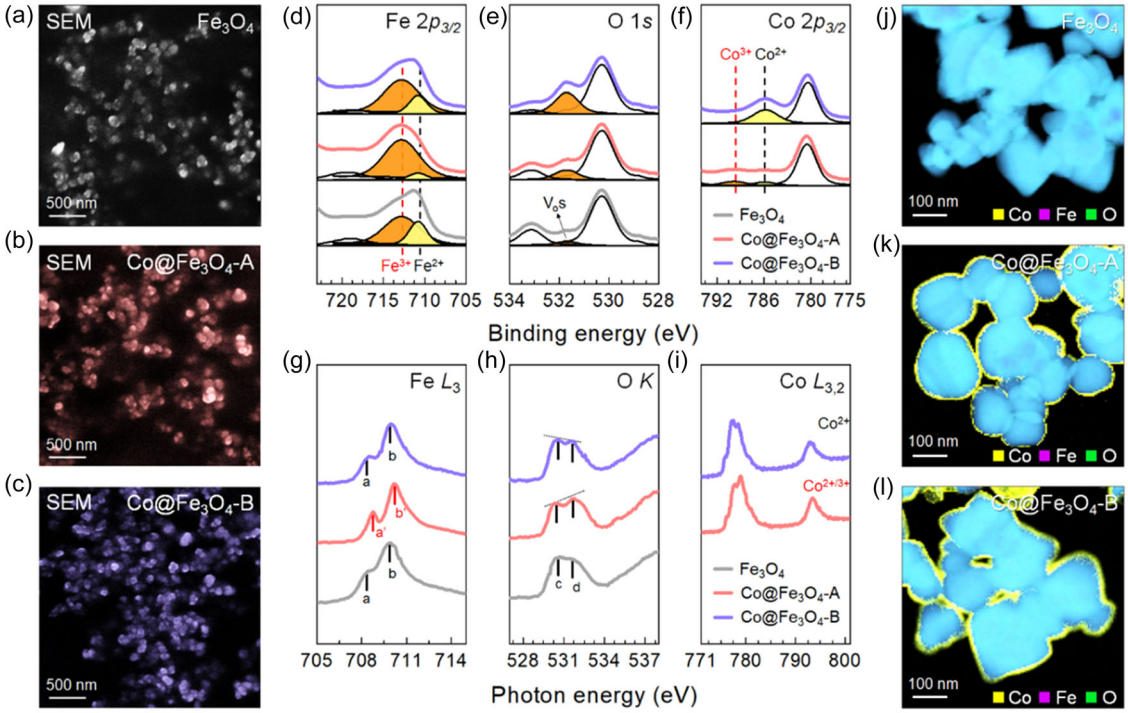


Figure 2. Electronic structure analysis. a–c) SEM images showing morphologies of the three samples. d–f) XPS spectra of Fe *2p*, O *1s* and Co *2p*. g–i) XAS spectra of Fe L_3, O *K*, and Co $L_{3,2}$ edges for the Fe_3O_4 (gray), $Co@Fe_3O_4$-A (red), and $Co@Fe_3O_4$-B (violet) NPs. j–l) Nanoscale EELS mapping of Fe *L* (purple), O *K* (green) and Co *L* (yellow) for the three samples.

(Figure 2d) revealed distinct differences in the oxidation states of Fe in the pH-treated samples. The deconvolution of the Fe 2*p* XPS results indicated that the relative ratio of Fe^{2+} to Fe^{3+} in the $Co@Fe_3O_4$-B NPs remained comparable to that of pristine Fe_3O_4. In contrast, the $Co@Fe_3O_4$-A NPs exhibited a decrease in Fe^{2+} and a concomitant increase in Fe^{3+}, which is consistent with the conversion of the α-Fe_2O_3 phase, as observed in Figure 1.

Next, deconvolution of the O 1*s* XPS spectra (Figure 2e) identified three distinct oxygen components: lattice oxygen (529.9 eV), oxygen vacancies (V_os, 531.7 eV), and adsorbed oxygen (532.7 eV).[16] Compared to the pristine Fe_3O_4, $Co@Fe_3O_4$-B NPs exhibited a higher percentage of V_os peaks. This suggests that the base treatment likely introduced defect structures, resulting in an increased abundance of V_os on the surface. The presence of V_os on the NP surface is presumed to act as a catalytic site for photocatalytic reaction. Analysis of the Co *2p* peak (Figure 2f) revealed a characteristic CoO peak[17] with uniform Co^{2+} in $Co@Fe_3O_4$-B NPs, potentially arising from the interaction between the CoO and Fe_3O_4 phases. $Co@Fe_3O_4$-A NPs displayed a heterogeneous mixture of Co^{3+} and Co^{2+} peaks, with Co^{3+} being predominant.

To further probe the electronic structural changes of the three NPs, soft XAS was conducted at the Fe *L*, O *K*, and Co *L* edges (Figure 2g-i). The Fe *L* edge profile of pristine Fe_3O_4 showed two primary features, marked as *a* (*a*′) and *b* (*b*′), where peak *a* (*a*′) at 707. 6 eV (708.1 eV; small shoulder at low energy) corresponds to the Fe^{2+} ions occupying octahedral sites, and peak *b* (*b*′) corresponds to the Fe^{3+} ions located in both tetrahedral (T_d) and octahedral (O_h) sites (Figure 2g).[15,18–20] In the $Co@Fe_3O_4$-A NPs, a shift of the Fe *L* edge peaks toward higher photon energies (*a*′ and *b*′) indicates an alteration in the electronic environment of the Fe ions compared to the peaks of pristine Fe_3O_4 and $Co@Fe_3O_4$-B NPs. In other word, that of $Co@Fe_3O_4$-B NPs was observed to be nearly unchanged while maintaining a relatively constant ratio of peaks (*a*/*b*) at the Fe *L* edge. This result indicates that the basic treatment did not significantly alter the oxidation state or occupancy of the Fe ions and preserved the inherent electronic structure of the Fe_3O_4 core, with minimal impact on the Fe ions within the crystal lattice.

The O *K* edge spectra (Figure 2h), divided into two regions, offer additional information about the electronic structure. The first region (529–534 eV) is associated with unoccupied O 2*p* states hybridized with Fe 3*d* states, and the second region (535–545 eV) corresponds to O 2*p* states hybridized with Fe 4*s* and 4*p* states. Notably, the intensity of the peak denoted as peak *d* in the first region, corresponding to t_{2g} and e_g states, varied among the pH-treated samples. For pristine Fe_3O_4 NPs, the intensity of peak *c* was greater than that of peak *d*.[19,21] The intensity of peak *c* for the $Co@Fe_3O_4$-B NPs was also higher than that of peak *d*, which is similar to the trend observed for pristine Fe_3O_4 NPs. Conversely, the intensity of peak *c* for $Co@Fe_3O_4$-A was lower than that of peak *d*, likely because of the electronic redistribution between O and Fe/Co. The Co *L*-edge spectra (Figure 2i) provided additional evidence of the different oxidation states of Co in the different samples. The Co oxidation state in $Co@Fe_3O_4$-B NPs exhibited predominantly Co^{2+} states, whereas $Co@Fe_3O_4$-A NPs indicated the coexistence of Co^{3+} and Co^{2+}.

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

This observation is noteworthy considering that the initial cobalt precursor, $Co(NO_3)_2{\cdot}xH_2O$, used for doping possessed a Co^{2+} oxidation state. Therefore, the pH treatment under basic conditions preserved the Co^{2+} state in $Co@Fe_3O_4$-B, whereas the pH treatment under acidic conditions induced different Co ion charge states in the NPs, leading to the oxidation of Co^{2+} to Co^{3+} in $Co@Fe_3O_4$-A. This is consistent with the XPS analysis (Figure 2d-f), where the pH treatment led to different oxidation states and occupancies of the Co dopants on the surface.

2.3. Surface Structure Characterization

EELS was employed to gain a deeper understanding of the local electronic and defect structures that may exist on the surface. Elemental mapping revealed that the Co doping was prominently confined to a surface layer with an approximate thickness of 10 nm (Figure 2j-l). The individual EELS results (Figure S3, Supporting Information) showed that the overlapping regions of Fe (purple) and O (green) resulted in a blue color due to additive color mixing in the elemental mapping (Figure 2j–l). This visual representation confirms the homogeneous distribution of Fe and O within the mapped NPs. High-resolution EELS spectrum imaging (SI) was performed on the treated samples to characterize the outskirt layers of individual NP. Subsequently, multiple linear least-squares (MLLS) fitting of the O *K* edge was used to identify the distribution of defects and associated changes in the electronic structure.

Figure 3a and b confirmed the formation of a chemically modified core–shell structure, with site-specific spectra acquired from the core and surface regions of individual NPs exhibiting subtle variations in the Co *L*, O *K*, and Fe *L* edges. Specifically, the Fe *L* spectra (Figure 3c) displayed features attributed to crystal-field splitting, resulting in two distinct peaks (t_{2g} and e_g). The energy splitting between these peaks at the L_3 edge (1.8 eV) indicated a uniform and high oxidation state with Fe^{3+} in both

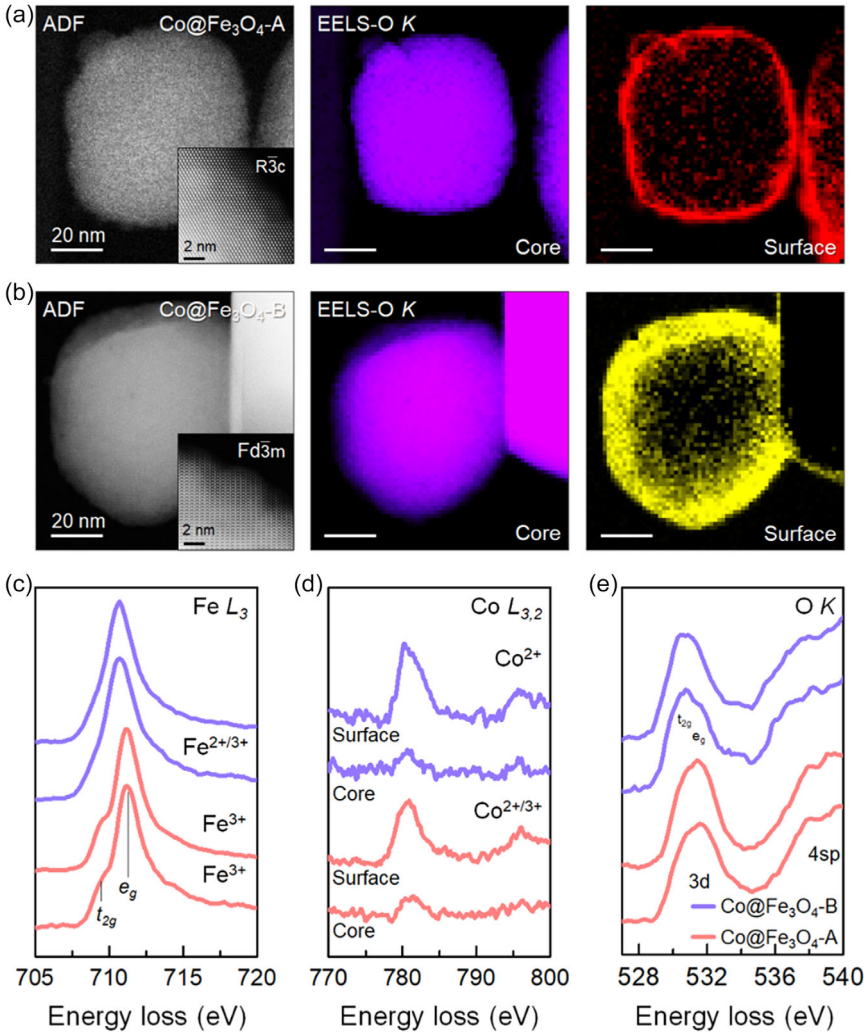


Figure 3. Electron energy loss spectroscopy (EELS) revealing the doping behavior of Co ions on the surface of the NPs. a,b) ADF images and MLLS fitting of O *K* for high-resolution EELS SI revealing the defect distribution and core–shell structure of $Co@Fe_3O_4$-A and $Co@Fe_3O_4$-B NPs. The inset ADF images of the atomic-resolution results show surface structures consistent with trigonal ($R\bar{3}c$) or inverse spinel cubic ($Fd\bar{3}m$) phase without dislocations or boundaries. c–e) Fe L_3, Co $L_{3,2}$, and O *K* edge spectra of core and surface regions of $Co@Fe_3O_4$-A (red) and $Co@Fe_3O_4$-B (violet) NPs.

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

regions for Co@Fe_3O_4-A NPs. In contrast, the t_{2g} and e_g fine structures were not resolved at the Fe L_3 edge of Co@Fe_3O_4-B NPs. Subtle variations in the Co L spectra (Figure 3d) indicate distinct doping behaviors between Co@Fe_3O_4-B NPs and Co@Fe_3O_4-A NPs. Co@Fe_3O_4-B NPs exhibited a higher L_3/L_2 intensity ratio and slight shifts in the L_3 and L_2 edges (≈0.8 and 0.5 eV, respectively) compared to Co@Fe_3O_4-A NPs. These changes reflect the differences in the electronic configuration and local environment of the Co ions between the samples.[22] This observation implies the possible substitution of Fe^{2+} by Co^{2+} within the Fe_3O_4 lattice. As a result, the surface-confined Co ions exhibited distinct charge states; the Co^{2+} state primarily existed in Co@Fe_3O_4-B NPs, whereas the Co^{2+} and Co^{3+} states coexisted in Co@Fe_3O_4-A NPs. Examination of the O K edge (Figure 3e) reveals disparities between the two samples, indicative of hybridization between the unoccupied 3d and 4sp states of cations and O 2p states. Although the dopants were localized at the surface in the Co@Fe_3O_4-A NPs, the spectral feature of the O K edge remained unchanged throughout the NP. By contrast, the spectra of Co@Fe_3O_4-B NPs exhibited an inverse slope at the O K edge compared to Co@Fe_3O_4-A NPs. The core O K spectrum of Co@Fe_3O_4-B NPs reflects two peaks associated with the t_{2g} and e_g states, followed by two broad peaks associated with Fe 4sp. Surface spectral changes include a reduction in the t_{2g} peak and a shift of the remaining 3d-related peaks toward higher energy, which is likely attributed to increased contributions from unoccupied states. This suggests the presence of Co^{2+} within the Fe_3O_4 lattice, and the possibility of surface-localized CoO_6 with Co^{2+} substituting for Fe^{2+}.[23] Consequently, Co@Fe_3O_4-B NPs possesses a more heterogeneous oxygen environment between the core and surface within the individual NPs, altering the electronic structure and bonding of oxygen atoms. Prior multiscale spectroscopic analyses highlighted the prevalence of V_os or local defects within the crystal structure. This heterogeneity underscores the potential presence of a stoichiometric core and a modified shell structure with Co substitutional doping and V_os concentrated on the NP surface, which is likely attributed to treatment under basic conditions (OH^- ions).

The modifications in the surface electronic structure were further corroborated by bandgap measurements using UV-vis diffuse reflectance spectroscopy (Figure S4a, Supporting Information). Pristine Fe_3O_4 exhibited a bandgap of 1.85 eV, with analogous values (≈1.7 eV) observed for the other samples. This substantiates the potential for photocatalytic activity in all three samples under visible-light irradiation. The Raman spectra of Co@Fe_3O_4-B NPs exhibited a red shift in a peak associated with the A_{1g} mode (≈2.6 cm^{-1}) for the Fe_3O_4 phase (Figure S4b, Supporting Information). Given that the A_{1g} mode in the Fe_3O_4 phase is ascribed to the symmetric stretching of the oxygen along the Fe—O bonds,[24] this indicates the weakening of the Fe—O bonding. The insets showing the atomic-resolution images in ADF (Figure 3a,b) confirm the typical structure of $R\bar{3}c$ and $Fd\bar{3}m$ symmetry without dislocations or boundaries. To further evaluate the surface characteristics of the samples, nitrogen adsorption–desorption measurements were conducted using the BET method (Figure S5, Supporting Information). The measured surface areas were 31.1 $m^2\ g^{-1}$ for Fe_3O_4, 24.3 $m^2\ g^{-1}$ for Co@Fe_3O_4-A and 41.1 $m^2\ g^{-1}$ for Co@ Fe_3O_4-B, indicating that base treatment enhanced the accessible surface area despite the increase in particle size.

We propose that the base treatment method with Co-ion doping is effective for surface modification, which agrees with the observed lattice contraction of Co@Fe_3O_4-B (Figure S2, Supporting Information) and peak shifts in the XRD analysis (Figure 1b,c). These findings suggest a correlation between strategic treatments and surface-engineered structures, which are anticipated to influence the photocatalytic and magnetic behaviors of Fe_3O_4 NPs. To evaluate the efficiency of charge separation for the photogenerated electron–hole pairs in the NPs, the photocurrent response was measured (Figure S6, Supporting Information). The results revealed a clear trend in photocurrent intensity, increasing in the order of Fe_3O_4, Co@Fe_3O_4-A, and Co@Fe_3O_4-B NPs. The enhancement for Co@Fe_3O_4-B (1130% compared to Fe_3O_4) is likely attributed to the presence of a high concentration of V_os on the surface. These V_os act as trapping sites for photogenerated electrons, facilitating charge separation and enhancing photocatalytic activity. To further support the presence of V_os, electron paramagnetic resonance (EPR) analysis was performed on three samples (Figure S7, Supporting Information). The Co@Fe_3O_4-B exhibited the most intense signal at a g-factor of 1.99, which is characteristic of unpaired electrons localized in oxygen-deficient sites. These findings corroborate that the base treatment promotes the formation of defect-rich surface states. Conversely, the negligible change for Co@Fe_3O_4-A NPs indicates a minimal contribution of surface defects to charge separation. Given the strong correlation between charge carrier dynamics and catalytic activity, the observed variations in photocurrent response are expected to be directly related to the photocatalytic reactivity of three samples.

2.4. Photocatalytic Performance

To assess the catalytic performance of the NPs, the photocatalytic conversion reaction was performed by exposing the liquid suspension containing three model molecules (furfuraldehyde, 5-hydroxymethylfurfural (HMF), and toluene) under visible-light irradiation ($\lambda \geq 445$ nm) with 230 mW cm^{-2} of illumination intensity. The photocatalytic degradation (PCD) profiles of three molecules highlight the differences in performance between the different NP constructs (**Figure 4**). Both pristine Fe_3O_4 and Co@Fe_3O_4-A NPs exhibited limited PCD efficiencies, decomposing only 69% and 79% of the furfuraldehyde, respectively (Figure 4a). This suggests that neither the pristine Fe_3O_4 structure nor the acidic treatment with Co-doping was sufficient to promote significant photocatalytic activity. In contrast, Co@Fe_3O_4-B NPs demonstrated an enhanced photocatalytic performance, achieving 97% decomposition of furfuraldehyde within 36 h (Figure 4a). Furthermore, Co@Fe_3O_4-B NPs showed 94% degradation of HMF and 78% of toluene, respectively (Figure 4b,c). These results represent substantial enhancements compared to pristine Fe_3O_4 NPs, with improvements of 41% for furfuraldehyde degradation, 45% for HMF degradation and 73% for toluene degradation. These enhancements are quantitatively summarized in **Table 1**, which presents the relative

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

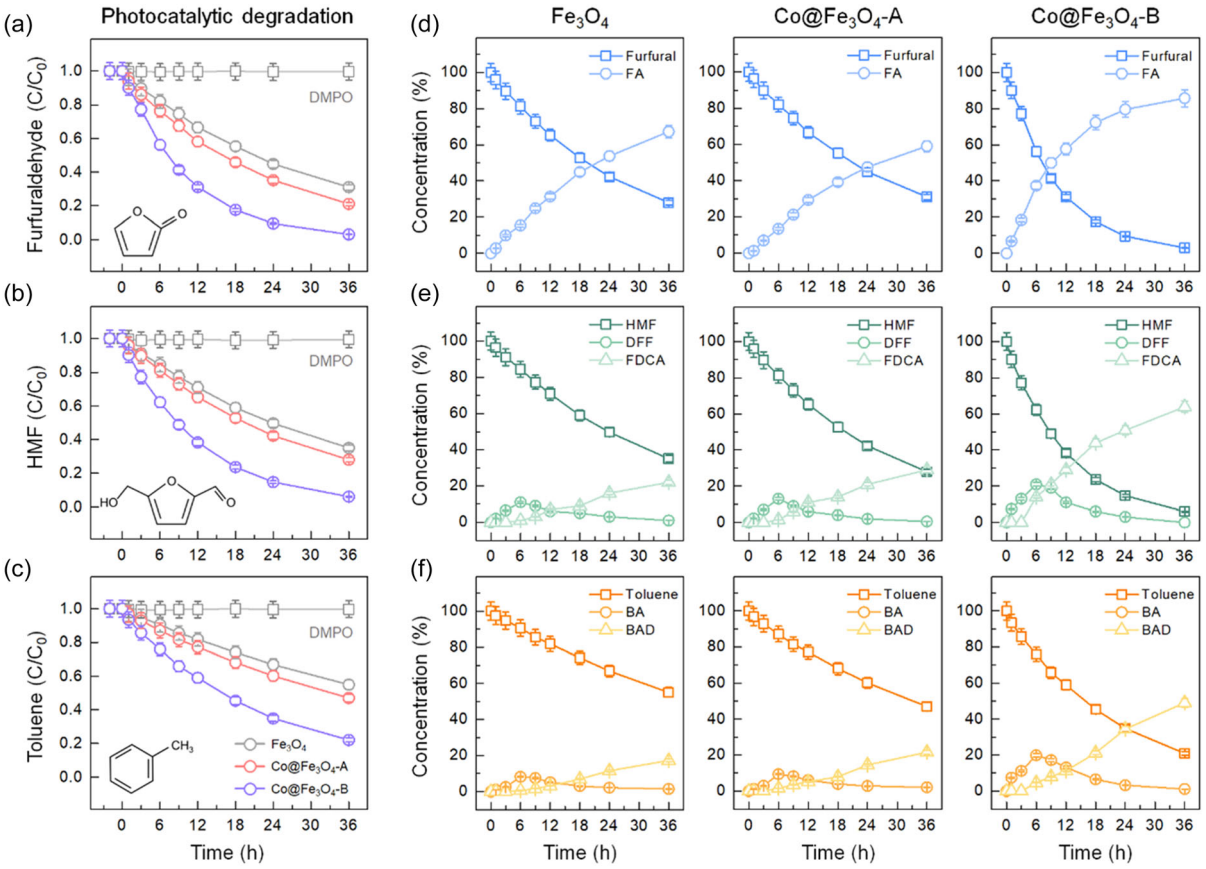


Figure 4. Photocatalytic conversion reaction of a) furfuraldehyde, b) HMF, and c) toluene under visible-light irradiation with pristine Fe_3O_4 (gray), Co@Fe_3O_4-A (red), and Co@Fe_3O_4-B (violet) NPs. DMPO was introduced as a radical scavenger. Experimental conditions: [catalysts] = 0.5 g L^{-1}, $[C]_0 = 25$ mM of furfuraldehyde, HMF and toluene, $\lambda \geq 445$ nm with 230 mW cm^{-2} of illumination intensity. d–f) Concentration profiles of derivates during photocatalytic reaction obtained from high-performance liquid chromatography (HPLC) with Fe_3O_4 (left), Co@Fe_3O_4-A (middle), and Co@Fe_3O_4-B (right panel) NPs. To ensure reproducibility, the experiments were performed three times under identical conditions.

Table 1. Photocatalytic conversion reaction of furfuraldehyde, HMF and toluene for 36 h under visible-light irradiation ($\lambda \geq 445$ nm) with 230 mW cm^{-2} of illumination intensity.

Samples	Furfuraldehyde (C/C_0)	HMF (C/C_0)	Toluene (C/C_0)
Fe_3O_4	0.31 ± 0.016	0.35 ± 0.018	0.55 ± 0.028
Co@Fe_3O_4-A	0.21 ± 0.011	0.28 ± 0.014	0.47 ± 0.024
Co@Fe_3O_4-B	0.03 ± 0.002	0.06 ± 0.003	0.22 ± 0.011

concentrations of the reactants (C/C_0), where C_0 is the initial concentration (25 mM) and C is the remaining concentration at a given time. A lower C/C_0 value corresponds to higher photocatalytic efficiency. This finding implies that the superior activity of the Co@Fe_3O_4-B NPs underscores the crucial role of base treatment in maximizing the concentration of V_os and optimizing the electronic properties for efficient photocatalysis. To gain deeper insights of the kinetics of photocatalytic reactions, three degradations were analyzed (Figure S8, Supporting Information). A linear relationship between $\ln(C/C_0)$ and the reaction time was observed for all three samples, indicating that degradation process followed pseudo-first-order kinetics. The constants of first-order rate (k) of Co@Fe_3O_4-B NPs were calculated to be 0.097 h^{-1} for furfuraldehyde (Figure S8a, Supporting Information), 0.078 h^{-1} for HMF (Figure S8b, Supporting Information) and 0.042 h^{-1} for toluene (Figure S8c, Supporting Information), which demonstrated a substantially elevated rate constant, reflecting ≈300% increase in conversion rate of furfuraldehyde, ≈270% increase in HMF and ≈250% increase in toluene compared with its pristine counterpart.

As present in Figure 4d-f, the results indicate that byproduct formation can be tracked using HPLC analysis. The yields of derivatives agree well with the observed PCR efficiencies. **Table 2** presents the corresponding yields of the oxidation derivatives, indicating the percentage concentration of specific derivatives formed after 36 h under visible-light irradiation. Based on these findings, the plausible mechanisms of the oxidative reaction for the three test molecules can be inferred as follows: the trapping of electrons and O_2 by oxygen vacancies formed at the NP surface, promoting the generation of superoxide radicals ($\bullet O_2^-$). These radicals, along with the photogenerated holes (h^+), drive the conversion of (a) furfuraldehyde to furoic acid (FA), (b) HMF to 2,5-furandicarboxylic acid (FDCA) and

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2. Concentration yield of derivates for photocatalytic conversion reaction for 36 h under visible-light irradiation ($\lambda \geq 445$ nm) with 230 mW cm^{-2} of illumination intensity.

Samples	Furoic acid (%)	FDCA (%)	Benzaldehyde (%)
Fe_3O_4	59.1 ± 3.0	22.0 ± 1.1	17.2 ± 1.0
Co@Fe_3O_4-A	67.3 ± 3.4	29.0 ± 1.5	21.6 ± 1.1
Co@Fe_3O_4-B	85.9 ± 4.8	64.0 ± 3.2	49.1 ± 2.5

(c) toluene to benzaldehyde (Scheme S1, Supporting Information). To confirm the role of $\bullet O_2^-$ in the photocatalytic system, 5,5-dimethyl-1-pyrroline N-oxide (DMPO) was introduced as a radical scavenger to inhibit the photocatalytic activity (Figure 4a-c).

The magnetic properties of the pristine Fe_3O_4 and Co@Fe_3O_4 NPs were evaluated by vibrating sample magnetometry (VSM) (Figure S9a, Supporting Information). The pristine Fe_3O_4 NPs exhibited a saturation magnetization of ≈80 emu g^{-1}. The shape of the magnetization curve of the Co@Fe_3O_4-A NPs exhibited no hysteresis and exhibited a weak magnetic moment, attributed to phase transition to paramagnetic α-Fe_2O_3 at 300 K.[25] Conversely, Co@Fe_3O_4-B NPs exhibited a saturation magnetization of 112 emu g^{-1}, surpassing that of pristine Fe_3O_4. This distinct behavior is crucial for recovering and reusing photocatalysts, a key factor for sustainable and cost-effective processes. Catalysts with high magnetic moments such as Co@Fe_3O_4-B NPs are highly advantageous because the magnetic nature of the NPs enables their facile and efficient separation from the reaction mixture through the use of an external magnet. This facile recovery not only simplifies the purification process but also minimizes material loss, contributing to the overall efficiency and economic viability of the photocatalytic system. To evaluate the possibility of recycling these catalysts, a recovery rate test was performed by retrieving the NPs using a magnet (Figure S9b, Supporting Information). Each PCR test, followed by an NP retrieval cycle, was repeated 10 times. As expected, the Co@Fe_3O_4-B NPs demonstrated the highest recovery rate after 10 cycles (Figure S6b, Supporting Information). Furthermore, the photocatalytic performance was sustained (>98%) even after 10 cycles, demonstrating their excellent stability and reusability (Figure S10, Supporting Information).

Given the impact of pH treatment after Co doping and on the photocatalytic and magnetic properties of the Fe_3O_4 NPs, we propose that a chemically modified core–shell structure was formed on Co@Fe_3O_4-B NPs. Specifically, $CoFe_2O_4$-like surface modifications occurred. This engineered surface enhanced the magnetic responsiveness of the NPs, facilitating their recovery and reuse. This not only streamlines the purification process but also fosters the sustainability and cost-effectiveness of the photocatalytic system. Moreover, this surface structure effectively minimizes the common challenges often encountered with conventional $CoFe_2O_4$ photocatalysts, such as ion leaching and agglomeration, while concurrently amplifying the photocatalytic performance. By confining the active species to the surface, this unique structure promotes stability and prevents the loss of catalytic activity over time. In summary, our findings suggest that base treatment after Co ion doping into Fe_3O_4 NPs plays a pivotal role in inducing surface modification with V_os, leading to enhanced photocatalytic activity, magnetic recoverability, and long-term stability. To visually summarize the overall strategy and mechanistic insights, a schematic illustration is presented in Scheme S2, Supporting Information. This approach holds great promise for the design of efficient photocatalysts for various applications including biomass conversion and environmental remediation.

3. Conclusion

This study elucidated the intricate mechanisms underlying the augmented photocatalytic activity of Fe_3O_4 following pH treatment after Co ion doping. Modulation of the pH coupled with the strategic incorporation of cobalt were used to engineer the crystallographic structure and electronic properties of Fe_3O_4. This manipulation engenders oxygen vacancies and induces distinct oxidation states, leading to enhanced photocatalytic activity under visible-light irradiation. Notably, the Co@Fe_3O_4-B NPs exhibited noticeable improvements in both the photocatalytic efficiency and magnetic properties, presenting a compelling solution for facile catalyst recovery and reusability. This research underscores the pivotal role of surface modification in optimizing Fe_3O_4 for applications pertaining to renewable energy and sustainable chemical processes and provides valuable insights into the design of high-performance photocatalysts.

4. Experimental Section

Preparation of Fe_3O_4 and Co@Fe_3O_4 NPs: Fe_3O_4 NPs were synthesized following a previously established method.[21,26] $FeCl_3$ (1.9 g) and $FeCl_2{\cdot}4H_2O$ (1.6 g) were dissolved in 200 mL of deionized water. The dissolution process was conducted under a nitrogen atmosphere with continuous and vigorous stirring at 25 °C to ensure homogeneity and prevent oxidation. Subsequently, 5 mL of polyethylene glycol (PEG 300) was added to the solution, serving as a capping agent to control the growth of the NPs. The mixture was then heated to 95 °C in a water bath and held at this temperature for 1 h with constant stirring. The fabricated solutions were then sealed with a Teflon lining and subjected to hydrothermal treatment at 220 °C for 9 h. Upon completion of the reaction, the autoclave was allowed to cool to 25 °C. The relevant chemical reaction is expressed in Equation (3)

$$Fe^{2+} + 2Fe^{3+} + 8OH^- \rightarrow Fe_3O_4 + 4H_2O \quad (3)$$

The resulting brown suspensions (Fe_3O_4) were washed multiple times with distilled water and dried at 95 °C for 48 h to remove any residual water. Co-doped Fe_3O_4 NPs were prepared by dispersing 3.0 g of the as-prepared Fe_3O_4 NPs and 0.54 g of $Co(NO_3)_2{\cdot}6H_2O$ (Co^{2+} ion state) in 100 mL of double-distilled water. This mixture was transferred to a Teflon-lined autoclave and heated at 220 °C for 9 h in a convection oven, resulting in a gel-like mixture. Next, Co@Fe_3O_4 was subjected to pH treatment by adjusting the pH level to 1.0 or 13.0 using HNO_3 or KOH, respectively. After ≈2 h, the synthesized gel mixture became transparent. The treated mixtures were subsequently transferred to Teflon-lined autoclaves and heated again at 220 °C for 9 h. The resulting Co@Fe_3O_4 NPs were filtered, washed with double-distilled water, and dried to remove residues.

Characterization of Crystal and Electronic Structures: The XRD patterns of the Fe_3O_4 and the two different pH-treated Co@Fe_3O_4 NPs were acquired using a D8 Advance TRIO/TWIN diffractometer (Bruker, USA) with

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Ni-filtered Cu Kα radiation. Ultraviolet-visible (UV-vis) diffuse reflectance spectra were acquired using a Shimadzu UV-2600i UV-vis spectrophotometer equipped with integrated spheres. A custom-designed Raman spectroscopy setup was employed and calibrated using the Si peak at 520.89 cm^{-1} as an internal reference. The light source used was a 532 nm laser with a 1.02 mW power. Photocurrent experiments (EIS experiments) were also conducted by fabricating a working electrode by drop-casting the photocatalyst dispersion onto fluorine-doped tin oxide (FTO) glass. The dispersion (25 μL) was then dropped onto the desired area (0.1 cm^2) of clean FTO glass and dried at room temperature. A light-emitting diodes (LEDs) (6 W) was used as the light source for photocurrent measurements at a wavelength of 445 nm with an illumination intensity of 230 mW cm^{-2} at the sample. The photocurrent densities of the samples were measured using two-electrode systems comprising working electrodes fabricated as described above and Pt gauze (0.8 × 0.8 cm) as the counter electrode using a multichannel potentiostat (WonATech, Korea). To compare the electronic structures of the three tested samples, XPS and work function analyses were performed at the 8A1 beamline in the Pohang Accelerator Laboratory (Korea). Fe 2*p*, O 1*s*, and Co 2*p* core-level spectra were obtained using photon energies of 800, 600, and 900 eV, respectively. The binding energies of the core-level electrons were compared with those of a clean Au 4*f* core at the same photon energy. X-ray absorption spectroscopy (XAS) measurements were also performed at the 8A1 beamline in the Pohang Accelerator Laboratory, and the Fe *L*-edge, O *K*-edge, and Co *L*-edge XAS profiles were obtained.

Microscopic Characterization: The morphologies of the pristine Fe_3O_4, Co@Fe_3O_4-A, and Co@Fe_3O_4-B NPs were investigated using aberration-corrected transmission electron microscopy (TEM) and scanning TEM (STEM). A spherical aberration-corrected microscope (ARM200CF, JEOL Ltd.) operated at an accelerating voltage of 200 kV was used in the bright-field TEM (BFTEM) and annular dark-field (ADF) modes to investigate the average particle size and crystal structure. Atomic-resolution ADF imaging was performed with a semi-convergence angle of 24 mrad and a detector angle range of 68–280 mrad. To map the crystal phases and symmetry, scanning electron nanobeam diffraction (SEND) was employed at an accelerating voltage of 300 kV under semi-convergence angle of 1.68 mrad (GrandARM300CF, JEOL Ltd) using a precession angle of 0.9° and precession frequency of 100 Hz (ASTAR, NanoMEGAS). Crystal mappings were acquired at 300 × 300 pixels for the pristine, 500 × 500 pixels for Co@Fe_3O_4-A, and 350 × 350 pixels for Co@Fe_3O_4-B with a resolution of 7.5 nm/pixel, and then cropped to a unified field-of-view of 2.3 × 1.6 μm. Phase analysis was performed using the template-matching method for diffraction indexing implemented in the software package of ASTAR. Core-loss EELS spectral imaging (SI) datasets for chemical mapping of the Fe *L*, O *K*, and Co *L* edges were acquired using a Gatan System (GIF Quantum ER 965, Gatan) in STEM mode. SI was performed at a resolution of 200 × 200 pixels, with a scanning step of 1 nm/pixel, scanning speed of 0.5 s, and energy dispersion of 0.25 eV/channel. Principal component analysis (PCA) was applied using commercial software (MSA, HREM Research) to filter statistical noise in the EELS SI datasets. High-precision spectral analysis of the Fe *L*, O *K*, and Co *L* edges was conducted using a Gatan Continuum HR-1066 instrument on a monochromated STEM instrument (JEOL MonoARM200, DJ105@KBSI, JEOL Ltd.) at the Korea Basic Science Institute (KBSI). The acquisition was performed with energy dispersion of 0.05 eV/channel, scanning step of 0.5 nm/pixel and scanning speed of 0.5 s. The energy loss in the core-loss EELS dataset was calibrated by tracking the energy drift of the zero-loss spectra recorded concurrently with that of the core-loss spectra.

Photocatalytic and Magnetic Property Measurements: The photocatalytic conversion reaction (PCR) of the three samples was evaluated under visible-light irradiation. Suspensions of the NPs (0.5 g L^{-1}) were prepared in an aqueous solution containing the target substrate (25 mM furfuraldehyde, 5-hydroxymethylfurfural (HMF), or toluene) and stirred in the dark for 2 h to allow the system to reach adsorption–desorption equilibrium. The reactor containing the NP solution was placed 10 cm in front of LEDs (λ = 445 nm, output power = 6 W, illumination intensity = 230 mW cm^{-2}; Thorlab) under magnetic stirring at ≈110 rpm. The concentration of furfuraldehyde, HMF and toluene was measured using a UFLC LC-20AD (Shimadzu, Japan) high-performance liquid chromatography system. The magnetic properties (magnetic moment and coercivity) of the Fe_3O_4 and Co@Fe_3O_4 NPs were measured using a vibrating sample magnetometry (VSM) instrument (MicroSense, USA) at 300 K.

Statistical Analysis: Spectroscopic data for XAS, XPS, and EELS were normalized to a [0,1] range to facilitate direct comparison across samples. Quantitative data are reported as the mean ± standard deviation (SD). To ensure reproducibility, all photocatalytic property measurements of conversion rates, product yields, photocurrent intensity, and catalyst recovery were performed in triplicate (n = 3) under identical conditions.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Y.-H.K., M.S., and E.-B.P. contributed equally to this work. This work was supported by National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (MSIT) of the Korean government (RS-2024-00346153 for H. Lee and 2022R1C1C1009116 for M. Shin). Y.-M.K. acknowledges support from the Technology Innovation Program (RS-2024-00433399) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and the Institute for Basic Science (IBS-R036-D1). This work was supported by the Sungkyunkwan President Fellowship Program (2024).

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Young-Hoon Kim: conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); writing—original draft (lead); **Minjeong Shin**: data curation (equal); formal analysis (equal); investigation (equal); visualization (equal); **Eun-Byeol Park**: investigation (equal); **Sunyoung Hwang**: investigation (equal); **Jung A. Hong**: investigation (equal); **Jae Hyuck Jang**: investigation (equal); **Young-Min Kim**: conceptualization (equal); supervision (lead); writing—review & editing (lead); **Hangil Lee**: conceptualization (equal); supervision (lead); writing—review & editing (lead).

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

Fe_3O_4 nanoparticles, oxygen vacancies, photocatalytic properties, surface structural modification

Received: February 26, 2025
Revised: April 15, 2025
Published online: May 11, 2025

[1] a) K. Nakata, A. Fujishima, *J. Photochem. Photobiol. C-Photochem. Rev.* **2012**, *13*, 169; b) S. Liang, G. Sui, J. Li, D. Guo, Z. Luo, R. Xu, H. Yao,

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

C. Wang, S. Chen, *Int. J. Hydrog. Energy* **2022**, *47*, 11190; c) M. Schmitt, D. Becker, J. Lalevée, *Polymer* **2018**, *158*, 83; d) Z. Y. Wang, S. Y. Liu, J. A. Zhang, J. J. Yan, Y. P. Zhao, C. Mahoney, R. Ferebee, D. L. Luo, J. Pietrasik, M. R. Bockstaller, K. Matyjaszewski, K. Matyjaszewski, *Langmuir* **2017**, *33*, 12276.

[2] a) C. L. Pang, R. Lindsay, G. Thornton, *Chem. Soc. Rev.* **2008**, *37*, 2328; b) Y. Ma, X. L. Wang, Y. S. Jia, X. B. Chen, H. X. Han, C. Li, *Chem. Rev.* **2014**, *114*, 9987; c) J. Theerthagiri, S. Salla, R. A. Senthil, P. Nithyadharseni, A. Madankumar, P. Arunachalam, T. Maiyalagan, H. S. Kim, *Nanotechnology* **2019**, *30*, 392001; d) R. Katal, S. Masudy-Panah, M. Tanhaei, M. H. D. A. Farahani, J. Y. Hu, *Chem. Eng. J.* **2020**, *384*, 123384.

[3] a) W. Zhang, J. K. Yang, X. Wu, Y. C. Hu, W. H. Yu, J. X. Wang, J. X. Dong, M. Y. Li, S. Liang, J. P. Hu, R. V. Kumar, *Renew. Sust. Energ. Rev.* **2016**, *61*, 108; b) L. G. Zhang, Z. M. Xu, *Waste Manage.* **2019**, *96*, 175; c) E. Valadez-Renteria, J. Oliva, V. Rodriguez-Gonzalez, *Sci. Total Environ.* **2022**, *807*, 150820.

[4] A. G. Leonel, A. A. P. Mansur, H. S. Mansur, X. Liu, *Water Res.* **2021**, *190*, 116693.

[5] J. Kiwi, S. Rtimi, *Appl. Catal., B* **2021**, *281*, 119420.

[6] Y. Wang, X. Gu, J. Quan, G. Xing, L. Yang, C. Zhao, P. Wu, F. Zhao, B. Hu, Y. Hu, *Sci. Total Environ.* **2021**, *773*, 145476.

[7] a) X. Liu, J. Liu, S. H. Zhang, Z. D. Nan, Q. Shi, *J. Phys. Chem. C* **2016**, *120*, 1328; b) F. Chang, H. Y. Chen, X. Y. Zhang, B. Lei, X. F. Hu, *Sep. Purif. Technol.* **2020**, *238*, 116442; c) W. Wang, Q. Mao, H. H. He, M. H. Zhou, *Water Sci. Technol.* **2013**, *68*, 2367; d) J. Yu, F. Zhao, W. L. Gao, X. Yang, Y. M. Ju, L. Y. Zhao, W. S. Guo, J. Xie, X. J. Liang, X. Y. Tao, J. Li, Y. Ying, W. C. Li, J. W. Zheng, L. Qiao, S. B. Xiong, X. Z. Mou, S. L. Che, Y. L. Hou, *ACS Nano* **2019**, *13*, 10002; e) C. Q. Tan, X. Lu, X. Cui, X. C. Jian, Z. X. Hu, Y. J. Dong, X. Y. Liu, J. Huang, L. Deng, *Chem. Eng. J.* **2019**, *363*, 318.

[8] a) Y. J. Deng, X. L. Tian, G. H. Shen, Y. Gao, C. X. Lin, L. M. Ling, F. L. Cheng, S. J. Liao, S. G. Zhang, *J. Colloid Interface Sci.* **2020**, *567*, 410; b) J. Mock, B. Klingebiel, F. Köhler, M. Nuys, J. Flohre, S. Muthmann, T. Kirchartz, R. Carius, *Phys. Rev. Mater.* **2017**, *1*, 065407; c) W. Xiao, Y. Yang, X. Chi, B. H. Liu, Y. H. Du, P. Yang, T. S. Herng, D. Q. Gao, W. D. Song, Y. P. Feng, A. Rusydi, J. Ding, *ACS Appl. Mater. Interfaces* **2019**, *11*, 1049; d) D. L. Ni, J. W. Zhang, J. Wang, P. Hu, Y. Y. Jin, Z. M. Tang, Z. W. Yao, W. B. Bu, J. L. Shi, *ACS Nano* **2017**, *11*, 4256; e) W. S. Jang, Y. Jin, Y.-H. Kim, S.-H. Yang, S. J. Kim, J. A. Hong, J. Baik, J. Lee, H. Lee, Y.-M. Kim, *Appl. Catal., B* **2022**, *305*, 121083; f) Y.-H. Kim, S. Kim, K.-j. Kim, C. Kim, J. H. Jang, Y.-M. Kim, H. Lee, *J. Mater. Chem. A* **2020**, *8*, 25345.

[9] a) A. Lolli, V. Maslova, D. Bonincontro, F. Basile, S. Ortelli, S. Albonetti, *Molecules* **2018**, *23*, 2792; b) A. Ulyankina, S. Mitchenko, N. Smirnova, *Processes* **2020**, *8*, 647; c) R. Jaiswal, K. V. S. Ranganath, *J. Inorg. Organomet. Polym. Mater.* **2021**, *31*, 4504.

[10] a) C. Li, Y. Na, *ChemPhotoChem* **2021**, *5*, 502; b) M. E. Davis, *Top. Catal.* **2015**, *58*, 405; c) M. G. Davidson, S. Elgie, S. Parsons, T. J. Young, *Green Chem.* **2021**, *23*, 3154.

[11] a) Z. L. Liu, Y. J. Liu, K. L. Yao, Z. H. Ding, J. Tao, X. Wang, *J. Mater. Synth. Process* **2002**, *10*, 83; b) Y. H. Zheng, Y. Cheng, F. Bao, Y. S. Wang, *Mater. Res. Bull.* **2006**, *41*, 525.

[12] a) O. M. Lemine, *Superlattices Microstruct.* **2009**, *45*, 576; b) F. Wang, X. F. Qin, Y. F. Meng, Z. L. Guo, L. X. Yang, Y. F. Ming, *Mater. Sci. Semicond. Process* **2013**, *16*, 802.

[13] M. Pourbaix, *Atlas Of The Electrochemical Equilibria*, Pergamon Press, Oxford, UK **1966**.

[14] a) K. Sobańska, P. Pietrzyk, Z. Sojka, *ACS Catal.* **2017**, *7*, 2935; b) G. Han, J. Y. Kim, K.-J. Kim, H. Lee, Y.-M. Kim, *Appl. Surf. Sci.* **2020**, *507*, 144916.

[15] A. R. B. de Castro, P. T. Fonseca, J. G. Pacheco, J. C. V. da Silva, E. G. L. da Silva, M. H. A. Santana, *J. Magn. Magn. Mater.* **2001**, *233*, 69.

[16] X. Hao, R. Xing, *Mater. Lett.* **2024**, *377*, 137387.

[17] M. C. Biesinger, B. P. Payne, A. P. Grosvenor, L. W. M. Lau, A. R. Gerson, R. S. C. Smart, *Appl. Surf. Sci.* **2011**, *257*, 2717.

[18] B. Gilbert, J. E. Katz, J. D. Denlinger, Y. D. Yin, R. Falcone, G. A. Waychunas, *J. Phys. Chem. C* **2010**, *114*, 21994.

[19] D. Kwon, H. W. Nho, T. H. Yoon, *Colloid Surf. B-Biointerfaces* **2014**, *122*, 384.

[20] S. H. Sun, H. Zeng, D. B. Robinson, S. Raoux, P. M. Rice, S. X. Wang, G. X. Li, *J. Am. Chem. Soc.* **2004**, *126*, 273.

[21] H. T. Bui, S. M. Im, K. J. Kim, W. Kim, H. Lee, *Appl. Surf. Sci.* **2020**, *509*, 144853.

[22] a) S. Samira, J. Hong, J. C. A. Camayang, K. Sun, A. S. Hoffman, S. R. Bare, E. Nikolla, *JACS Au* **2021**, *1*, 2224; b) Z. Zhang, *Ultramicroscopy* **2007**, *107*, 598.

[23] M. Taskin, Z. Novotny, N. Comini, J. T. Diulus, M. Hengsberger, P. Krüger, J. Osterwalder, *Phys. Rev. B* **2023**, *108*, 155403.

[24] a) J. Zhang, P. Tan, W. Zhao, J. Lu, J. Zhao, *J. Raman Spectrosc.* **2011**, *42*, 1388; b) O. N. Shebanova, P. Lazor, *J. Solid State Chem.* **2003**, *174*, 424.

[25] a) H. Jiao, G. S. Jiao, *Mater. Lett.* **2009**, *63*, 2725; b) M. M. Rafi, K. S. Z. Ahmed, K. P. Nazeer, D. S. Kumar, M. Thamilselvan, *Appl. Nanosci.* **2015**, *5*, 515; c) D. Shi, H. Yang, S. F. Ji, S. Jiang, X. F. Liu, D. N. Zhang, *Procedia Engineer* **2015**, *102*, 1555; d) D. Kim, S. S. Min, Y. H. Choa, H. T. Kim, *J. Ind. Eng. Chem.* **2007**, *13*, 1137.

[26] A. H. M. Yusoff, M. N. Salimi, M. F. Jamlos', *AIP Conf. Proc.* **2017**, *1835*, 020010.

© 2025 The Author(s). Small Structures published by Wiley-VCH GmbH

26884062, 2025, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sstr.202500123 by Sookmyung Womens University, Wiley Online Library on [09/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Inorganic Chemistry Communications 177 (2025) 114287

Contents lists available at ScienceDirect

Inorganic Chemistry Communications

journal homepage: www.elsevier.com/locate/inoche

Impact of oxygen vacancy density in modified cobalt oxide nanoparticles on the efficiency of visible-light photocatalysis

Hyeri Jeon [a,b,1], Sunyoung Hwang [c,1], Taeyeon Kang [d,1], Seungwoo Hong [a,b,*], Hyun S. Ahn [d,*], Hangil Lee [c,*]

[a] Department of Chemistry & Nanoscience, Ewha Womans University, Seoul 03760, Republic of Korea
[b] Graduate Program in Innovative Biomaterials Convergence, Ewha Womans University, Seoul 03760, Korea
[c] Department of Chemistry, Sookmyung Women's University, Seoul 04310, Republic of Korea
[d] Department of Chemistry, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

ARTICLE INFO

Keywords:
Surface oxygen vacancy
Base treated CoO
Li-deposited CoO NPs
Photocatalytic activity
Photocatalyst

ABSTRACT

This study demonstrates enhanced photocatalysis on cobalt oxide nanoparticles (CoO NPs) by increasing the surface density of oxygen vacancies (V_os), achieved through base treatment and subsequent Li ion deposition. Surface Li-ion-doped CoO NPs exhibited significantly improved biomass oxidation capabilities. Li doping effectively scavenged oxygen from CoO NPs, resulting in a substantial increase in V_os and modifications to the surface structure. The induced surface defects, combined with energy engineering, enhanced the photocatalytic oxidation of biomolecules, such as 2,5-hydroxymethylfurfural. This study serves as a proof of concept, highlighting the potential of purpose-driven and mechanism-specific surface modification and defect engineering techniques, while emphasizing the need for their development and appropriate application.

1. Introduction

Photocatalysis research primarily focuses on materials engineering aimed at achieving high efficiency and cost-effectiveness. However, solar energy utilization by conventional photocatalysts, such as metal oxide nanoparticles (MO NPs), remains low, at approximately 4 %. This limitation is due to the wide band gap of these materials, typically around 3.2 eV (in the UV range), necessitating advancements to enhance photon absorption and utilization in the visible spectrum [1,2]. Recently, MO NPs have gained attention as cost-effective catalysts for photocatalytic biomass conversion reactions [3,4]. Nevertheless, their photocatalytic performance remains suboptimal, highlighting the need for developing synthetic and surface engineering methods to improve their visible-light active photocatalytic capabilities.

In recent years, the photocatalytic transformation of 2,5-hydroxymethylfurfural (HMF) into 2,5-furandicarboxylic acid (FDCA) has attracted significant scientific interest due to the immense potential of biomass-derived materials in addressing pressing environmental challenges [5,6]. Based on the well-established mechanism of photocatalytic nanoparticle (NP) reactions, photon absorption generates high-energy excited states, resulting in the formation of electron-hole pairs. Electrons are promoted to the conduction band (e^-_{CB}), while holes remain in the valence band (h^+_{VB}) [7]. The h^+_{VB} facilitates various photoreactions by (i) selectively oxidizing molecules such as HMF or (ii) generating highly reactive hydroxyl radicals ($OH^{\blacksquare}$) in aqueous solution.

Most photocatalytic MO NPs are n-type semiconductors with desirable physicochemical properties, including low cost and chemical inertness [8]. However, their bandgap is typically above 3.0 eV, limiting their photo-response and absorption efficiency in the visible light spectrum. This limitation underscores the need for strategies to enhance their intrinsic chemical properties to improve photocatalytic activity under visible light. A key focus in developing visible light-active photocatalysts is to either tune the bandgap of semiconducting MO NPs through doping or use NPs with an intrinsic bandgap capable of absorbing visible light. Additionally, creating a substantial number of oxygen vacancies (V_os) on the surface can generate extra charge carriers and enhance the oxidation of molecules interacting with the surface [9].

Cobalt oxide nanoparticles (CoO NPs) is an important catalyst candidate material for a number of energy conversion reactions. Most research on CoO NPs has focused on enhancing the efficiency of

* Corresponding authors at: Department of Chemistry & Nanoscience, Ewha Womans University, Seoul 03760, Korea (Seungwoo Hong).
E-mail addresses: hsw@ewha.ac.kr (S. Hong), ahnhs@yonsei.ac.kr (H.S. Ahn), easyscan@sookmyung.ac.kr (H. Lee).
[1] H.J., S.H. and T.K. contributed equally to this paper.

https://doi.org/10.1016/j.inoche.2025.114287
Received 6 January 2025; Received in revised form 19 February 2025; Accepted 5 March 2025
Available online 14 March 2025
1387-7003/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

electrocatalysts used in secondary batteries [10]. However, with a band gap of 1.4 eV, the photocatalytic properties of CoO NPs can be further improved through various surface modifications [11,12]. For example, creating V_os on the surface of MO NPs via base treatment or metal ion deposition is a straightforward post-synthesis modification strategy that enhances the catalytic efficiency of the parent material [13,14]. V_os act as mobile electron donors by altering the local valence states of neighboring cations at the atomic scale, thereby improving the photocatalytic activity of MO NPs [15].

This study focuses on developing an effective method to enhance the photocatalytic capabilities of CoO NPs by modulating the number of V_os through base treatment (CoO-B; base treated CoO NPs) and the introduction of Li-ion dopants (Li@CoO-B; Li doping after base treated CoO NPs). High-resolution transmission electron microscopy (HRTEM), X-ray absorption spectroscopy (XAS), and X-ray photoelectron spectroscopy (XPS) analyses reveal a discernable change in the density of V_os on the surface of both CoO-B and Li@CoO-B NPs as well as their surface structures. Compared to pristine CoO NPs, Li@CoO-B NPs demonstrate superior performance in the photocatalytic conversion of 2,5-hydroxymethylfurfural (HMF) to 2,5-furandicarboxylic acid (FDCA), highlighting the effectiveness of these modification methods in significantly improving photocatalytic properties under visible light.

2. Experimental

2.1. Sample preparation

The synthetic procedures for CoO NPs are summarized as follows: First, 0.2571 g of cobalt(II) acetylacetonate ($Co(acac)_2$: 1 mmol) was dissolved in 20 mL of octadecene (ODE) along with oleic acid (OA) and oleylamine (OAm) in a molar ratio of 4:5:10 to form a colloidal mixture. The mixture was vigorously stirred for 30 min at room temperature [16]. It was then sealed in a Teflon-lined autoclave and maintained at 220 °C for 24 h. After cooling to room temperature, the dark brown product was washed several times with absolute ethanol. To fabricate CoO-B NPs, 1.0 g of the pristine CoO NPs was dispersed in 100 mL of water and titrated with potassium hydroxide (KOH) to a target pH of 13.0. The titrated solution was stirred at 110 rpm for 3 h at 90 °C to achieve homogeneous mixing. This mixture was then sealed in a Teflon-lined autoclave and maintained at 220 °C for 9 h before cooling to room temperature. For the synthesis of 3 wt% Li-doped CoO-B NPs (Li@CoO-B NPs), 0.2983 g of Li-ion dopant, in the form of lithium nitrate ($LiNO_3$, 99 %), was added to CoO-B NPs dispersed in 100 mL of water. The resulting homogeneous and transparent solution was transferred to a Teflon-lined autoclave, sealed, and heated at 220 °C for 9 h in a convection oven. The CoO, CoO-B, and Li@CoO-B NPs were filtered and washed five times with double-distilled water (DDW). Finally, the washed samples were held at 95 °C for 48 h to remove any remaining water.

2.2. Characterizations

Atomic-scale structural images of pristine CoO, CoO-B, and Li@CoO-B NPs were acquired using high-resolution transmission electron microscopy (HRTEM; JEM-ARM200CF, JEOL Ltd., Japan) operated at an accelerating voltage of 200 kV. XRD patterns of the three NPs were obtained using a D8 Advance TRIO/TWIN diffractometer (Bruker, USA) with Ni-filtered Cu Kα radiation. Photocurrent and electrochemical impedance spectroscopy (EIS) experiments were conducted by preparing a working electrode through drop-casting a photocatalyst dispersion onto fluorine-doped tin oxide (FTO) glass. Specifically, 25 μL of the dispersion was deposited onto a 0.1 cm^2 area of clean FTO glass and allowed to dry at room temperature. A blue light-emitting diode (LED) was employed as the light source for photocatalytic reactions at a wavelength of 445 nm. (λ = 445 nm, output power = 6 W, Thorlabs, USA). Photocurrent densities of the samples were measured using a two-electrode system. The system consisted of the fabricated working electrodes and a platinum (Pt) gauze (0.8 × 0.8 cm) counter electrode, with measurements performed using a multichannel potentiostat (WonATech, Korea). Ultraviolet–visible diffuse reflectance (UV–Vis DRS) measurements were performed using a Shimadzu UV-2600i UV–Vis spectrophotometer equipped with integrating spheres. The sample morphologies were characterized using a field-emission scanning electron microscope (FE-SEM, FEI Inspect F50) operating at an acceleration voltage of 15 kV. Nitrogen adsorption isotherms were obtained at 77 K using a BELsorp-Max system (BEL) to measure the surface areas of the three CoO NPs. To compare the electronic structures of the three CoO NPs (CoO, CoO-B, and Li@CoO-B), Co 2*p* and O 1*s* core-level spectra were acquired using X-ray photoelectron spectroscopy (XPS), and Co *L*-edge and O *K*-edge spectra were obtained using X-ray absorption spectroscopy (XAS) at beamline 8A1 of the Pohang Accelerator Laboratory (PAL, Republic of Korea). Co *K*-edge data were analyzed using extended X-ray absorption fine structure (EXAFS) at the 8C beamline of PAL. The magnetic properties, including magnetic moment and coercivity, were measured at 300 K using a vibrating-sample magnetometer (VSM) instrument (MicroSense, USA). To detect the presence of V_os on the surfaces of the three CoO NPs, electron paramagnetic resonance (EPR) spectra were recorded using a JES-X320 instrument (JEOL Ltd.) under the following conditions: 345–350 mT range, 9446 MHz frequency, 10 mW power, 100 kHz modulation frequency, and 0.4 mT modulation width at 300 K.

2.3. Extended X-ray absorption fine structure

Co *K*-edge XAFS data were collected at beamline 8C of the PAL. Energy calibration was performed by measuring the first inflection point of the cobalt reference foil (7709.1 eV). EXAFS data were analyzed using *Athena* software [17]. The processed data were converted to photoelectron wave vector (*k*) space and weighted by k^3. The electron wavenumber is defined as in Equation (1) as follows:

$$k = [\{2m(E - E_0)\}/\hbar^2]^{1/2} \quad (1)$$

where E_0 denotes the threshold energy. *k*-space data were truncated near zero crossings and Fourier transformed into *R*-space. *Artemis* software was used for curve fitting. Curve fitting was performed using the ab initio-calculated phases and amplitudes from the FEFF8 program at the University of Washington. The *ab initio* calculated phases and amplitudes were used in the EXAFS equation (Eq. (2)). [18]

$$\chi(\mathrm{k}) = \mathrm{S_0^2}\sum \frac{\mathrm{N}}{\mathrm{kR_j^2}} \mathrm{f_{eff,j}}(\pi, \mathrm{k}, \mathrm{R_j}) \mathrm{e}^{-2\sigma_j^2 \mathrm{k}^2} \mathrm{e}^{-\frac{2\mathrm{R_j}}{\lambda_j(\mathrm{k})}} \sin(2\mathrm{kR_j} + \varphi_{\mathrm{ij}}(\mathrm{k})) \quad (2)$$

where N_j is the number of atoms in the *j*th shell, R_j is the distance between the absorbing atoms and the atoms in the *j*th shell; $f_{\mathrm{eff},j}$ is the *ab initio* amplitude function for *j*, and $e^{-2\sigma_j^2 k^2}$ is the Debye − Waller factor for shell *j* accounting for damping due to thermal and static disorder in the shell. The mean free path term ($e^{-2Rj/\lambda j(k)}$) accounts for losses due to inelastic scattering. The oscillations in the EXAFS spectrum are reflected in the $\sin(2kR_j + \varphi_{ij}(k))$ term, where $\varphi_{ij}(k)$ is the *ab initio* phase function for the shell *j*. S_0 is an amplitude reduction factor. The EXAFS equation was used to fit experimental data using *N*, E_0, *R*, and σ^2 as variable parameters, while S_0 was kept fixed.

2.4. Photocatalytic degradation activities and selective oxidation of HMF

Photocatalytic conversion activity (PCA) was evaluated by stirring suspensions of the three CoO NPs (0.5 g/L) in DDW containing HMF (25 mM) in the dark for 3 h to establish adsorption equilibrium. The reactors containing each solution were magnetically stirred at approximately 110 rpm in front of a blue light-emitting diode (LED) (λ = 445 nm, output power = 6 W, Thorlabs, USA). The oxidation products of HMF were analyzed using HPLC-MS (ULTIMATE 3000 RSLC SYSTEM

(Thermo) and Q-EXACTIVE ORBITRAP PLUS MS (Thermo)). Changes in HMF concentrations were measured by high-performance liquid chromatography (HPLC, Shimadzu UFLC LC-20AD pump).

3. Results and discussion

To investigate the surface morphological changes of CoO NPs due to the base treatment and an additional Li-ion doping, HRTEM analyses were conducted (Fig. 1a–c) [19,20]. The images and fast Fourier transform (FFT) results clearly show that CoO and the two modified CoO samples (i.e., CoO-B and Li@CoO-B) exhibit distinct interlayer spacing. The FFT patterns for all CoO-based NPs correspond to the (100), (010), and (−110) crystal planes of the *fcc* CoO phase. As shown in Fig. 1a–c, all three CoO NPs display a highly crystalline *fcc* CoO phase, with each NP exhibiting a single-crystal orientation.

To study the photocatalytic properties of the two modified CoO NPs (CoO-B and Li@CoO-B), their structural changes were analyzed. The change in the interatomic distance corresponding to the sample surface direction was calculated using FFT, as shown in Fig. 1d–f. The changes in distance along three crystal directions (i.e., [110], [010], and [−110]) for the (001) plane of CoO NPs were measured using the FFT method, as shown in Fig. 1d. The changes in the distance along the [110], [010], and [−110] crystal directions were 2.84, 2.84, and 2.84 Å, respectively, confirming the consistency of the results with literature values.

The purpose of this study was to vary the photocatalytic properties of the modified CoO NPs by the base treatment and additional Li ion doping. Therefore, the distances in three crystal directions (i.e [110], [010], and [−110]) for the (100) plane of crystalline CoO-B and

Table 1
Distances in three crystal directions in the (100) plane of CoO, CoO-B, and Li@CoO-B NPs.

Crystal directions in (100) plane of CoO	CoO	CoO-B	Li@CoO-B
[100]	2.84 Å	2.88 Å	4.67 Å
[010]	2.84 Å	2.34 Å	4.69 Å
[-110]	2.84 Å	2.45 Å	4.09 Å

*CoO denotes the reference distance.

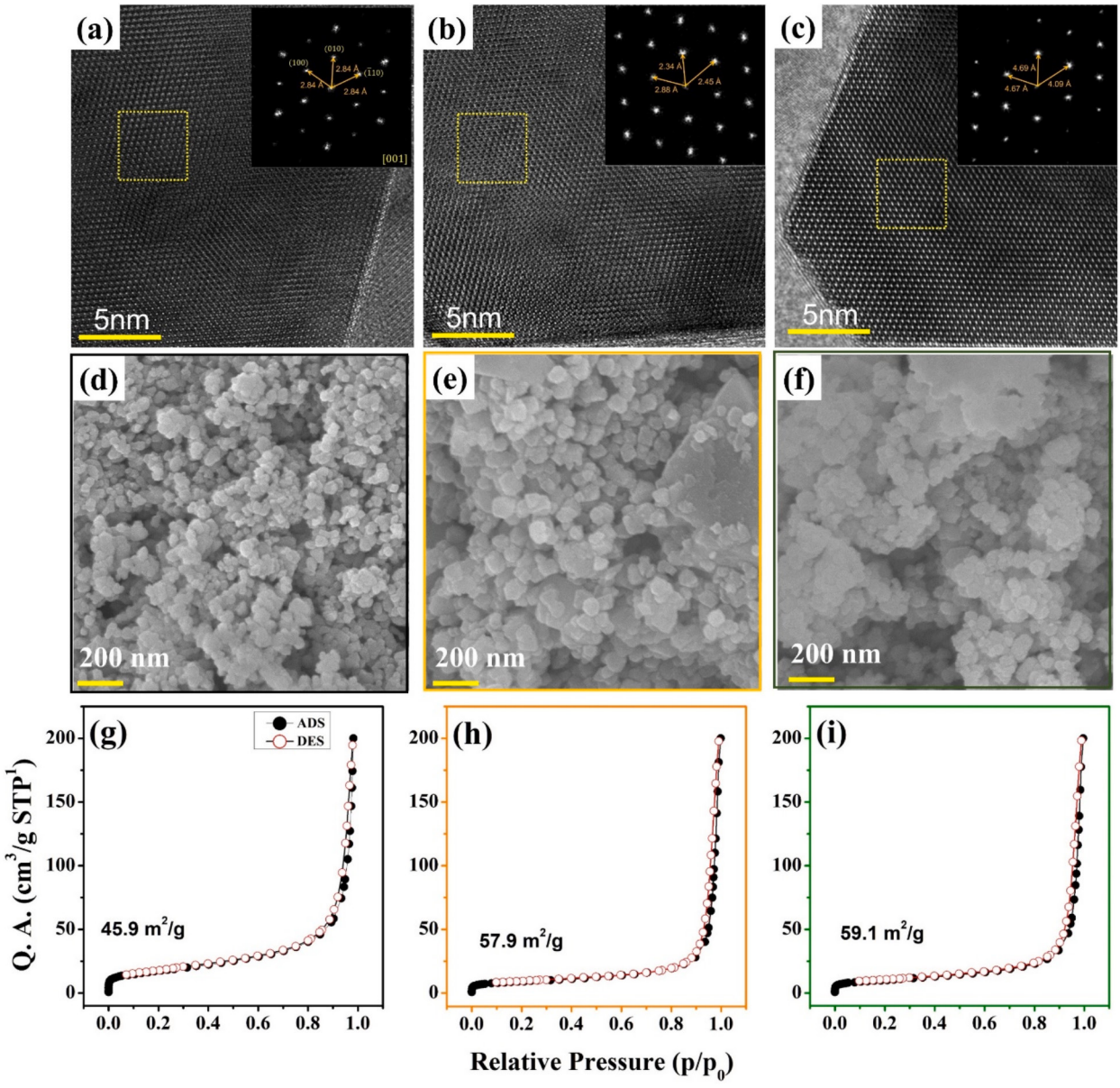


Fig. 1. HRTEM images and corresponding fast Fourier transform (FFT) images of (a) CoO, (b) CoO-B, and (c) 3 wt% Li@CoO-B NPs. SEM images of (d) CoO, (e) CoO-B, and (f) 3 wt% Li@CoO-B NPs. Brunauer–Emmett–Teller (BET) analysis of (g) CoO, (h) CoO-B, and (i) 3 wt% Li@CoO-B NPs.

Li@CoO-B NPs were also obtained, as summarized in Table 1.

The FFT results show changes in crystal distances for both the base treatment and additional Li-doped treatments, which are presumed to be related to changes in photocatalytic properties. For CoO-B NPs, lattice changes due to the formation of defect structures were detected, while for Li@CoO-B NPs, a significant lattice change was observed upon Li-ion doping. This surface reorganization is attributed to the formation of oxygen defect structures on the CoO NP surface, resulting from changes in the charge state after base treatment or Li-ion doping [21].

We also analyzed the SEM images (Fig. 1d–f) and BELsorp-Max system (BEL) measurements (Fig. 1g–i) to provide direct evidence supporting the experimental results. In contrast to the differences in crystallite distance observed by TEM, the SEM images of all three samples showed a rounded NP surface with a similar size of 50 nm. BET experiments were conducted to explore the relationship between the specific surface area of the NPs and their catalytic properties. The results showed that the average surface area of CoO-B and Li@CoO-B NPs was significantly larger than that of untreated CoO NPs. Thus, the BET experiments highlighted the substantial effect of modulation (CoO-B and Li@CoO-B) on the surface area of each sample, with the order of surface area being Li@CoO-B > CoO-B > CoO NPs. Based on the analysis of surface structure changes, significant photocatalytic changes are expected through the modulation of CoO NPs. Additionally, the elemental distribution of the synthesized NPs was analyzed via energy-dispersive X-ray spectroscopy (EDS), revealing that Co and O were homogeneously distributed throughout the NPs (Fig. S1).

Fig. 2a shows the XRD patterns of the three prepared CoO NPs. The diffraction peaks of the pristine CoO NPs at 36.5°, 42.4°, and 61.6° correspond to the (111), (200), and (220) crystal planes of the typical face-centered cubic CoO structure (JCPDS number 71–1178) [22]. The lattice parameters of the unit cells for pristine CoO are as follows: $a = b = c = 4.26$, $\alpha = \beta = \gamma = 90°$. A trace amount of cubic Co_3O_4 impurity (lattice parameters: $a = b = c = 8.08$, $\alpha = \beta = \gamma = 90°$) was also detected, with diffraction peaks appearing at 36.8°, 38.6°, 44.9°, 59.4°, and 65.2°, corresponding to the (220), (311), (400), (511), and (440) planes of Co_3O_4 NPs (JCPDS number 00–042-1467) [23]. Interestingly, for the two modified CoO NPs, the diffraction peaks corresponding to Co_3O_4 NPs became much more prominent compared to the pristine CoO NPs. The formation of a mixture of two different structures indicates that base treatment and Li ion doping on CoO NPs may lead to partial oxidation of Co^{2+} ions within CoO to Co^{3+} ions, generating defect structures (i.e., crystalline Co_3O_4 NPs) on the surface of CoO NPs. Moreover, the lattice parameter analysis of CoO and Co_3O_4 NPs revealed that the unit cells could vary depending on the treatment. For instance, in the case of CoO-B NPs, the lattice parameter of CoO remains identical to that of pristine CoO, while the lattice parameter of Co_3O_4 is determined to be $a = b = c = 8.1$, $\alpha = \beta = \gamma = 90°$. A slight extension of the unit cell was observed compared to the Co_3O_4 impurity in pristine CoO NPs. On the other hand, Li@CoO-B NPs resulted in a slight shrinking of the unit cell of Co_3O_4 ($a = b = c = 8.06$, $\alpha = \beta = \gamma = 90°$), while the unit cell of CoO NPs remained unchanged. Therefore, we propose that both CoO-B and Li@CoO-B NPs lead to the formation of Co_3O_4 NPs on the surface of CoO NPs through partial oxidation of Co^{2+} ions. The unit cells of the newly formed Co_3O_4 NPs can be modulated depending on the type of treatment. Although the peak intensity associated with Co_3O_4 NPs is smaller than that of pristine Co_3O_4 NPs, charge transfer can still be inferred. Partial oxidation of the Co^{2+} centers and formation of surface V_os was expected upon lithium and base treatments; however, the formation of defined domains of Co_3O_4 was incidental.

Pristine CoO NPs exhibit strong light absorption in the visible wavelength range due to their narrow bandgap [24]. The bandgap of the synthesized CoO NPs, as shown at the bottom of Fig. 2b, was 1.40 eV, according to the Tauc plot derived from UV–vis DR spectroscopy results. The optical properties and bandgap energies of the two heterostructured photocatalysts were confirmed by UV–vis DR absorption spectroscopy. The bandgaps of CoO-B and Li@CoO-B NPs were measured to be 1.38, and 1.34 eV, respectively, indicating that both samples can exhibit visible light photocatalytic activity [25]. To assess the charge separation efficiency of the photogenerated electron-hole pairs, the photocurrents of the pristine CoO, CoO-B, and Li@CoO-B NPs were examined under

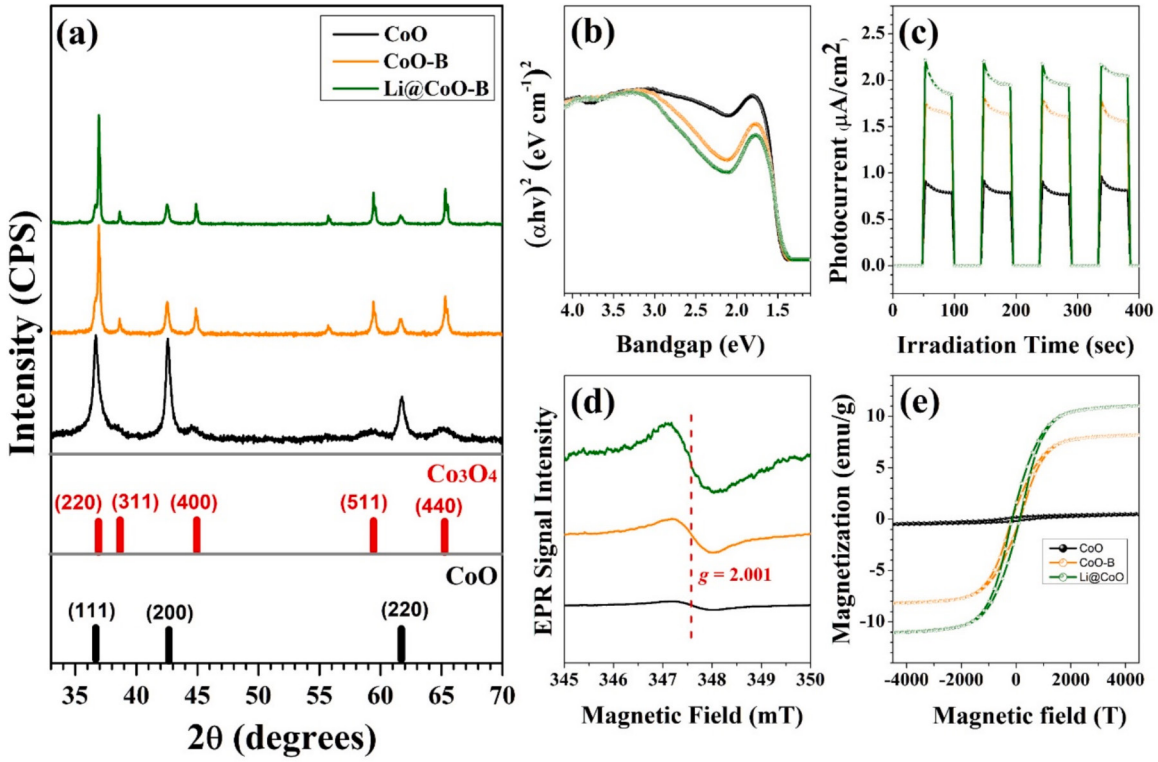


Fig. 2. (a) XRD, (b) UV–vis DRS, (c) the photocurrent measurements, (d) EPR, and (e) VSM curves of CoO (black), CoO-B (orange), and Li@CoO-B (olive) NPs.

visible light irradiation (λ = 445 nm), as shown in Fig. 2c. Transient photocurrent responses were measured under visible irradiation for four on–off cycles [26]. Upon 445 nm wavelength irradiation, all three CoO NPs generated photocurrents; however, the current densities immediately returned to zero when the light was turned off. Notably, a nearly two-fold increase in current density was observed for CoO NPs following base treatment or Li ion doping. This result supports the idea that the formation of V_os on the surface trapped electrons at specific defect sites, reducing the recombination of photogenerated charge carriers and resulting in a significantly higher current compared to the pristine CoO NPs. The enhancement can be predominantly attributed to the variations in the surface density of V_os leading to the reduction of recombination rate rather than other factors such as changes in the bulk band structure. To prove this mechanistic point, similar photocurrent measurements were performed in 0.025 M Na_2SO_3, a well-known hole scavenger (Fig. S6). Here, all three samples exhibited nearly identical photocurrent density, indicating that the overall light absorption and charge transport through the bulk band structure were identical across all samples (in conditions of facile hole scavenging at the particle-electrolyte interface). More importantly, the improvement in photocurrent due to the formation of V_os aligns well with the photocatalytic and electrocatalytic activity of CoO NPs.

We also conducted electron paramagnetic resonance (EPR) analysis to provide evidence of increased V_os in the three CoO NPs (Fig. 2d) [27]. The pristine CoO NPs, exhibiting a range of non-stoichiometry and surface defects, showed an EPR signal centered at $g = 2.001$, attributed to the Zeeman effect caused by a single electron trapped in anionic oxygen vacancies. For CoO-B, the EPR signal at $g = 2.001$ was also observed. Spin quantification by integrating the signal revealed that the V_os concentration was three times greater than that of the pristine CoO NPs [28]. In contrast, the EPR spectra of Li@CoO-B NPs displayed a stronger signal at $g = 2.001$, and spin quantification indicated a five-fold higher concentration of oxygen vacancies. These results confirm the presence of abundant V_os in the two modified CoO NPs compared to the pristine CoO NPs. Overall, the modification of CoO NPs by lithium ion deposition proved more effective than base treatment, directly contributing to the increased V_os on the surface of the modified CoO NPs. The magnetic properties of the pristine CoO, CoO-B, and Li@CoO-B NPs were assessed using a VSM (Fig. 2e). As is well known, CoO NPs exhibit paramagnetic properties, while Co_3O_4 NPs display weak ferromagnetic properties at room temperature [29]. As expected, the pristine CoO NPs exhibited typical paramagnetic behavior, while CoO-B and Li@CoO-B NPs showed weak ferromagnetic properties due to the formation of Co_3O_4 NPs on the surface. Notably, Li@CoO-B NPs exhibited stronger magnetic properties compared to CoO-B, a trend consistent with the photocurrent results shown in Fig. 2c. In sum, there are multiple factors that could contribute to the formation of V_os within CoO NPs upon Li ion doping or base treatment [30]. Due to a lower oxidation state of Li ion compared to that of the host Co ion, the charge neutrality of the nanoparticle irreversibly created stable V_os by vertu of the equation given below:

$$CoO + Li^{+}_{dopant} \rightarrow 2Li_{at\ \ substituted\ \ Co\ \ site} + O_{at\ \ regular\ \ O\ \ site} + V_o$$

Larger ionic radius of Li ion (0.90 Å) compared that of the host Co ion (0.79 Å for low-spin; 0.88 Å for high-spin) plausibly engendered the lattice distortion. In addition, addition of KOH during hydrothermal synthesis of CoO NPs could help the ionic conduction by creating aqueous electrolytic milieu as well as acting as stabilizer for basic reduction.

To establish the correlation between V_os induced by base treatment or Li ion deposition and photocatalytic performance, we evaluated the photocatalytic activities of pristine CoO, CoO-B, and Li@CoO-B in the photocatalytic conversion activity (PCA) of 2,5-hydroxymethylfurfural (HMF) [31]. As shown in Fig. 3a and Table 2, the PCA rates of HMF for CoO-B and Li@CoO-B NPs were significantly higher than those for pristine CoO NPs. The PCA followed the order: Li@CoO-B > CoO-B > CoO NPs. These results highlight the critical role of charge transfer, facilitated by surface modification and the increased abundance of V_os, in enhancing the photocatalytic performance of CoO NPs [32].

A linear correlation between $\ln(C/C_0)$ and reaction time was obtained for the three CoO NPs, and the degradation reaction processes were fitted with pseudo-first-order kinetics to determine the first-order

Table 2
C/C_0 values for PCA of HMF degradation and the quantities of FDCA obtained from HMF using CoO, CoO-B, and Li@CoO-B NPs after 24 h of 445 nm irradiation.

445 nm	CoO	CoO-B	Li@CoO-B
HMF degradation (C/C_0)	0.79 ± 0.039	0.35 ± 0.018	0.02 ± 0.003
FDCA (%)	8	45	59

* Each C and C_0 indicate time-dependent and initial concentrations of HMF.

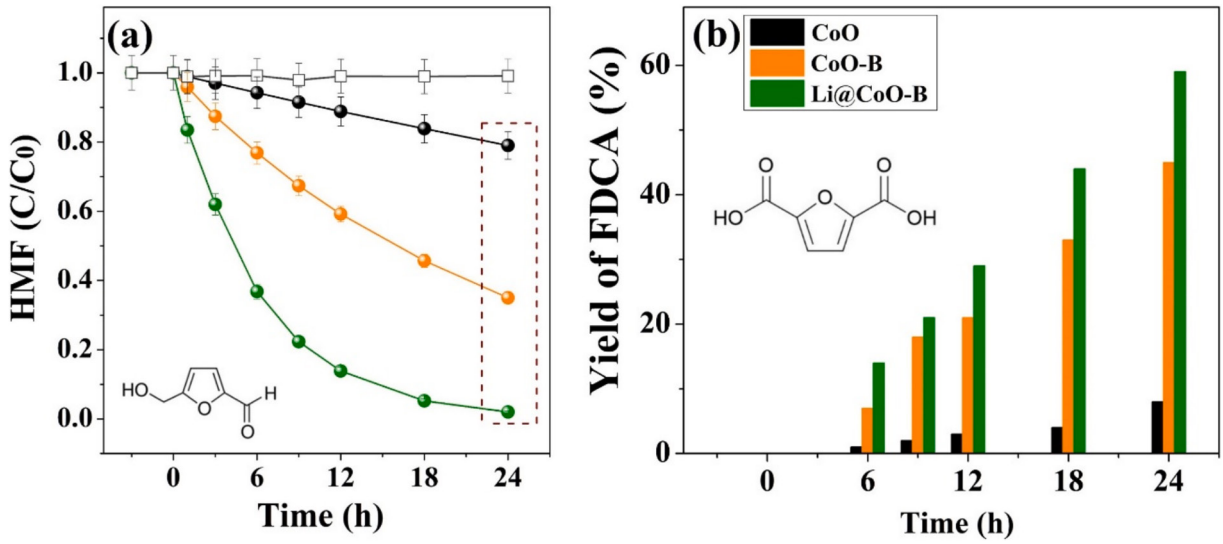


Fig. 3. Photocatalytic conversion activities of (a) HMF using three different CoO NPs with and without DMPO (gray): pristine CoO (black), CoO-B (orange), and Li@CoO-B (olive) NPs. Experimental conditions: [catalyst] = 0.5 g/L, λ = 445 nm, $[HMF]_0$ = 25 mM. (b) Efficiency of FDCA conversion from HMF over time under 445 nm irradiation.

rate constants (k) for the photocatalytic reaction of each CoO NPs (Fig. S2). The k values for HMF PCA were calculated as 0.00958, 0.0437, and 0.126 h^{-1} for pristine CoO (black), CoO-B (orange), and Li@CoO-B (olive), respectively. Kinetically, Li@CoO-B NPs accelerated the HMF PCA reaction rate by 16.01 times compared to CoO NPs.

Understanding the yield of the final product and the exact mechanistic pathway of HMF photolysis is crucial for assessing photocatalytic performance. HMF contains two distinct functional groups: an alcohol group (–OH) and an aldehyde group (–CHO). The selective oxidation of –OH or –CHO leads to the formation of 2,5-diformylfuran (DFF) or 5-hydroxymethyl-2-furancarboxylic acid (HMFCA) as intermediates, respectively [33]. These intermediates are further oxidized to produce the final product, 2,5-furandicarboxylic acid (FDCA) (Fig. S3). In this system, we observed that in all samples, the –OH group of HMF was first selectively oxidized to –CHO, resulting in the formation of DFF as an intermediate, which is then converted to FDCA as the final product, despite changes in the ratio between the two substances [34]. The V_os formed on the surface of the photocatalytic MO NPs can trap electrons and O_2 to promote the formation of $\cdot O^{2-}$. Consequently, HMF is converted to FDCA via photocatalytic oxidation under the action of $\cdot O^{2-}$ and h^+. To confirm the involvement of reactive oxygen species, such as $\cdot O^{2-}$, in the photocatalytic system, we used 5,5-dimethyl-1-pyrroline N-oxide (DMPO) as a $\cdot O^{2-}$ radical scavenger, which completely inhibited the photocatalytic activity (Fig. 3a). As shown, the preferential reaction of DMPO with the radicals formed during the photoreaction interfered with the photoreaction of all three CoO NPs, supporting our proposed mechanism.

In conclusion, the selective oxidation of HMF using the three samples dispersed in an aqueous solution confirmed that, under all conditions, the reaction pathway from HMF to DFF proceeds to the final product, FDCA (Scheme S1). Additionally, the photocatalytic activity of the three tested NPs remained above 97 % after even ten reuse cycles (Fig. S4).

The chemical composition and state were analyzed using XPS (Fig. 4a and 4b). The Co 2p spectra of CoO NPs displayed two prominent peaks at binding energies of 781.2 eV (marked A) and 796.8 eV, along with two distinct satellite peaks corresponding to the Co^{2+} oxidation states of the Co $2p_{3/2}$ and Co $2p_{1/2}$ orbitals [35,36]. In contrast, both CoO-B and Li@CoO-B NPs exhibited distinct electronic structures compared to the pristine CoO NPs. The binding energies for the Co $2p_{3/2}$ and Co $2p_{1/2}$ peaks of CoO-B and Li@CoO-B appeared at 783 ± 0.1 eV (marked B) and 798 ± 0.5 eV, respectively. The binding energy difference between the Co $2p_{3/2}$ and Co $2p_{1/2}$ peaks was calculated to be approximately 15 eV, a characteristic feature of the Co_3O_4 phase [37]. Although neither synthetic method produced bulk Co_3O_4 NPs, a mixed charge state of Co^{3+} and Co^{2+} was found at the surface of the modified CoO NPs, partially revealing an electronic structure similar to that of Co_3O_4 NPs. The deconvoluted O 1 s XPS spectra of pristine CoO, CoO-B, and Li@CoO-B NPs displayed three distinct oxygen components, corresponding to Co-O (529.9 eV), V_{oS} (531.7 eV), and –OH (532.7 eV), respectively. Among the three CoO NPs, CoO-B and Li@CoO-B NPs showed a significant increase in the proportion of V_os peaks [38], indicating that surface charge modulation from + 2 to + 3 states led to

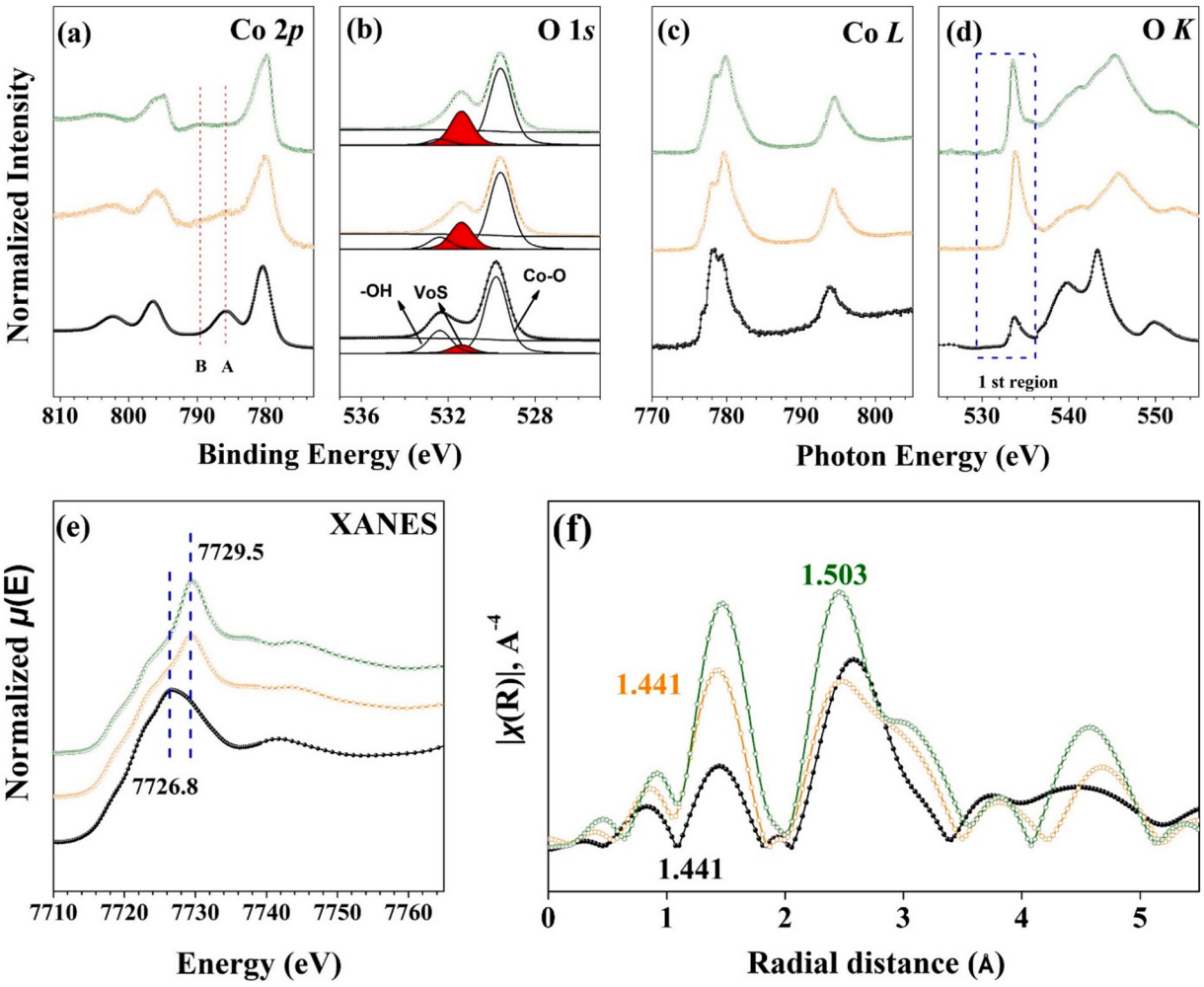


Fig. 4. XPS spectra of (a) Co 2p, and (b) O 1s core level spectra; XAS spectra of (c) Co L-edge and (d) O K-edge; (e) overlaid XANES data; and (f) Fourier-transform EXAFS data of CoO (black), CoO-B (orange), and Li@CoO-B (olive).

an abundance of V_os.

To gain insight into the electronic structure changes at the surface of CoO NPs due to base treatment and Li ion doping, we obtained soft XAS profiles at the Co *L*-edge and O *K*-edge (Fig. 4c and d). We analyzed the Co *L*-edge spectra to examine how the electronic structure of CoO NPs is altered by base treatment and Li ion doping. As shown in Fig. 4c, the oxidation state of cobalt in CoO NPs is primarily Co^{2+}, while Co^{2+} and Co^{3+} coexist on the surface of CoO-B and Li@CoO-B NPs. This result aligns with the findings for Co_3O_4 NPs, as shown by XRD in Fig. 2a. Notably, base treatment and Li ion doping led to the formation of different Co ion charge states on the CoO surface. Specifically, Co^{2+} was converted to Co^{3+} on CoO-B and Li@CoO-B NPs. The O *K*-edge spectra of pristine CoO, CoO-B, and Li@CoO-B NPs can be divided into two regions, as shown in Fig. 4d. The first region (529–536 eV; blue box) corresponds to the transition of electrons from the O 1 *s* state to the unoccupied O 2*p* state hybridized with the 3*d* states of the transition metal. The second region (537–545 eV) is attributed to the O 2*p* state hybridized with the Co 4 *s* and 4*p* states. This study focuses on the changes in the intensity ratio of peaks in the first region. The peak for pristine CoO NPs corresponds to the t_{2g} and e_g states, respectively. For pristine CoO NPs, the peak intensity is smaller than that of CoO-B and Li@CoO-B NPs, which is consistent with previous results [39]. These differences in the O *K*-edge spectra between CoO and CoO-B or Li@CoO-B NPs are likely due to the varying oxidation states and occupancy of Co ions on the surface of the CoO NPs.

Furthermore, an in-depth investigation of the atomic-scale bonding structures was conducted using X-ray absorption near-edge structure (XANES) spectra to examine the electronic properties and oxidation states of cobalt centers in the three different CoO NPs (Fig. 4e) [40]. The Co *K*-edge absorption energy of pristine CoO NPs was observed at 7726.8 eV, while that of CoO-B and Li@CoO-B NPs appeared at 7729.5 eV. This 2.7 eV shift in the Co *K*-edge absorption energy strongly indicates a higher oxidation state in the two modified CoO NPs compared to pristine CoO NPs. The increased absorption energy is comparable to that of the Co_3O_4 spinel structure, in which the main oxidation state of Co was calculated to be approximately 2.65, with 8 Co^{2+} cations in tetrahedral geometry and 16 Co^{3+} cations in octahedral geometry [41,42]. Thus, both base and lithium ion treatments on pristine cobalt oxide resulted in an increase in the average oxidation state of the cobalt centers. Additionally, the increased pre-edge intensities in the synthesized CoO NPs suggest a significant deviation from centrosymmetric, likely due to structural alterations leading to the formation of tetrahedral cobalt centers [43].

The local coordination environments of cobalt centers in differently synthesized CoO NPs were investigated using extended X-ray absorption fine structure (EXAFS) analysis (Fig. 4f). The first peak, 1.44 Å < r < 1.50 Å, arises from backscattering by neighboring oxygen atoms and corresponds to Co^{2+} and Co^{3+} cations in tetrahedral and octahedral geometries. In contrast to the spectrum of pristine CoO NPs, the other synthesized CoO NPs (CoO-B and Li@CoO-B) displayed second and third peaks at approximately 2.45 Å and 2.95 Å, due to backscattering by neighboring cobalt atoms related to the Co_{oct}-Co_{oct} and Co_{tet}-Co_{oct} coordination shells, respectively [44]. This clearly indicates the presence of two different structures in the other synthesized cobalt oxides. Since the Co-O interatomic distances in different geometries were too close to be resolved, they overlapped into a main Co-O shell with interatomic bond distances of 2.11 Å, 1.89 Å, and 1.92 Å for pristine CoO, CoO-B, and Li@CoO-B NPs, respectively. While the pristine CoO NPs contained only Co(II) ions, the other synthesized CoO NPs exhibited a mixed valence of Co(II) and Co(III) ions after base treatment or Li-ion doping. This clearly resulted in shorter Co-O bond distances in the synthesized cobalt oxides (1.89–1.92 Å) compared to pristine CoO (2.11 Å) due to the decrease in the ionic radii of cobalt centers [45]. Notably, the Co-O bond distances varied moderately depending on the synthetic method used for the pristine cobalt oxide. When the pristine cobalt oxide was doped with lithium, the Co-O bond distance increased compared to that of the base-treated CoO NPs. The elongation of the Co-O bond distances reflected an increased number of defect sites within Li@CoO-B NPs. In summary, Co-O bond distances decreased with the increasing average oxidation state of the cobalt centers, and depending on the synthetic procedure, these distances could be modulated, as shown by the Fourier-transform of the k^3-weighted EXAFS spectrum (Fig. S5).

As shown in Scheme 1, all three samples exhibit similar bandgaps and can absorb light in the visible region, highlighting their potential as visible light photocatalysts. However, a significant difference exists due to variations in the density of V_os, a key indicator of photocatalytic activity. Specifically, Li@CoO-B NPs exhibit the highest V_os density, resulting in the best selective oxidation reactivity of HMF under visible light irradiation. In a typical photocatalytic reaction, carriers are generated near the surface (VB) due to the charge separation [46]. Because of the short carrier diffusion length, surface defects such as V_os act as trap states, enhancing photochemical reactivity and driving the photoreaction [47,48]. This explains the significantly improved visible light photocatalytic properties of Li@CoO-B NPs, which have the highest V_os density.

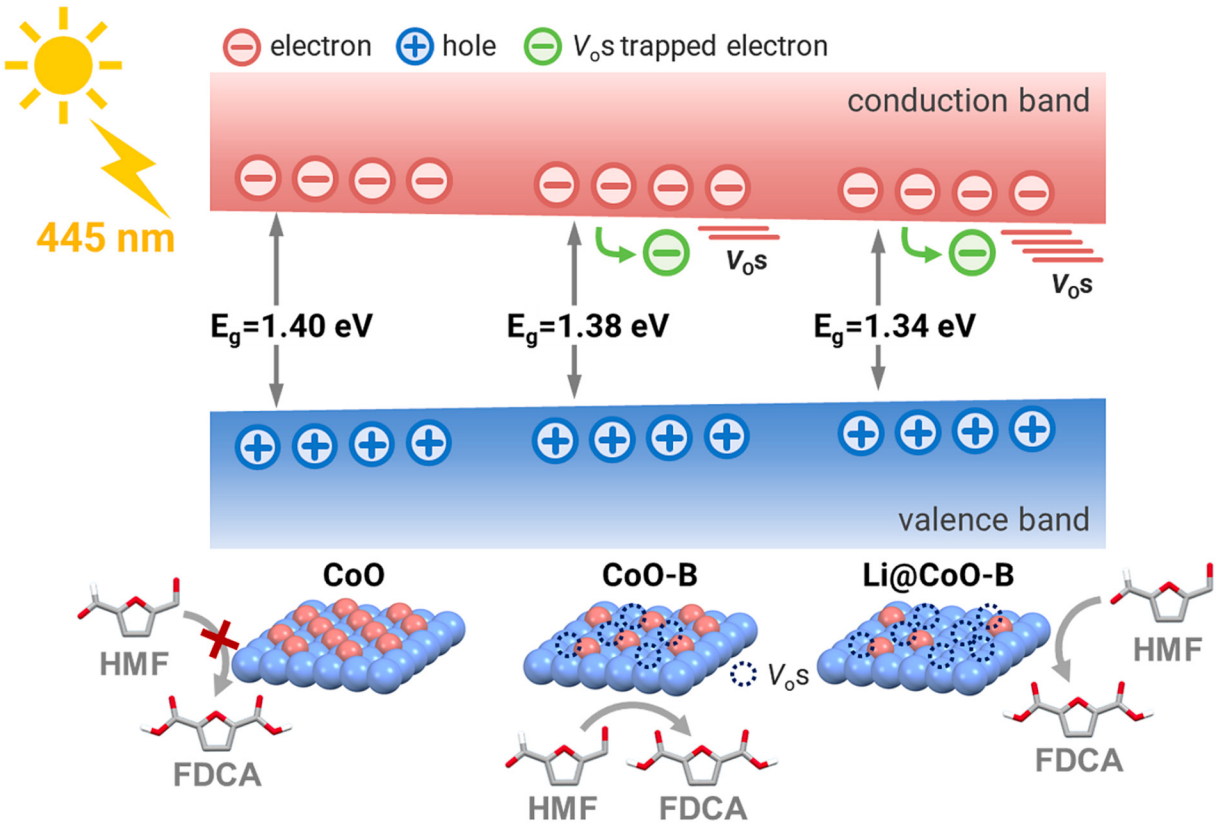
Based on the experimental results, we infer that base treatment and Li ion doping effectively generate V_os through smooth charge transfer by altering the charge state of cobalt in CoO NPs (from Co^{2+}) [49]. In other words, both modulation methods are expected to form Co_3O_4 on the surface of CoO NPs, creating V_os due to the charge mismatch with the Co^{2+} ions. In conclusion, the increase in photocatalytic conversion activity can be attributed to the rise in V_os, which results from the charge mismatch caused by these modulation treatments [50]. Additionally, considering the Shannon effective ionic radii, the ionic sizes of the cations involved are 0.72 Å for Co^{2+}, 0.74 Å for Li^+, and 0.60 Å for Co^{3+} [51]. Therefore, lattice deformation from cation substitution during both modulation treatments may contribute to the formation of atomic defects that alter the electronic structure of the host CoO NPs, significantly enhancing their visible light photocatalytic properties.

4. Conclusion

This study successfully synthesized CoO NPs with varying V_os through base treatment and Li ion deposition. Various spectroscopic techniques, including XPS and EXAFS, revealed that the abundance of V_os defects in base-treated or Li-ion-deposited CoO NPs significantly enhanced their photocatalytic properties. Our findings suggest that surface modification-induced V_os defects provide a promising approach for developing novel visible light-active photocatalysts, enhancing biomass conversion efficiency. Furthermore, this strategy is crucial for improving photocatalyst properties, depending on the chosen modification method.

CRediT authorship contribution statement

Hyeri Jeon: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sunyoung Hwang:** Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Taeyeon Kang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Seungwoo Hong:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Conceptualization. **Hyun S. Ahn:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Hangil Lee:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.



Scheme 1. Depiction of the distinct reaction and charge transport mechanisms in photocatalysis by (a) CoO, (b) CoO-B, and (c) Li@CoO-B NPs.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (RS-2024-00346153 for H. Lee, RS-2023-00272031 for H. Jeon, 2021R1A6A1A10039823 for S. Hong, and RS-2024-00337549 for H. S. Ahn). H. S. Ahn also acknowledges support from the Global - Learning & Academic research institution for Master's·PhD students, and Postdocs (LAMP) Program (RS-2024-00442483) funded by the Ministry of Education. This research was also supported by the Technology Innovation Program (20022479) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

Appendix A. Supplementary data

EDS mapping images, Fourier transformed EXAFS spectrum, LC-MS product analysis, HMF decomposition data and calculation of the rate constant of the photocatalytic reaction of each NPs. (PDF).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.inoche.2025.114287.

Data availability

Data will be made available on request.

References

[1] V. Etacheri, C. Di Valentin, J. Schneider, D. Bahnemann, S.C. Pillai, Visible-Light Activation of TiO_2 Photocatalysts: Advances in Theory and Experiments, J. Photochem. Photobiol. C: Photochem. Rev. 25 (2015) 1–29.
[2] T. Iihoshi, T. Ohwaki, J.J.M. Vequizo, A. Yamakata, Improvement of Photocatalytic Activity under Visible-Light Irradiation by Heterojunction of Cu Ion Loaded WO_3 and Cu Ion Loaded N-TiO_2, Appl. Catal. B: Environ. 248 (2019) 249–254.
[3] C. Li, J. Li, L. Qin, P. Yang, D.G. Vlachos, Recent Advances in the Photocatalytic Conversion of Biomass-Derived Furanic Compounds, ACS Catal. 11 (2021) 11336–11359.
[4] Q. Wang, K. Dastafkan, C. Zhao, Design Strategies for Non-Precious Metal Oxide Electrocatalysts for Oxygen Evolution Reactions, Curr. Opin. Electrochem. 10 (2018) 16–23.
[5] C. Chen, L. Wang, B. Zhu, Z. Zhou, S.I. El-Hout, J. Yang, J. Zhang, 2, 5-Furandicarboxylic Acid Production via Catalytic Oxidation of 5-Hydroxymethylfurfural: Catalysts, Processes and Reaction Mechanism, J. Energy Chem. 54 (2021) 528–554.
[6] P. Zhu, W. Zhang, Q. Li, H. Xia, Visible-Light-Driven Photocatalytic Oxidation of 5-Hydroxymethylfurfural to 2, 5-Furandicarboxylic Acid over Plasmonic Au/ZnO Catalyst, ACS Sustain. Chem. Eng. 10 (2022) 8778–8787.
[7] J. Schneider, M. Matsuoka, M. Takeuchi, J. Zhang, Y. Horiuchi, M. Anpo, D. W. Bahnemann, Understanding TiO_2 Photocatalysis: Mechanisms and Materials, Chem. Rev. 114 (2014) 9919–9986.
[8] W. Yu, J. Zhang, T. Peng, New Insight into the Enhanced Photocatalytic Activity of N-, C-and S-Doped ZnO Photocatalysts, Appl. Catal. B: Environ. 181 (2016) 220–227.
[9] M. Xue, D. Pan, J. Zhao, J. Chen, Optically Tunable Transient Plasmons in InSb Nanowires, Adv. Mater. 35 (2023) 2208952.
[10] L. Liao, Q. Zhang, Z. Su, Z. Zhao, Y. Wang, Y. Li, X. Lu, D. Wei, G. Feng, Q. Yu, X. Cai, J. Zhao, Z. Ren, H. Fang, F. Robles-Hernandez, S. Baldelli, J. Bao, Efficient Solar Water-Splitting Using a Nanocrystalline CoO Photocatalyst, Nat. Nanotechnol. 9 (2014) 69–73.
[11] Y. Tang, L. Dong, S. Mao, H. Gu, T. Malkoske, B. Chen, Enhanced Photocatalytic Removal of Tetrabromobisphenol A by Magnetic CoO@Graphene Nanocomposites under Visible-Light Irradiation, ACS Appl. Energy Mater. 1 (2018) 2698–2708.
[12] F. Guo, W. Shi, C. Zhu, H. Li, Z. Kang, CoO and g-C_3N_4 Complement Each Other for Highly Efficient Overall Water Splitting under Visible Light, Appl. Catal. B: Environ. 226 (2018) 412–420.
[13] H.T. Bui, S. Weon, J.W. Bae, E.J. Kim, B. Kim, Y.Y. Ahn, K. Kim, H. Lee, W. Kim, Oxygen Vacancy Engineering of Cerium Oxide for the Selective Photocatalytic Oxidation of Aromatic Pollutants, J. Hazard. Mater. 404 (2021) 123976.
[14] Y. Yamazaki, N. Doshita, K. Mori, Y. Kuwahara, H. Kobayashi, H. Yamashita, Effect of Oxygen Vacancies and Crystal Phases in Defective Pt/ZrO_{2-x} on Its

Photocatalytic Activity toward Hydrogen Production, Cat. Sci. Technol. 14 (2024) 397–404.
[15] L. Hao, H. Huang, Y. Zhang, T. Ma, Oxygen Vacant Semiconductor Photocatalysts, Adv. Funct. Mater. 31 (2021) 2100919.
[16] R. Shi, G. Chen, W. Ma, D. Zhang, G. Qiu, X. Liu, Shape-Controlled Synthesis and Characterization of Cobalt Oxides Hollow Spheres and Octahedra, Dalton Trans. 41 (2012) 5981–5987.
[17] B. Ravel, M.J. Newville, Athena, Artemis, Hephaestus: data analysis for X-ray absorption spectroscopy using IFEFFIT, J. Synchrot. Radiat. 12 (2005) 537.
[18] J.J. Rehr, R.C. Albers, Theoretical approaches to x-ray absorption fine structure, Rev. Mod. Phys. 72 (2000) 621.
[19] Y. Zhang, J. Zhu, X. Song, X. Zhong, Controlling the Synthesis of CoO Nanocrystals with Various Morphologies, J. Phys. Chem. C 112 (2008) 5322–5327.
[20] X. Kuang, B. Kang, Z. Wang, L. Gao, C. Guo, J.Y. Lee, X. Xun, Q. Wei, Sulfur-Doped CoO Nanoflakes with Loosely Packed Structure Realizing Enhanced Oxygen Evolution Reaction, Chem. - Eur. J 24 (2018) 17288–17292.
[21] Y. Wei, Z. Zhang, C. Mei, J. Tan, Z. Wang, J. Li, L. Gan, Facet Dependence and Vacancy-Controlled Reversibility of the Spinel-to-Rocksalt Phase Transformation at Co_3O_4 Surfaces, Chem. Mater. 35 (2023) 4461–4470.
[22] W. Shi, F. Guo, C. Zhu, H. Wang, H. Li, H. Huang, Y. Liu, Z. Kang, Carbon Dots Anchored on Octahedral CoO as A Stable Visible-Light-Responsive Composite Photocatalyst for Overall Water Splitting, J. Mater. Chem. A 5 (2017) 19800–19807.
[23] I. Shaheen, K.S. Ahmad, C. Zequine, R.K. Gupta, A.G. Thomas, M.A. Malik, Modified Sol-Gel Synthesis of Co_3O_4 Nanoparticles Using Organic Template for Electrochemical Energy Storage, Energy 218 (2021) 119502.
[24] Y. Zhu, T. Wan, X. Wen, D. Chu, Y. Jiang, Tunable Type I and II Heterojunction of CoO_x Nanoparticles Confined in g-C_3N_4 Nanotubes for Photocatalytic Hydrogen Production, Appl. Catal. B: Environ. 244 (2019) 814–822.
[25] Z. Lin, C. Du, B. Yan, G. Yang, Two-Dimensional Amorphous CoO Photocatalyst for Efficient Overall Water Splitting with High Stability, J. Catal. 372 (2019) 299–310.
[26] H. Ying, Z. Huang, G. Dong, Y. Zhang, The Charge Transfer Pathway of CoO QDs/g-C_3N_4 Composites for Highly Efficient Photocatalytic Hydrogen Evolution, J. Photochem. Photobiol. A: Chem. 415 (2021) 113305.
[27] K. Zhang, G. Zhang, J. Qu, H. Liu, Disordering the Atomic Structure of Co (II) Oxide via B-Doping: An Efficient Oxygen Vacancy Introduction Approach for High Oxygen Evolution Reaction Electrocatalysts, Small 14 (2018) 1802760.
[28] H. Chu, D. Zhang, P. Feng, Y. Gu, P. Chen, K. Pan, H. Xie, M. Yang, Engineering Oxygen Vacancies in CoO@Co_3O_4/C Nanocomposites for Enhanced Electrochemical Performances, Nanoscale 13 (2021) 19518–19526.
[29] X. He, X. Song, W. Qiao, Z. Li, X. Zhang, S. Yan, W. Zhong, Y. Du, Phase-and Size-Dependent Optical and Magnetic Properties of CoO Nanoparticles, J. Phys. Chem. C 119 (2015) 9550–9559.
[30] A. Sarkar, G.G. Khan, The Formation and Detection Techniques of Oxygen Vacancies in Titanium Oxide-based Nanostructures, Nanoscale 11 (2019) 3414.
[31] D.T. Hoang, D. Lim, M.J. Kang, H. Yu, V.N. Pham, J.H. Park, K. Ihm, H.S. Kim, H. S. Ahn, H. Lee, Multi-Purpose Improvements in Catalytic Activity for Li-Ion Deposited TiO_2, SnO_2, and CeO_2 Nanoparticles through Oxygen-Vacancy Control, Adv. Sustain. Syst. 2400094 (2024).
[32] K. Chen, T. Jiang, T. Liu, J. Yu, S. Zhou, A. Ali, S. Wang, Y. Liu, L. Zhu, X. Xu, Zn Dopants Synergistic Oxygen Vacancy Boosts Ultrathin CoO Layer for CO_2 Photoreduction, Adv. Funct. Mater. 32 (2022) 2109336.
[33] Y. Yang, T. Mu, Electrochemical Oxidation of Biomass Derived 5-Hydroxymethyl-furfural (HMF): Pathway, Mechanism Catalysts and Coupling Reactions, Green Chem. 23 (2021) 4228–4254.
[34] M. Sajid, X. Zhao, D. Liu, Production of 2, 5-Furandicarboxylic Acid (FDCA) from 5-Hydroxymethylfurfural (HMF): Recent Progress Focusing on the Chemical-Catalytic Routes, Green Chem.. 20 (2018) 5427–5453.
[35] W. Fu, Y. Du, J. Jing, C. Fu, M. Zhou, Highly Selective Nitrate Reduction to Ammonia on CoO/Cu Foam via Constructing Interfacial Electric Field to Tune Adsorption of Reactants, Appl. Catal. B: Environ. 324 (2023) 122201.
[36] K. Zhang, X. Xia, S. Deng, D. Xie, Y. Lu, Y. Wang, J. Wu, X. Wang, J. Tu, N-doped CoO Nanowire Arrays as Efficient Electrocatalysts for Oxygen Evolution Reaction, J. Energy Chem. 37 (2019) 13–17.
[37] S.Y. Zhang, Z. Li, X. Shen, J. Shan, J. Zhan, H. Zhou, X. Yi, H.Y. Lian, Y. Liu, Formulating the Li Sites of Li@CoO-Bx Composites for Achieving High-Efficiency Oxidation Removal of Formaldehyde over the Ag/Li@CoO-Bx Catalyst Under Ambient Conditions, Environ. Res. 235 (2023) 116683.
[38] H. Wang, E.M. Feng, Y.M. Liu, C.Y. Zhang, High-Performance Hierarchical Ultrathin Sheet-Based CoOOH Hollow Nanospheres with Rich Oxygen Vacancies for the Oxygen Evolution Reaction, J. Mater. Chem. A 7 (2019) 7777–7783.
[39] H.C. Choi, Y.M. Jung, I. Noda, S.B. Kim, A Study of the Mechanism of the Electrochemical Reaction of Lithium with CoO by Two-Dimensional Soft X-ray Absorption Spectroscopy (2D XAS), 2D Raman, and 2D Heterospectral XAS−Raman Correlation Analysis, J. Phys. Chem. B 107 (2003) 5806–5811.
[40] D. Gajdek, P.A. Olsson, S. Blomberg, J. Gustafson, P.A. Carlsson, D. Haase, E. Lundgren, L.R. Merte, Structural Changes in Monolayer Cobalt Oxides under Ambient Pressure CO and O_2 Studied by In Situ Grazing-Incidence X-ray Absorption Fine Structure Spectroscopy, J. Phys. Chem. C 126 (2022) 3411–3418.
[41] P. Prieto, J.F. Marco, A. Serrano, M. Manso, J. de la Figuera, Highly Oriented (111) CoO and Co_3O_4 Thin Films Grown by Ion Beam Sputtering, J. Alloy. Compd. 810 (2019) 151912.
[42] E. Mijit, K. Chen, J.E.F. Rodrigues, Z. Hu, L. Nataf, A. Trapananti, A.D. Cicco, F. Baudelet, Crystal and Electronic Structure of Co_3O_4 Spinel under Pressure Probed by XANES and Raman Spectroscopy, Phys. Rev. B 103 (2021) 024105.
[43] K. Lienau, C.A. Triana, L. Reith, S. Siol, G.R. Patzke, Microwave-Hydrothermal Tuning of Spinel-Type C Co_3O_4 Water Oxidation Catalysts, Front. Chem. 8 (2020) 473.
[44] F.T. Haase, A. Bergmann, T.E. Jones, J. Timoshenko, A. Herzog, H.S. Jeon, C. Rettenmaier, B.R. Cuenya, Size Effects and Active State Formation of Cobalt Oxide Nanoparticles during the Oxygen Evolution Reaction, Nat. Energy 7 (2022) 765–773.
[45] Y. Tong, P. Chen, T. Zhou, K. Xu, W. Chu, C. Wu, Y. Xie, A Bifunctional Hybrid Electrocatalyst for Oxygen Reduction and Evolution: Cobalt Oxide Nanoparticles Strongly Coupled to B, N-decorated Graphene. Angew. Chem. Int. Ed. 56 (2017) 7121–7125.
[46] B. Choudhury, S. Bayan, A. Choudhury, P. Chakraborty, Narrowing of Band Gap and Effective Charge Carrier Separation in Oxygen Deficient TiO_2 Nanotubes with Improved Visible Light Photocatalytic Activity, J. Colloid Interface Sci.. 465 (2016) 1–10.
[47] L. Wang, D. Lv, F. Dong, X. Wu, N. Cheng, J. Scott, X. Xu, W. Hao, Y. Du, Boosting Visible-Light-Driven Photo-Oxidation of BiOCl by Promoted Charge Separation via Vacancy Engineering, ACS Sustain. Chem. Eng. 7 (2019) 3010–3017.
[48] D. Cui, L. Wang, K. Xu, L. Ren, L. Wang, Y. Yu, Y. Du, W. Hao, Band-Gap Engineering of BiOCl with Oxygen Vacancies for Efficient Photooxidation Properties under Visible-Light Irradiation, J. Mater. Chem. A 6 (2018) 2193–2199.
[49] J. Lian, D. Yin, S. Zhao, X. Zhu, Q. Liu, X. Zhang, X. Zhang, Core-Shell Structured Ag-CoO Nanoparticles with Superior Peroxidase-Like Activity for Colorimetric Sensing Hydrogen Peroxide and o-Phenylenediamine, Colloids Surf. A Physicochem. Eng. Asp 603 (2020) 125283.
[50] V.N. Pham, H. Jeon, S. Hong, H. Lee, Selective Oxidation of Biomass Molecules via ZnO Nanoparticles Modified Using Charge Mismatch of the Doped Co Ions, Inorg. Chem. 61 (2022) 16887–16894.
[51] S. Phadke, R. Mysyk, M. Anouti, Effect of Cation (Li^+, Na^+, K^+, Rb^+, Cs^+) in Aqueous Electrolyte on the Electrochemical Redox of Prussian Blue Analogue (PBA) Cathodes, J. Energy Chem. 40 (2020) 31–38.

Hyeri Jeon completed his doctoral training in Chemistry at Sookmyung Women's University, followed by a postdoctoral work at Ewha Womans University in 2023.

Sunyoung Hwang completed her Bachelor's degree program in Chemistry at Sookmyung Women's University, followed by a Master's program at Sookmyung Women's University in 2025.

Taeyeon Kang received his B.S. degree in Chemistry from Yonsei University in 2023. He is currently pursuing a Ph.D. at Yonsei University, specializing in the electrochemical synthesis of nanoparticles for catalytic applications. His research interests include nanomaterial synthesis, electrocatalysis, and heterogeneous catalysis.

Seungwoo Hong completed his doctoral training in Chemistry at Sogang University, followed by a postdoctoral work at Harvard University. He then began his independent career at Sookmyung Women's University in 2016, and moved to Ewha Womans University in 2023, where he currently serves as an associate professor.

Hyun S. Ahn completed his doctoral training in Chemistry at the University of California at Berkeley, followed by a postdoctoral work at the University of Texas at Austin. He then began his independent career at Yonsei University in 2017, where he currently serves as an associate professor.

Hangil Lee completed his doctoral training in Chemistry at the KAIST, followed by a postdoctoral work at the national synchrotron light source. He then began his independent career at Sookmyung women's university in 2008, where he currently serves as a professor.

Dalton Transactions

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Dalton Trans.*, 2025, **54**, 5546

Enhancing visible-light-driven photocatalysis of Pd- and Pt-doped WO_3 nanoparticles: the role of oxygen vacancies and bandgap narrowing†

Hyeri Jeon,‡[a,b] Dung Thanh Hoang,‡[c] Hyejin Yu,‡[d] Sunyoung Hwang,[c] Hyun Sung Kim, *[d] Seungwoo Hong *[a,b] and Hangil Lee *[c]

Received 25th December 2024,
Accepted 27th February 2025

DOI: 10.1039/d4dt03540b

rsc.li/dalton

The global energy demand has driven the development of efficient and cost-effective visible-light-activated photocatalysts for the synthesis of fine chemicals. However, most high-performance photocatalysts possess bandgaps exceeding ~3.0 eV, limiting their photocatalytic efficiency under visible light. In this study, Pd- and Pt-doped WO_3 nanoparticles were synthesized. Doping induced oxygen vacancies, which act as electron traps, reducing the bandgap and enhancing visible-light-driven photocatalytic activity. The photocatalytic performance was examined using hydroxymethylfurfural and benzyl alcohol as model substrates. The product yields for both substrates in the presence of Pd-doped WO_3 nanoparticles exceeded 95%. This work demonstrates a simple strategy for enhancing the solar-energy-driven photocatalytic efficiency of metal oxide nanoparticles, promoting sustainable fine chemical synthesis.

Open Access Article. Published on 03 March 2025. Downloaded on 3/25/2025 10:21:40 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Introduction

Owing to increasing energy demands and depleting fossil fuel reserves, researchers are focusing on harnessing solar energy through photocatalysis. Solar energy is a renewable, clean, and inexhaustible resource.[1,2] Consequently, solar-energy-driven photocatalysis is a sustainable, green method. Recent research on photocatalysis has focused on converting a wide range of molecules into valuable chemicals. Metal oxide nanoparticles (MO NPs) exhibit excellent photocatalytic activity and stability and are relatively inexpensive; however, their bandgap (>3.0 eV) significantly limits their visible-light-driven photocatalytic potential.[3,4]

During a photocatalytic NP reaction, the photocatalyst absorbs a photon and reaches a high-energy excited state. The absorbed energy enables electron (e_{CB}^-) transfer from the valence band to the conduction band, leaving holes (h_{VB}^+) in the valence band.[5] These h_{VB}^+ can (i) selectively oxidize useful molecules such as hydroxymethylfurfural (HMF) and benzyl alcohol (BA) and (ii) generate highly reactive hydroxyl radicals ($HO^\bullet$) in an aqueous solution. Similarly, e_{CB}^- can react with O_2 to produce superoxides ($HOO^{\bullet-}$), which are converted into H_2O_2 in an aqueous solution and ultimately generate $HO^\bullet$.[6] HMF conversion and BA oxidation have significant environmental, energy-related, and industrial implications. The conversion of HMF, a key biomass-derived platform chemical, into valuable biofuels, fine chemicals, and polymers is a sustainable solution for mitigating climate change by promoting a circular economy. The oxidation of BA to benzaldehyde and other fine chemicals is a common step during the production of fragrances, flavors, and active pharmaceutical ingredients in the cosmetics, food, and pharmaceutical industries, respectively. The photocatalytic conversion of HMF and BA addresses urgent environmental and industrial requirements and promotes greener, more sustainable chemical processes.

Among the various MO NPs, n-type semiconductor photocatalysts, such as tungsten oxide (WO_3) NPs, have garnered tremendous attention over the past few decades owing to their advantageous physicochemical properties, including low cost, chemical inertness, and high exciton binding energy (~60 MeV).[7–10] Pure WO_3 NPs exhibit a bandgap of ~3.3 eV, which achieves a limited light response and absorption efficiency in the visible spectrum; therefore, we must develop strategies for optimizing their visible-light-driven photocatalysis. Doping WO_3 NPs with precious metal ions such as Pd and Pt would optimize their bandgaps for visible-light-driven photocatalysis. Moreover, doping creates abundant oxygen vacancies (V_Os) on the WO_3 NP surfaces, which produce additional electronic

[a]*Department of Chemistry & Nanoscience, Ewha Womans University, Seoul, 03760, Republic of Korea*
[b]*Graduate Program in Innovative Biomaterials Convergence, Ewha Womans University, Seoul 03760, Korea*
[c]*Department of Chemistry, Sookmyung Women's University, Seoul 04310, Republic of Korea*
[d]*Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea*

†Electronic supplementary information (ESI) available: STEM-EDS, BET analysis, enlarged XRD pattern, LC-MS, reaction pathway, kinetics, recycle experiments. See DOI: https://doi.org/10.1039/d4dt03540b

‡These authors contributed equally to this work.

This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 03 March 2025. Downloaded on 3/25/2025 10:21:40 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

charge carriers and promote the oxidation of molecules on the surface.[11,12]

In this study, we doped WO_3 NPs with Pd and Pt (Pd@WO_3 and Pt@WO_3) and analyzed their morphology, chemical bonding structure, and electronic structure *via* high-resolution transmission electron microscopy (HRTEM), powder X-ray diffraction (PXRD), Raman spectroscopy, and X-ray photoelectron spectroscopy (XPS). Subsequently, we assessed their visible-light-driven photocatalytic performance during the oxidation of HMF and BA.[13]

Experimental

Preparation of WO_3 and NM@WO_3 NPs

To synthesize pristine WO_3 NPs, two precursor solutions, sodium tungstate dihydrate ($Na_2WO_4{\cdot}2H_2O$; 3.0 g) and HCl (0.1 M) in 100 mL of distilled deionized water (DDW), were prepared separately and gradually mixed with gentle stirring for 1 h until a yellow precursor was obtained.[14]

The mixed solution was then placed in an autoclave at 220 °C for 10 h. The produced WO_3 NPs were filtered and washed twice with DDW to remove any residues and dried at 90 °C for 48 h. After synthesizing the pristine WO_3 NPs, we obtained Pd@WO_3 and Pt@WO_3 NPs through the following procedure. The desired amounts of each noble metal ion source (5 mol% of $PdCl_2$ and $PtCl_2$) were added to the synthesized WO_3 NPs (1.0 g) and stirred continuously for 3 hours at 90 °C. The mixed solution was transferred to autoclaves and heated at 220 °C for 10 h in a convection oven. The produced NM@WO_3 NPs were filtered and washed five times with DDW to remove any residues and dried at 90 °C for 48 h. They were then washed 10 times with DDW and five times with ethanol, followed by drying at 95 °C. After synthesis, each sample was subjected to quantitative analysis of Pd and Pt using X-ray fluorescence (XRF), and the results indicated values close to the intended amounts, measuring 5.3% and 5.1%, respectively.

Characterization

The morphologies of the samples were observed using a JEM-ARM200F (JEM-ARM200F (NEOARM), JEOL Ltd, Japan) equipped with an energy-dispersive X-ray spectrometer. X-ray diffraction (XRD) patterns for pristine WO_3, Pd@WO_3, and Pt@WO_3 NPs were acquired using a D8 Advance TRIO/TWIN diffractometer (Bruker, USA) with Ni-filtered Cu Kα radiation at Sookmyung Women's University and a D/MAX-2200/VPC Rigaku at the National Research Facilities and Equipment Center (NanoBio·Energy Materials Center) at Ewha Womans University. Quantitative analysis of Pd and Pt on WO_3 was performed using an X-ray fluorescence spectrometer (Rigaku/ZSX-Primus IV) (Table S1†). Raman spectra were acquired by using a custom-made setup and were calibrated using the Si peak (520.89 cm^{-1}) as an internal reference. The light source used was a 532 nm laser with 1.02 mW power. Ultraviolet–visible diffuse reflectance spectroscopy (UV-Vis DRS) measurements were performed using a Shimadzu UV-2600i UV-Vis spectrophotometer equipped with integrating spheres. The nitrogen adsorption isotherms of the samples were collected at 77 K using a BELsorp-Max system (BEL) to measure the surface area of the three tested WO_3 NPs. To detect the presence of V_Os on the surface of the WO_3 NPs, electron paramagnetic resonance (EPR; JES-X320, JEOL Ltd) spectra, ranging from 300 to 400 mT, were obtained at 300 K at a frequency of 9446 MHz, a power of 10 mW, a modulation frequency of 100 kHz, and a width of 0.4 mT. To compare the electronic structures of the three tested samples, we performed XPS and work-function analysis at the 8A1 beamline in the Pohang Accelerator Laboratory. The W 4f, Pt 4f, Pd 3d, and O 1s core-level spectra were obtained using photon energies of 150, 200, 450, and 600 eV, respectively. The binding energies of the core-level electrons were compared to that of the clean Au 4f core level at the same photon energy.

Photocatalytic measurement

The photocatalytic conversion activities of WO_3, Pd@WO_3 and Pt@WO_3 NPs were determined using the oxidation of 25 mM 2,5-hydroxymethylfurfural (HMF, Sigma-Aldrich, 99%, 0.315 g) and 25 mM benzyl alcohol (BA Sigma-Aldrich, 99%, 0.270 g) in 100 mL of DDW.[15] No uncommon hazards are noted. The prepared pristine WO_3 and the two modified WO_3 NPs (0.015 g) were dispersed in distilled water (30 mL) under sonication. This aqueous suspension was stirred for 2 h to allow the equilibrium adsorption of HMF and BA on the three WO_3 NPs. A 300 W Xe arc lamp (Oriel) equipped with a 10 cm IR water filter and a cutoff filter (λ > 420 nm) was used as the light source. The photocatalytic conversion activities of HMF and BA were analyzed using high-performance liquid chromatography (HPLC-20AD pump, Shimadzu). The oxidation products of HMF and BA were analyzed *via* a HPLC-MS (ULTIMATE 3000 RSLC SYSTEM (Thermo) and Q-EXACTIVE ORBITRAP PLUS MS (Thermo)).

Results and discussion

The HRTEM images in Fig. 1 show the phase distributions and crystal structures of the three WO_3 NPs. The corresponding fast Fourier transform (FFT) and line profiles were simultaneously obtained at the atomic scale (inset of Fig. 1). Fig. 1a displays a low-magnification TEM image of the WO_3 NPs (yellow box). The corresponding FFT pattern indicates a single-crystal structure. The TEM image confirms a clear lattice fringe of 0.372 nm, which is consistent with the interplanar spacing of the (020) plane of WO_3. This indicates a highly crystalline hexagonal WO_3 phase (Fig. 1a).[16,17]

The interatomic distances on the surfaces of Pd@WO_3 and Pt@WO_3 NPs were determined from their FFT patterns (Fig. 1b and c, respectively).[18] Pd and Pt ions were observed in the crystalline structures on the WO_3 NP surfaces. The HRTEM image of Pd@WO_3 NPs (Fig. 1b, yellow box) shows two distinct lattice fringes, corresponding to the WO_3 NPs and crystallized PdO on the WO_3 NP surfaces. The (002) lattice fringe spacing

Open Access Article. Published on 03 March 2025. Downloaded on 3/25/2025 10:21:40 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

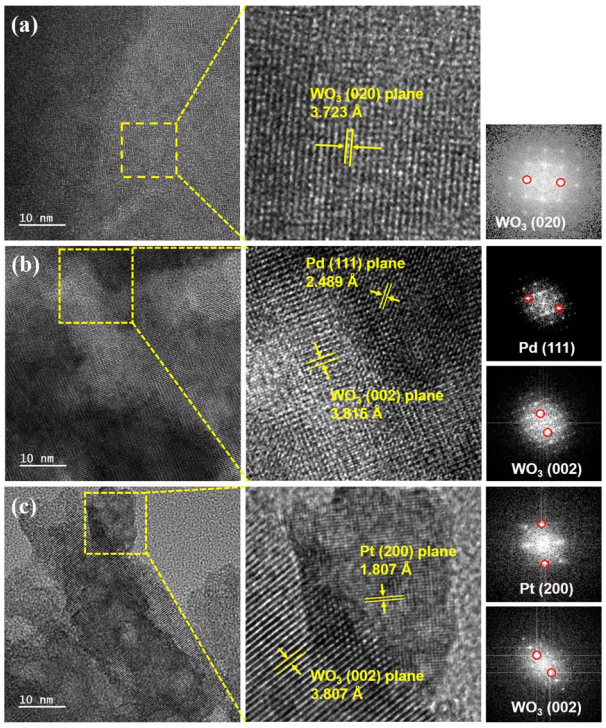


Fig. 1 High-resolution transmission electron microscopy images and the corresponding fast Fourier transform images of (a) WO_3, (b) 5 wt% $Pd@WO_3$, and (c) 5 wt% $Pt@WO_3$ nanoparticles (NPs).

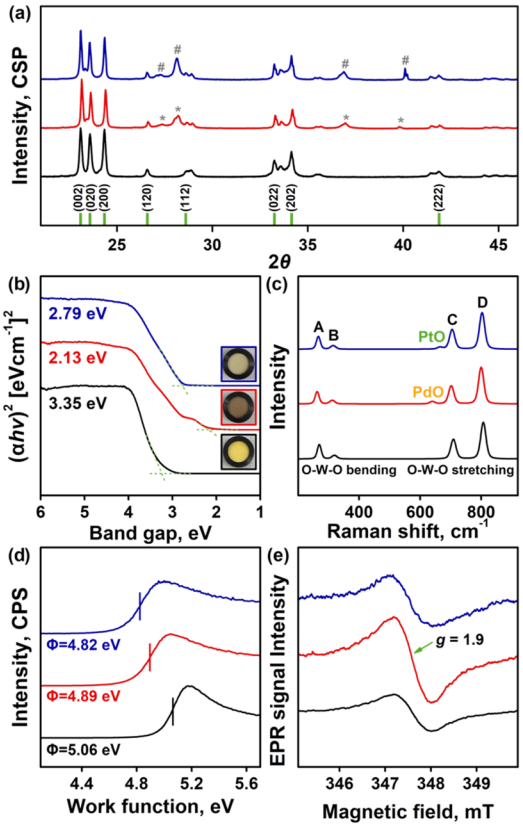


Fig. 2 (a) PXRD, (b) UV-Vis DRS spectra, (c) Raman spectra, (d) work function (ϕ) measurements, and (e) EPR data of pristine WO_3 (black), $Pd@WO_3$ (red), and $Pt@WO_3$ (navy blue).

of WO_3 was 0.382 nm, which is consistent with that of pristine WO_3 NPs. The (111) lattice fringe spacing of Pd was 0.249 nm, which is larger than the (111) interplanar distance in the face-centered cubic (fcc) Pd structure (0.225 nm). This indicates that the Pd ions were successfully doped into WO_3 in the form of PdO.[19] Similarly, the HRTEM image of $Pt@WO_3$ NPs shows two lattice fringes (Fig. 1c). The (002) lattice fringe spacing of WO_3 was identical (0.381 nm) to that of the pristine WO_3 NPs. The Pt (200) lattice fringe spacing was reduced to 0.181 nm, which is smaller than the (200) interplanar distance of the fcc Pt structure (0.19 nm).[20] These findings suggest that $Pt@WO_3$ and $Pd@WO_3$ NPs may exhibit different catalytic properties. The elemental distribution of the synthesized NPs was examined using energy-dispersive X-ray spectroscopy. Pd, Pt, W, and O were homogeneously distributed across the NPs (Fig. S1†).

The N_2 adsorption–desorption isotherms of the NPs were acquired to determine their Brunauer–Emmett–Teller (BET) surface areas.[21] The surface areas of WO_3, $Pd@WO_3$, and $Pt@WO_3$ NPs were determined to be 26.3, 35.9, and 32.1 m^2 g^{-1}, respectively. The doped NPs had higher surface areas (Fig. S2†). This could be attributed to the differences in their ionic radii (W^{6+}, 74 pm; Pt^{2+}, 94 pm; and Pd^{2+}, 100 pm), which likely cause surface defects on the WO_3 NP surfaces upon Pt and Pd doping.

The crystallinity and crystal planes of the three WO_3 NPs were analyzed *via* PXRD (Fig. 2a).[22] All samples exhibited characteristic 2θ peaks at 23.09°, 23.58°, 24.34°, 26.59°, 28.60°, 33.25°, 34.15°, and 41.88°, corresponding to the reflections from the (002), (020), (200), (120), (112), (022), (202), and (222) planes of WO_3, respectively. These prominent reflection patterns clearly indicate good crystallization (JCPDS No. 83-0950).[23] These patterns are consistent with the formation of monoclinic WO_3 (m-WO_3).[24] The lattice parameters of the unit cells of pristine WO_3 were as follows: a = 7.68, b = 7.72, c = 7.75, $\alpha = \gamma = 90°$, and β = 90.77°. However, the unit cells of $Pt@WO_3$ were slightly larger than those of pristine WO_3 (a = 8.35, b = 7.30, c = 7.85, $\alpha = \gamma = 90°$, and β = 106.1°). In contrast, the unit cells of $Pd@WO_3$ were slightly smaller than those of pristine WO_3 (a = 7.59, b = 7.53, c = 5.54, $\alpha = \gamma = 90°$, and β = 105.1°).[25] These results suggest that doping did not induce significant phase transitions. The new peaks in the PXRD patterns of $Pd@WO_3$ (marked *) and $Pt@WO_3$ (marked #) were attributed to PdO and PtO, respectively. However, the formation of these metal oxides on the WO_3 NP surfaces did not significantly alter the overall structure of WO_3.[26] The high-intensity peaks corresponding to WO_3 in the PXRD patterns of $Pd@WO_3$ and $Pt@WO_3$ NPs slightly shifted to larger angles owing to metal doping (Fig. S3†). This also confirms the presence of defects on the WO_3 surface (*i.e.*, a smaller unit cell).

The Scherrer equation (eqn (1)) was used to determine the average crystallite size of the NPs based on the peak with the

This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 03 March 2025. Downloaded on 3/25/2025 10:21:40 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

highest intensity (~22.9°), corresponding to the [002] reflection.

$$D = 0.9\lambda/B \cos\theta \quad (1)$$

where D is the crystallite size (nm), λ is the wavelength of the X-ray radiation (0.154 nm), θ is Bragg's angle, and B is the full width at half maximum (FWHM) of the 2θ peak. The nanoscale crystallite sizes are summarized in Table S2.† The crystallite sizes of the three NPs were in the range of 60.5–63.2 nm.[27]

The UV-visible diffuse reflectance spectra (UV-vis DRS) of the NPs were analyzed using the Kubelka–Munk function, which is proportional to the absorption coefficient. The bandgap (E_g) was determined using Tauc's plot in combination with the Kubelka–Munk function.[28] The band gap was then estimated using Tauc's equation (eqn (2)), where hn is the photon energy and A is a proportionality constant.

$$(F(R)hn)^2 = A(hn - E_g) \quad (2)$$

A plot of $(F(R)hn)^2$ *versus* hn was constructed, and the linear region was extrapolated to the energy axis to determine the band gap energy as shown in Fig. 2b. The calculated approximate bandgaps for WO_3, Pd@WO_3, and Pt@WO_3 NPs were 3.35, 2.13, and 2.79 eV, respectively.[29]

The vibrational frequencies of the W–O bonds in WO_3, Pd@WO_3, and Pt@WO_3 NPs were obtained through Raman spectroscopy (Fig. 2c). All three samples demonstrated W–O–W–O–W bending modes at 272 and 325 cm^{-1} (denoted as A and B), asymmetric O–W–O vibrations at 718 cm^{-1} (denoted as C), and symmetric O–W–O vibrations at 807 cm^{-1} (denoted as D).[30] These active vibration modes confirm the presence of a single, dominant WO_3 NP phase and are consistent with the XRD results. However, the peak positions of the Pd@WO_3 and Pt@WO_3 NPs downshifted compared to those of the WO_3 NPs.[31] This suggests that Pt and Pd doping lowered the W–O bond stretching energy, which confirms the formation of a defect structure on the surface of the WO_3 NPs.[32]

The charge transfer between WO_3 and the doped Pd and Pt ions in the defect structure was determined *via* work function measurements (Fig. 2d).[33] Doping WO_3 NPs with 5 wt% Pd and Pt ions shifted the secondary electron edge energies from 5.06 eV to 4.89 and 4.82 eV for Pd@WO_3 and Pd@WO_3, respectively. This negative shift in the work function (~0.2 eV) suggests a smoother charge carrier transfer, whether by light or electrons. This indicates the higher photocatalytic potential of the doped WO_3 NPs.[34] Electron paramagnetic resonance (EPR) spectroscopy was performed to confirm the increase in V_Os in the Pd@WO_3 and Pt@WO_3 NPs (Fig. 2e).[35] The EPR signal centered at g = 1.9 was attributed to the Zeeman effect induced by a single electron trapped within the anionic V_Os on the WO_3 NP surfaces. An EPR signal at g = 1.9 was also observed for Pt@WO_3 NPs. The spin quantification calculated by integrating this signal revealed that Pt@WO_3 had 1.3 times the V_Os of the WO_3 NPs.[36] Moreover, the Pd@WO_3 NPs exhibited an EPR signal at g = 1.9. The calculated spin quantification revealed that Pd@WO_3 NPs had 3 times the V_Os of the WO_3 NPs. These results confirm the abundance of V_Os in the Pt@WO_3 and Pd@WO_3 NPs compared to that in the WO_3 NPs.

The bonding configurations of the W and O atoms and the electronic structural changes induced by the doped Pd and Pt on the WO_3 NP surfaces were examined *via* XPS (Fig. 3).[37] The W 4f core-level XPS profile exhibited two W $4f_{7/2}$ peaks at 34.8 and 32.3 eV, corresponding to W^{6+} in WO_3 and W^{x+} in the WO_3 defects (WO_x), respectively (Fig. 3a).[38] The WO_x peak intensity was significantly higher for Pd@WO_3 and Pt@WO_3 NPs, which could be attributed to the abundant V_Os on their surfaces.[39] However, the Pd@WO_3 NPs had a higher number of defects than the Pt@WO_3 NPs. The O 1s core-level XPS profile showed three peaks at 530.5, 531.7, and 532.8 eV, corresponding to WO_3, V_Os, and –OH, respectively (Fig. 3b).[40] The peak intensity at 531.7 eV for Pd@WO_3 NPs was higher than those of Pt@WO_3 and WO_3 NPs. This confirms that the Pd@WO_3 NPs have a higher number of defects and V_Os than Pt@WO_3 NPs. These results suggest that Pd@WO_3 NPs could exhibit the best photocatalytic performance among the three WO_3 NPs.[41] The Pd 3d and Pt 4f core-level XPS profiles are shown in Fig. 3c and d, respectively. The peaks at 73.2 and 77.5 eV in the Pt XPS profile of Pt@WO_3 NPs correspond to Pt^{2+} in PtO, while those at 72.1 and 74.5 eV correspond to PtO_x formed because of V_Os. The peaks at 343.5 and 338.0 eV in the Pd XPS profile of Pd@WO_3 NPs correspond to Pd^{2+} in PdO, while those at 341.5 and 335.5 eV correspond to PdO_x formed because of the V_Os on the WO_3 NP surfaces. The formation of PtO and PdO on the surfaces of the WO_3 NPs reduced the number of O atoms on the WO_3 NP surfaces. However, a higher amount of PdO_x was formed on the Pd@WO_3 NPs than that of PtO_x on the Pt@WO_3 NPs (Fig. 3c and d). This indicates a higher number of V_Os on the Pd@WO_3 NPs than on the Pt@WO_3 NPs.[42] This suggests that the Pd@WO_3 NPs have the narrowest bandgap among the samples and could exhibit the highest photocatalytic performance.[43–45]

Fig. 4 shows the X-ray absorption near-edge structure (XANES) spectra and soft X-ray absorption spectra (XAS) of the

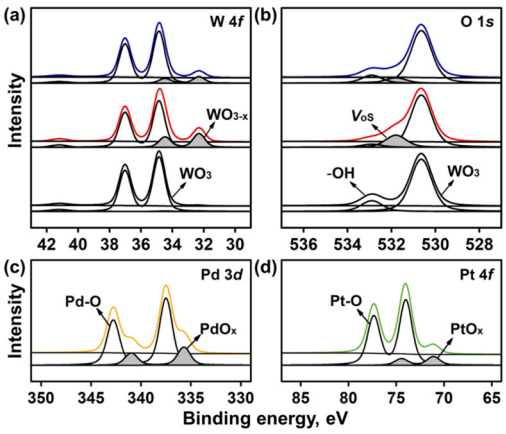


Fig. 3 (a) W 4f and (b) O 1s X-ray photoelectron spectra of WO_3 (black), Pd@WO_3 (red), and Pt@WO_3 NPs (navy blue). (c) Pd 3d and (d) Pt 4f core-level spectra.

Open Access Article. Published on 03 March 2025. Downloaded on 3/25/2025 10:21:40 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

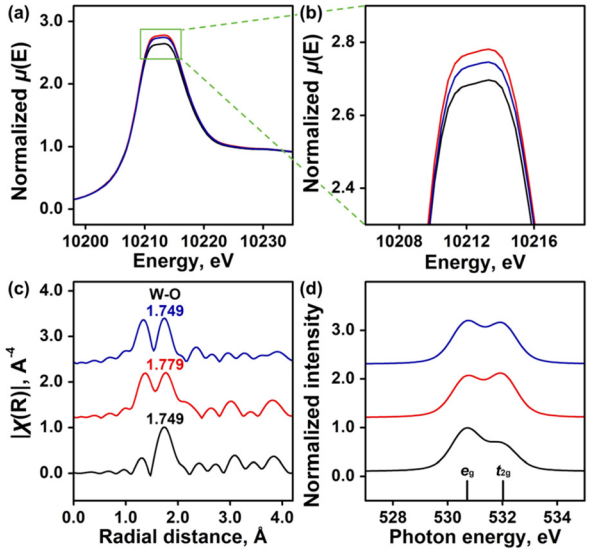


Fig. 4 (a) Overlaid W L_3-edge X-ray absorption near-edge structure (XANES), (b) magnified W L_3-edge XANES, (c) Fourier transform extended X-ray absorption fine structure (EXAFS), and (d) O K-edge spectra of WO_3 (black line), Pd@WO_3 (red), and Pt@WO_3 NPs (navy blue).

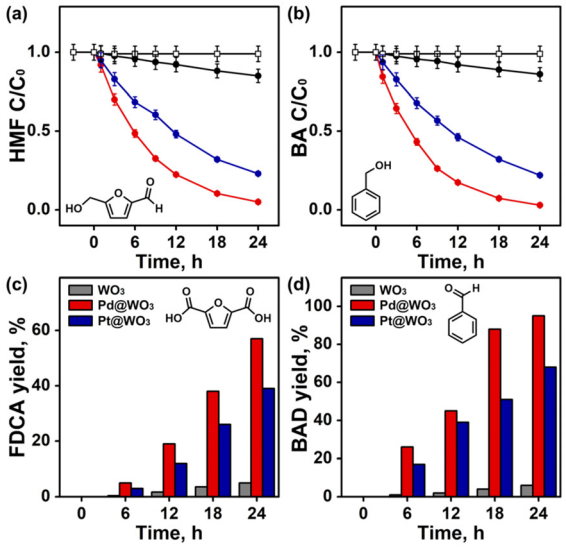


Fig. 5 Photocatalytic conversion activities of pristine WO_3, Pd@WO_3, and Pt@WO_3 NPs during the conversion of (a) hydroxymethylfurfural (HMF) and (b) benzyl alcohol (BA) with and without 5,5-dimethyl-1-pyrroline *N*-oxide (DMPO; gray). Reaction conditions: [catalyst] = 0.5 g L^{-1}, $\lambda \geq 420$ nm, and $[HMF]_0 = [BA]_0 = 25$ mM. Yields of (c) 2,5-furandicarboxylic acid (FDCA) from HMF and (d) benzaldehyde (BAD) from BA with increasing time of 420 nm irradiation.

three WO_3 NPs. W L_3-edge XANES was used to study the electronic transitions from the $2p_{3/2}$ orbitals to the vacant d orbitals of the WO_3 NPs.[46,47] All NPs exhibited a rising edge peak at ~10 212.5 eV, suggesting that the W^{6+} ions in all three NPs are the dominant metal ions (Fig. 4a). However, higher peak intensities were observed for the Pd@WO_3 and Pt@WO_3 NPs, indicating an abundance of V_Os (Fig. 4b). This difference in the intensities suggested that the amount of V_Os increased in the order of WO_3 < Pt@WO_3 < Pd@WO_3 NPs, which was consistent with the spectroscopic results described above. Fig. 4c exhibits the Fourier transform (FT) of the W L_3-edge EXAFS $k^3\chi$ data of the NPs in the range of k = 3–11 $Å^{-1}$.[48] The peak at 1.749–1.779 Å (phase-corrected values) corresponds to the average bond lengths of the W–O coordination shells in the WO_6 octahedron. Very interestingly, W–O bonds within Pt@WO_3 NPs remained identical to those in the pristine WO_3 NPs, while Pd doping noticeably affected the first W–O coordination shell, leading to an increase of 0.03 Å. Fig. 4d shows the O K-edge XAS spectra of the three WO_3 NPs. It was found that the peak intensity originating from the e_g orbital decreased upon doping with Pt^{2+} and Pd^{2+} ions, while the t_{2g} signal intensity significantly increased. This result suggested changes in the electronic structures of the unoccupied states within WO_3 NPs and, thereby, alterations in their defect structures.[49] The increase in t_{2g} signal intensity accompanied by a decrease in e_g signal intensity indicated that the changes in the absorption intensity ratio between e_g and t_{2g} orbitals generated additional defect structures in WO_3 NPs through a mixing of tetrahedral and octahedral geometries.

Fig. 5 shows the photocatalytic conversion efficiencies of HMF and BA in the presence of the three WO_3 NPs under visible-light irradiation and the yields of 2,5-furandicarboxylic acid (FDCA) and benzaldehyde (BAD) from HMF and BA, respectively.[50,51] When the radical scavenger 5,5-dimethyl-1-pyrroline *N*-oxide (DMPO) was used, HMF and BA were not converted into the desired products under visible-light irradiation, indicating that the radical scavenger consumed the generated radical species. Moreover, the WO_3 NPs did not convert HMF or BA, presumably because of their large bandgap. It is also noteworthy that there are no reported cases of HMF and BA conversion using other transition metal ion-doped WO_3 (Table S3†). While some transition metals may exhibit similar defect-engineering properties, Pd and Pt were chosen due to their superior stability and catalytic performance under visible light.

However, the Pd@WO_3 and Pt@WO_3 NPs showed higher photocatalytic degradation (PCD) of HMF and BA under visible-light irradiation compared to the WO_3 NPs. The Pd@WO_3 NPs exhibited the highest PCD efficiency, which is consistent with the XPS and DRS results. This enhanced PCD efficiency is attributed to the extensive defect structures on the Pd@WO_3 NP surfaces. Table 1 summarizes the 24 h PCD profiles of HMF and BA for the three WO_3 NPs under 420 nm irradiation.

The yields of FDCA (Fig. 5c) are consistent with the PCD efficiency of HMF (Fig. 5a). The Pd@WO_3 NPs demonstrated the highest FDCA yield. As the selective oxidation of the –OH or –CHO groups of HMF to produce FDCA is a key photocatalytic conversion reaction, we must identify the functional groups that participate in the oxidation of HMF in the presence of WO_3, Pd@WO_3, and Pt@WO_3 NPs (Fig. S4†). In all three biomass conversion reactions, the –OH group of HMF

This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 03 March 2025. Downloaded on 3/25/2025 10:21:40 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 1 Photocatalytic degradation (PCD, C/C_0 [a]) efficiencies of HMF and BA and yields of FDCA and BAD for WO_3, Pd@WO_3, and Pt@WO_3 NPs after 24 h of 420 nm irradiation

	WO_3	Pd@WO_3	Pt@WO_3
C/C_0 of HMF	0.85 ± 0.042	0.050 ± 0.003	0.23 ± 0.012
C/C_0 of BA	0.86 ± 0.003	0.030 ± 0.002	0.22 ± 0.011
FDCA (%)	5.0	57	39
BAD (%)	6.0	95	68

[a] C and C_0 are the time-dependent and initial concentrations of HMF and BA.

was selectively oxidized to –CHO, affording 2,5-diformylfuran (DFF) as the sole intermediate, which was further oxidized to FDCA, regardless of the variations in the ratio of the two species (Scheme 1).[52]

We analyzed the liquid chromatography-mass spectrometry profiles of the photocatalytic biomass conversion reactions on the three WO_3 NPs and derived a plausible catalytic mechanism. The V_Os on the WO_3 NP surfaces trap electrons, which react with the O_2 present in air-saturated water, affording $O_2^{\bullet-}$ species. Radical trapping experiments with DMPO confirmed the generation of $O_2^{\bullet-}$ species (gray circles in Fig. 5a and b). These $O_2^{\bullet-}$ species oxidize HMF to DFF and then to FDCA with the assistance of h^+. Both Pd@WO_3 and Pt@WO_3 NPs showed enhanced oxidation efficiencies for the conversion of BA to BAD (Fig. 5d). This is attributed to the higher number of defects on their surfaces.[53] Table 1 summarizes the yields of FDCA and BAD from 25 mM HMF and BA, respectively, over 24 h obtained using the three WO_3 NPs under visible-light irradiation.

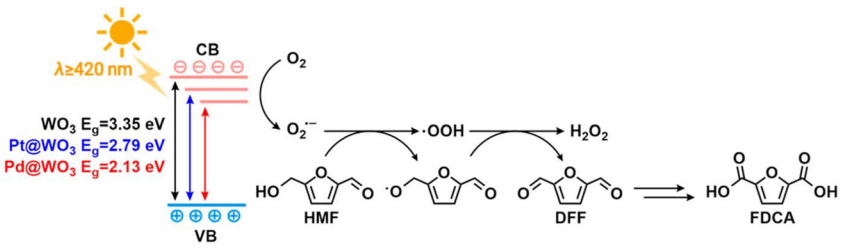
We observed a linear correlation between $\ln(C/C_0)$ and the reaction time for the three WO_3 NPs, indicating pseudo-first-order kinetics. The calculated first-order rate constants (k) (Fig. S5†) for the HMF PCD for pristine WO_3, Pd@WO_3, and Pt@WO_3 NPs were 0.0067, 0.13, and 0.061 h^{-1}, respectively. The k values of the BA PCD for pristine WO_3, Pd@WO_3, and Pt@WO_3 NPs were 0.0063, 0.15, and 0.063 h^{-1}, respectively. This indicates that the reaction rates of Pd@WO_3 NPs are over 17 (for PCD of HMF) and 22 (for PCD of BA) times higher than those of the WO_3 NPs. The photocatalytic activity of the three NPs was sustained (>97%) even after ten cycles (Fig. S6†).

The improved photocatalytic activity of WO_3 NPs under visible-light irradiation can be partially attributed to the narrow bandgap of Pd@WO_3 and Pt@WO_3 NPs, along with an increase in V_Os. The Pd@WO_3 NPs exhibited higher reactivity than the Pt@WO_3 NPs in the photocatalytic HMF conversion owing to their higher density of V_Os. In typical photocatalysts, charge carriers are generated near the surface due to exciton formation and subsequent charge separation. Owing to the short carrier diffusion length, surface defects often serve as trap states that enhance the photocatalytic reactivity, depending on the density of V_Os and the associated bandgap changes (Scheme 1).

The different valence states of Pd^{2+} and Pt^{2+} compared to W^{6+} in the host WO_3 facilitate the generation of V_Os through efficient charge transfer,[54] enhancing photocatalytic oxidation.[55] The Shannon effective ionic radii of W^{6+}, Pd^{2+}, and Pt^{2+} were 74 pm, 94 pm, and 100 pm, respectively. This suggests that cation doping, especially with Pd^{2+}, induces lattice strain in the host WO_3 NPs, acts as a reducing agent, and generates atomic defects, altering their electronic structure.[56,57] In addition, the different valence states of the doped cations compared to W^{6+} in the host WO_3 indicate the crucial role of V_Os in the photocatalytic activity of the doped WO_3 NPs. Overall, the Pd@WO_3 NPs showed the highest photocatalytic activity.

Conclusions

WO_3 NPs intrinsically exhibit very weak catalytic activity under visible light. In this study, we successfully engineered WO_3 NPs by doping with Pd and Pt ions. Atomic-scale analysis confirmed the generation of oxygen vacancies V_Os. Despite structural similarity in bulk phases, these defects, induced by local asymmetries in the electron and hole concentrations, lowered the semiconductor bandgap and dramatically enhanced photocatalytic conversion. The Pd@WO_3 and Pt@WO_3 NPs efficiently converted HMF to FDCA and BA to BAD in neutral water as the formal oxidant under visible light. In particular, the Pd@WO_3 NPs exhibited >95% PCD rates for HMF and BA. The yields of FDCA and BAD were 57% and 95%, respectively. This study highlights the crucial role of V_Os and local electronic asymmetries in tuning metal oxide photocatalysts. Future efforts should focus on optimizing the dopant concentration and exploring alternative transition metal dopants to further enhance catalytic efficiency. The development of high-performance, visible-light-driven photocatalysts will advance the utilization of solar energy.



Scheme 1 Illustration of the distinct reaction and charge transport mechanisms in photocatalysis by WO_3, Pd@WO_3, and Pt@WO_3 NPs.

Open Access Article. Published on 03 March 2025. Downloaded on 3/25/2025 10:21:40 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Author contributions

Hyeri Jeon, Dung Thanh Hoang, and Hyejin Yu: formal analysis, investigation, visualization, and writing – review & editing. Hyun Sung Kim, Hangil Lee, and Seungwoo Hong: conceptualization, formal analysis, funding acquisition, project administration, writing – original draft, and writing – review & editing.

Data availability

The data supporting this article have been included as part of the ESI.†

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (RS-2024-00346153). This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (2021R1A6A1A10039823 for S. Hong and RS-2023-00272031 for H. Jeon). H. S. Kim acknowledges the National Research Foundation of Korea (NRF) grant, funded by the Korean government (MSIT) (NRF-2022R1A6A1A03051158).

References

1 W. M. Cheng and R. Shang, *ACS Catal.*, 2020, **10**, 9170–9196.
2 X. Lang, X. Chen and J. Zhao, *Chem. Soc. Rev.*, 2014, **43**, 473–486.
3 Y. Yang, S. Niu, D. Han, T. Liu, G. Wang and Y. Li, *Adv. Energy Mater.*, 2017, **7**, 1700555.
4 M. Wang, Z. Hou, A. A. Al Kheraif, B. Xing and J. Lin, *Adv. Healthcare Mater.*, 2018, **7**, 1800351.
5 C. B. Ong, L. Y. Ng and A. W. Mohammad, *Renewable Sustainable Energy Rev.*, 2018, **81**, 536–551.
6 D. Tang, G. Lu, Z. Shen, Y. Hu, L. Yao, B. Li, G. Zhao, B. Peng and X. Huang, *J. Energy Chem.*, 2023, **77**, 80–118.
7 P. Shandilya, S. Sambyal, R. Sharma, P. Mandyal and B. Fang, *J. Hazard. Mater.*, 2012, **428**, 128218.
8 C. M. Wu, S. Naseem, M. H. Chou, J. H. Wang and Y. Q. Jian, *Front. Mater.*, 2019, **6**, 49.
9 G. Longobucco, L. Pasti, A. Molinari, N. Marchetti, S. Caramori, V. Cristino, R. Boaretto and C. A. Bignozzi, *Appl. Catal., B*, 2017, **204**, 273–282.
10 C. Zhai, M. Zhu, D. Bin, H. Wang, Y. Du, C. Wang and P. Yang, *ACS Appl. Mater. Interfaces*, 2014, **6**, 17753–17761.
11 S. Sun, G. Li, S. Zhu, W. Meng, L. Xu, J. Jiang, F. Wang and X. Li, *J. Mater. Chem. A*, 2024, **12**, 17463–17470.
12 W. Meng, S. Sun, D. Xie, S. Dai, W. Shao, Q. Zhang, C. Qin, G. Liang and X. Li, *Mol. Catal.*, 2024, **553**, 113768.
13 C. R. Lhermitte, N. Plainpan, P. Canjura, F. Boudoire and K. Sivula, *RSC Adv.*, 2021, **11**, 198–202.
14 H. Jeon, J. H. Park, S. Han, S. H. Ahn, J. Baik, H. Lee, H. S. Ahn and S. Hong, *Appl. Surf. Sci.*, 2021, **567**, 150834.
15 H. Jeon, D. T. Hoang, J. Baik, S. Hong and H. Lee, *Inorg. Chem.*, 2024, **63**, 12370–12376.
16 K. Sato, H. Yamashita and Y. Kojima, *Jpn. J. Appl. Phys.*, 2024, **63**, 04SP76.
17 Y. Li, X. Wei, X. Yan, J. Cai, A. Zhou, M. Yang and K. Liu, *Phys. Chem. Chem. Phys.*, 2016, **18**, 10255–10261.
18 S. H. Cho, J. M. Suh, B. Jeong, T. H. Lee, K. S. Choi, T. H. Eom, S. W. Choi, G. B. Nam, Y. J. Kim and H. W. Jang, *Small*, 2024, **20**, 2309744.
19 Z. Xi, D. P. Erdosy, A. Mendoza-Garcia, P. N. Duchesne, J. Li, M. Muzzio, Q. Li, P. Zhang and S. Sun, *Nano Lett.*, 2017, **17**, 2727–2731.
20 C. A. García-Negrete, T. C. Rojas, B. R. Knappett, D. A. Jefferson, A. E. H. Wheatley and A. Fernándeza, *Nanoscale*, 2014, **6**, 11090.
21 Z. Han, J. Ren, J. Zhou, S. Zhang, Z. Zhang, L. Yang and C. Yin, *Int. J. Hydrogen Energy*, 2020, **45**, 7223–7233.
22 T. Thilagavathi, D. Venugopal, R. Marnadu, J. Chandrasekaran, D. Thangaraju, B. Palanivel, M. S. Hamdy, M. Shkir and H. E. Ali, *J. Phys. Chem. Solids*, 2021, **154**, 110066.
23 B. Liu, D. Cai, Y. Liu, D. Wang, L. Wang, Y. Wang, H. Li, Q. Li and T. Wang, *Sens. Actuators, B*, 2014, **193**, 28–34.
24 C. V. Ramana, S. Utsunomiya, R. C. Ewing, C. M. Julien and U. Becker, *J. Phys. Chem. B*, 2006, **110**, 10430–10435.
25 M. Saleem, J. Iqbal, A. Nawaz, B. Islam and I. Hussain, *Int. J. Appl. Ceram. Technol.*, 2020, **17**, 1918–1929.
26 T. N. Anh, N. T. Hien, D. T. H. Linh, N. T. Hanh, L. T. Do, N. H. N. M. Hoang, D. V. Quang and V. D. Dao, *Inorg. Chem. Commun.*, 2024, **161**, 112100.
27 P. M. Kodam, P. A. Ghadage, D. Y. Nadargi, K. P. Shinde, I. S. Mulla and J. S. Park, *Ceram. Int.*, 2022, **48**, 17923–17933.
28 J. Zhang, H. Liu, Y. Wang and X. Zhao, *J. Phys. Chem. Lett.*, 2018, **9**, 6814–6817.
29 R. J. Bose, N. Illyasukutty, K. S. Tan, R. S. Rawat, M. V. Matham, H. Kohler and V. M. Pillai, *Appl. Surf. Sci.*, 2018, **440**, 320–330.
30 F. Mehmood, J. Iqbal, T. Jan and Q. Mansoor, *J. Alloys Compd.*, 2017, **728**, 1329–1337.
31 C. Xie, W. Chen, S. Du, D. Yan, Y. Zhang, J. Chen, B. Liu and S. Wang, *Nano Energy*, 2020, **71**, 104653.
32 J. Kaur, K. Anand, N. Kohli, A. Kaur and R. C. Singh, *Chem. Phys. Lett.*, 2018, **701**, 115–125.
33 P. Mandyal, R. Sharma, S. Sambyal, N. Islam, A. Priye, M. Kumar, V. Chauhan and P. Shandilya, *J. Water Process Eng.*, 2024, **59**, 105008.
34 S. S. Kalanur, R. Singh and H. Seo, *Appl. Catal., B*, 2021, **295**, 120269.

 This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 03 March 2025. Downloaded on 3/25/2025 10:21:40 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

35 K. Wang, L. Luo, C. Wang and J. Tang, *Chin. J. Catal.*, 2023, **46**, 103–112.
36 Y. Niu, B. Zhao, Y. Liang, L. Liu and J. Dong, *Ind. Eng. Chem. Res.*, 2020, **59**, 7389–7397.
37 S. Fardindoost, F. Rahimi and R. Ghasempour, *Int. J. Hydrogen Energy*, 2010, **35**, 854–860.
38 Y. Zhu, C. Blackman, P. Zhou, S. K. Ayyala, J. A. Covington, Y. Shen, J. Liang, X. Zhong, C. Knapp and Y. Zhou, *J. Alloys Compd.*, 2023, **936**, 167930.
39 B. Li, X. Zhao, X. H. Luo, W. X. Zhang, C. H. Wen, L. Y. Xu, C. Tang, M. F. Luo and J. Chen, *Appl. Surf. Sci.*, 2024, **650**, 159225.
40 E. Gioria, S. Li, A. Mazheika, R. N. d'Alnoncourt, A. Thomas and F. Rosowski, *Angew. Chem., Int. Ed.*, 2023, **62**, e202217888.
41 S. I. Han, M. Kumar, R. Yeasmin, C. Park, G. Jung, H. Kim, A. S. Khan, H. Dang and H. Seo, *Sens. Actuators, B*, 2024, **404**, 135259.
42 X. Liu, J. Han, X. Qiao, H. Cai, Y. Zhao, Z. Zhang, B. Zhai, T. Ni, C. Zhao and Y. Zhu, *ACS Appl. Mater. Interfaces*, 2024, **16**, 22155–22165.
43 X. Wei, X. He, P. Wu, F. Gong, D. Wang, S. Wang, S. Lu, J. Zhang, S. Xiang, T. Kai and P. Ding, *Int. J. Hydrogen Energy*, 2021, **46**, 27974–27996.
44 T. Li, Y. Shen, X. Zhong, S. Zhao, G. Li, B. Cui, D. Wei and K. Wei, *J. Alloys Compd.*, 2020, **818**, 152927.
45 F. Naaz and T. Ahmad, *Langmuir*, 2023, **39**, 9300–9314.
46 M. Fernández-García, A. Martínez-Arias, A. Fuerte and J. C. Conesa, *J. Phys. Chem. B*, 2005, **109**, 6075–6083.
47 S. Yamazoe, Y. Hitomi, T. Shishido and T. Tanaka, *J. Phys. Chem. C*, 2008, **112**, 6869–6879.
48 B. Ravel and M. J. Newville, *J. Synchrotron Radiat.*, 2005, **12**, 537–541.
49 F. Frati, M. O. J. Y. Hunault and F. M. F. de Groot, *Chem. Rev.*, 2020, **120**, 4056–4110.
50 A. Kumar and R. Ananthakrishnan, *ACS Appl. Energy Mater.*, 2022, **5**, 2706–2719.
51 H. Zhang, Z. Feng, Y. Zhu, Y. Wu and T. Wu, *J. Photochem. Photobiol., A*, 2019, **371**, 1–9.
52 F. Neaţu, R. S. Marin, M. Florea, N. Petrea, O. D. Pavel and V. I. Pârvulescu, *Appl. Catal., B*, 2016, **180**, 751–757.
53 Y. Su, Z. Han, L. Zhang, W. Wang, M. Duan, X. Li, Y. Zheng, Y. Wang and X. Lei, *Appl. Catal., B*, 2017, **217**, 108–114.
54 H. Liu, K. Tian, J. Ning, Y. Zhong, Z. Zhang and Y. Hu, *ACS Catal.*, 2019, **9**, 1211–1219.
55 B. Lei, W. Cui, P. Chen, L. Chen, J. Li and F. Dong, *ACS Catal.*, 2022, **12**, 9670–9678.
56 B. D. Mukri, U. V. Waghmare and M. S. Hegde, *Chem. Mater.*, 2013, **25**, 3822–3833.
57 R. Iimura, T. Hasegawa and S. Yin, *Inorg. Chem.*, 2022, **61**, 2509–2516.

Applied Surface Science 686 (2025) 162088

Contents lists available at ScienceDirect

Applied Surface Science

journal homepage: www.elsevier.com/locate/apsusc

Full Length Article

Enhanced photocatalytic and electrocatalytic properties of IrO_2 nanoparticles via Cr and Co ion doping: Insights into surface oxygen defect structures

Hyeri Jeon [a,1], Dung Thanh Hoang [b,1], Gyuri Kim [a], In Young Kim [a,*], Hangil Lee [b,*], Seungwoo Hong [a,*]

[a] Department of Chemistry & Nanoscience, Ewha Womans University, Seoul 03760, Korea
[b] Department of Chemistry, Sookmyung Women's University, Seoul 04310, Republic of Korea

ARTICLE INFO

Keywords:
Transition metal-doped IrO_2
Oxygen vacancy
Photocatalysis
Electrocatalysis

ABSTRACT

The growing global energy demand underscores the urgent need for cost-effective visible light-activated photocatalysts capable of efficiently converting diverse molecules into energy-relevant compounds. In parallel, advancements in catalyst synthesis methodologies have intensified research efforts aimed at developing multifunctional catalysts that synergistically integrate photocatalytic and electrocatalytic functionalities. A prevailing strategy to enhance catalytic efficiency in nanoparticles (NPs) focuses on increasing the concentration of active sites, particularly oxygen vacancies, which are critical for improving both photocatalytic and electrocatalytic performance.

To address this challenge, IrO_2 nanoparticles were employed as electrocatalysts and doped with transition metal ions, specifically Cr and Co, to systematically investigate their dual photocatalytic and electrocatalytic properties. Structural and electronic characterizations confirmed that the doping process effectively induced the formation of oxygen vacancies. Among the doped systems, Cr-doped IrO_2 (Cr@IrO_2) NPs exhibited superior photocatalytic activity, which was attributed to significant improvements in visible light absorption compared to both undoped IrO_2 and Co-doped IrO_2 NPs. This enhancement is primarily ascribed to the incorporation of Cr ions on the IrO_2 surface, which introduced a higher density of oxygen vacancies due to charge mismatch phenomena, thereby amplifying visible light absorption intensity and catalytic efficiency.

1. Introduction

Amidst rising energy needs and diminishing fossil fuel stores, there is a mounting enthusiasm for tapping into solar energy through photocatalysis as an encouraging green approach poised to play a vital role in our imminent transition to renewable energy. [1–3] Notably, visible light, constituting the bulk of sunlight on the Earth, stands out as a sustainable, pristine, and boundless asset, driving ongoing research; this clearly aims at transforming diverse molecules into valuable commodities. For example, Chen and co-workers investigated photocatalytic hydrogen evolution enhanced by oxygen vacancies and defect engineering under visible light. [4–9] Hence, developing a photocatalyst that is exceptionally efficient, long-lasting, and affordable stands as a paramount objective in modern green chemistry. [10,11].

Iridium oxide nanoparticles (IrO_2 NPs), recognized for their unparalleled electrocatalytic prowess for water oxidation also known as oxygen evolution reaction owing to exceptional activity and stability, display a band gap of approximately 1.4 eV. [12,13] We aim to take advantage of this semiconducting characteristic to develop a resilient and high-efficiency visible light photo- and electrocatalyst. [14,15] Over the past few decades, multifunctional photo- and electrocatalysis has surfaced as a compelling approach in solar-to-chemical conversion, leveraging its distinctive ability to integrate electrocatalysis and photocatalysis for enhanced solar energy capture, charge dynamics, and catalytic processes. Yet, challenges persist, such as limited reaction sites and low product selectivity, hindering its advancement toward viable applications. [16,17].

Given that defects in solid materials effectively introduce

* Corresponding authors.
E-mail addresses: easyscan@sookmyung.ac.kr (I.Y. Kim), hsw@ewha.ac.kr (S. Hong).
[1] These authors contributed equally.

https://doi.org/10.1016/j.apsusc.2024.162088
Received 2 September 2024; Received in revised form 6 December 2024; Accepted 11 December 2024
Available online 12 December 2024
0169-4332/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

coordinately unsaturated sites, which provide significant number of highly active sites for molecular chemisorption and activation, enhancing the defect structure on the surface of IrO_2 NPs through rational defect engineering holds promising route for creating a multi-functional catalyst. This approach will combine the strengths of established electrocatalyst (i.e., IrO_2 NPs) with the potential for photocatalytic applications (i.e., defects on the surface of IrO_2 NPs). [18,19] One of the synthetic strategies to create defects and enhance quantum efficiency of photocatalysts is by doping with foreign elements such as transition metals and introducing defects such as oxygen vacancies (V_os). [20,21] It is widely accepted that the recombination of photoexcited electrons and holes significantly hampers photocatalytic activity. Concurrently, lattice defects (e.g., V_os) can modulate optical properties through hybridization electron distributions, thereby improving photogenerated carrier separation and prolonging light carrier lifetime. Moreover, the oxygen vacancy serves as an electron trap within the photocenter, enhancing electron redirection to the catalyst surface and thereby suppressing the recombination of electron-hole pairs. [22,23] Hence, optimizing the concentration of oxygen vacancies in IrO_2 nanoparticles by doping is an important key to creating active atomic sites that actively engage in the photocatalytic process. The formation of oxygen vacancies in doped IrO_2 nanoparticles hinges on the effective charge transfer between the doped transition metal ions and the host IrO_2 NPs, contingent upon their respective valence states. [24].

In this regard, two transition metal ions with different valence states (Cr: Cr^{3+} and Co: Co^{2+}) were selected as dopants to investigate the relationship between the structure and electronic properties of doped IrO_2 NPs. Cr@IrO_2 and Co@IrO_2 NPs were synthesized and characterized various spectroscopic techniques. The differences in morphology, chemical bonding structure and electronic structure were analyzed using a variety of spectroscopic methods, including scanning electron microscopy (SEM), powder X-ray diffraction (PXRD), Raman spectroscopy and high-resolution X-ray photoelectron spectroscopy (HRXPS). Their visible light active photo- and electrocatalytic activities were compared with the pristine IrO_2 NPs in the photoconversion of biomass and electro-oxidation reactions of cysteine. Their varying reactivity was rationalized through the structural analysis combined with the consideration of charge valence of the doped metal ions.

2. Experimental

2.1. Materials

Potassium hexachloroiridate (K_2IrCl_6; 99.99 % CAS-No 16920–56-2), the two dopant precursors ($Cr(NO_3)_3{\cdot}9H_2O$; 99.99 % CAS-No 7789–02-8 and $Co(NO_3)_2{\cdot}6H_2O$; 99.9 % CAS-No 10026–22-9), and sodium hydroxide (NaOH; 98.0 % CAS-No 1310–73-2) were purchased from Sigma-Aldrich Co. and used without further purification unless otherwise indicated. All aqueous solutions were prepared using double distilled water (DDW).

2.2. Preparation of IrO_2, Cr@IrO_2, Co@IrO_2 NPs

The IrO_2 NP solution was prepared by a modification of the Wohler method. [25] K_2IrCl_6 (150 mg, 0.3 mmol) was dissolved in water in a flask to make 100 mL of the corresponding aqueous solution, to which 1.5 mL of an aqueous NaOH solution was added. The flask was sealed with a foil, the solution in the flask was mixed and immediately afterwards the foil seal was removed and the flask was placed in a 95℃ oil bath for 1 h. The resulting solution was then cooled to room temperature in an ice bath and stored in a capped flask for over 24 h. Upon heating, the initially dark green solution became increasingly transparent, but then dark blue at 95℃, as desired. An increasing number of particles could be seen over time after heating. When stored at room temperature, the upper layers of the heated solution separated from the upper level of the upper stream, and sediment eventually formed over this separation. However, when heated to 95℃, no such sediment was found. To prepare 5 wt% doped Cr@IrO_2 and Cr@IrO_2 NPs with different metal contents, the desired amount of dopant ($Cr(NO_3)_3{\cdot}nH_2O$ and $Co(NO_3)_2{\cdot}nH_2O$) was determined by mole fraction calculation and then added to a solution of IrO_2 NPs under continuous stirring. The mixed solution was transferred to a Teflon-lined autoclave and then placed in a convection oven preheated at 220℃ for 9 h. The product was washed several times with copious amounts of DDW and collected by precipitation.

2.3. Characterizations

X-ray diffraction (XRD) patterns for the pristine IrO_2, Cr@IrO_2, and Co@IrO_2 NPs were acquired using a D8 Advance TRIO/TWIN diffractometer (Bruker, USA) with Ni-filtered Cu Kα radiation at Sookmyung Women's University and D/MAX-2200/VPC Rigaku at the National Research Facilities and Equipment Center (NanoBio•Energy Materials Center) at Ewha Womans University. The morphologies of the samples were characterized by using a field-emission scanning electron microscope (FE-SEM, FEI Inspect F50) operating at an acceleration voltage of 15 kV. The nitrogen adsorption isotherms of the samples were collected at 77 K using a BELsorp-Max system (BEL) to measure the surface area of the three tested IrO_2 NPs. To detect the existence of V_os on the surface of the IrO_2 NPs, electron paramagnetic resonance (EPR; JES-X320, JEOL Ltd.) spectra, ranging from 300 to 400 mT, were obtained at 300 K at a frequency of 9446 MHz, power of 10 mW, the modulation frequency of 100 kHz, and a width of 0.4 mT. To compare the electronic structures of the three tested samples, we conducted XPS and work-function analysis at the 8A1 beamline in the Pohang Accelerator Laboratory. The Ir 4*f* and O 1 *s* core-level spectra were obtained using photon energies of 150 and 600 eV, respectively. The binding energies of the core-level electrons were compared to that of the clean Au 4*f* core-level at the same photon energy. We also acquired Cr *L*-edge and Co *L*-edge X-ray absorption spectra (XAS) at the 8A1 beamline in the Pohang Accelerator Laboratory. Raman spectra were obtained by using a spectrometer (Horiba, ARAMIS) with an Ar^+ ion CW (514.5-nm-wavelength) laser. Ultraviolet–visible (UV–Vis) diffuse reflectance measurements were performed using a Shimadzu UV-2600i UV–Vis spectrophotometer equipped with integrated spheres.

2.4. Photocatalytic measurement

The photocatalytic conversion activities of the pristine IrO_2, Cr@IrO_2, and Co@ IrO_2 NPs were traced using the oxidation of 2,5-hydroxymethylfurfural (HMF Sigma-Aldrich, 99 %) and benzyl alcohol (BA Sigma-Aldrich, 99 %). No uncommon hazards are noted. The prepared three IrO_2 NPs (0.015 g) were dispersed in distilled water (30 mL) under sonication. This aqueous suspension was stirred for 2 h to allow the equilibrium adsorption of HMF and BA on the three IrO_2 NPs. A 300 W Xe arc lamp (Oriel) equipped with a 10 cm IR water filter and cutoff filter ($\lambda \geq 420$ nm) was used as the light source. The photoconversion activities of HMF and BA were analyzed using high-performance liquid chromatography (LC-20AD pump, Shimadzu). The oxidation products of HMF and BA were analyzed via HPLC-MS (ULTIMATE 3000 RSLC SYSTEM (Thermo) and Q-EXACTIVE ORBITRAP PLUS MS (Thermo)).

2.5. Electrocatalytic measurement

The electrochemical oxidation of cysteine (Cys) was investigated using glassy carbon electrodes (GCEs) modified with the three IrO_2 NPs. For each systems, a mass of 4.0 mg of NPs was dispersed into 2.0 mL of distilled water containing 100 μL Nafion, and then mixed by using an ultrasonic processor (Wise cleaner, DAIHAN Sci., Wonju, Korea) for 30 min to obtain the homogeneous NPs-Nafion mixture. After that, a volume of 30 μL of the mixture was placed on a GCE, and was dried at 80 °C in a pre-heated oven for 2 h. A cyclic voltammogram of 20 mM Cys in

PBS was obtained for each NPs-Nafion modified GCE. The electrochemical experiments were performed by using a CHI617B potentiostat (CH Instruments, Austin, TX) with a three-electrode cell placed in a Faraday cage. A GCE with a diameter of 2 mm was used as the working electrode, a Pt wire with a diameter of 0.5 mm was used as the counter electrode, and the reference electrode was Ag/AgCl (3 M KCl). All working electrodes were polished with alumina (0.05 μm) paste on micro-cloth pads (Buehler, Lake Bluff, IL) prior to use. After measurements of electrochemical reaction of Cys with NPs-Nafion modified GCE, the catalytic activities were compared. The current values at 0.7 V (because of new small shoulder current) were selected to compare the electrocatalytic activities of the three tested IrO_2 NPs.

3. Results and Discussion

Fig. 1a showed the XRD patterns of IrO_2, $Cr@IrO_2$ and $Co@IrO_2$ NPs, respectively. The XRD spectra of all samples exhibited characteristic peaks at 2θ values of 28.2, 34.9, and 40.1°, corresponding to reflections from the (110), (101), and (200) planes of rutile IrO_2 with good crystallinity, respectively (JCPDS card No. 15870). [26,27] These findings confirmed that no significant phase transitions occurred in both samples, regardless of the transition metal ion dopants. Meanwhile, the XRD patterns of $Cr@IrO_2$ and $Co@IrO_2$ NPs revealed that the peaks noticeably shifted towards larger angles due to the doping of transition metal ions; for instance, the peak at 28.2° and 40.1° from IrO_2 NPs were high angle shifts to 28.25° and 40.64° due to Co^{2+} ion doping (marked #) and 28.46° and 40.84° upon Cr^{3+} ion doping (marked *), respectively (Fig. 1a and S1). This indicated that the overall structures of IrO_2 NPs remained intact upon metal ion doping while $Cr@IrO_2$ NPs exhibited the most significant shift of peaks among the three samples tested, particularly evident in the (110) reflected peak. [28,29] The observed positive peak shift reflected slight shrinking of the unit cell induced by the formation of defect structure on the IrO_2 surface in these two samples and this result directly correlated with their photocatalytic properties (vide infra). [30].

Application of the Scherrer formula revealed that changes in the intensity and width of XRD peaks provide insights into both the size distribution and structure of the samples. [31] Since the intense and interconnected nature of the (110) main reflection peak at approximately 28.2° across all three samples, they were used to determine primary particle size for each sample. The average size of primary crystallites for the three samples was calculated using the Scherrer equation (1):

$$D = 0.9\lambda/B{\cdot}cos\theta \quad (1)$$

where D is the crystallite size (nm), λ is the wavelength of the X-ray radiation (0.154 nm), θ is the Bragg angle and B is the full width at half maximum (FWHM) in radians of the peak at 2θ.

The crystallite sizes of three samples were calculated and summarized in Table S1. Given the primary particle size calculated by the Scherrer equation, SEM images in Fig. 1c-1e exhibited larger crystallites for all three samples, which were regarded as secondary particles. The

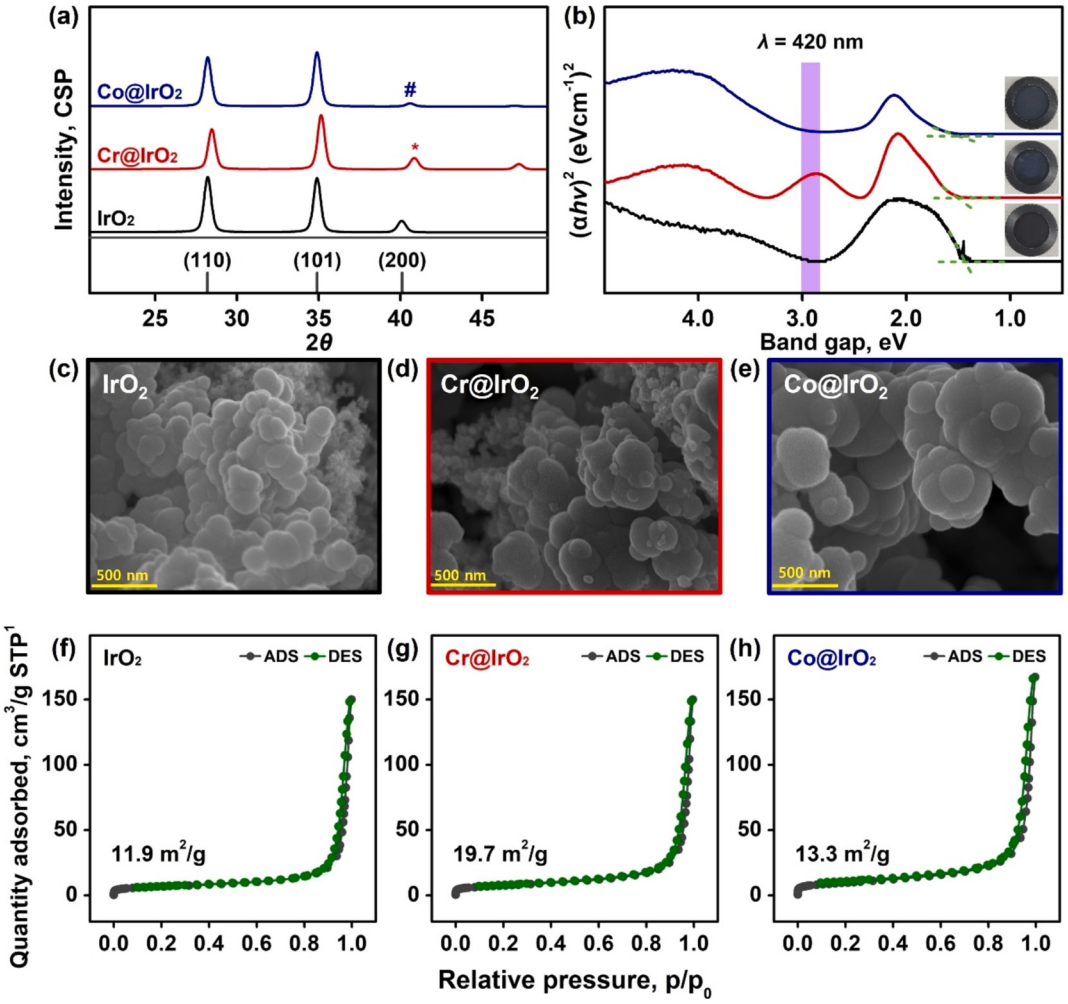


Fig. 1. (a) XRD, (b) DRS spectra, (c)-(e) SEM images, and (f)-(h) BET analysis of the pristine IrO_2 (grey), $Cr@IrO_2$ (red), and $Co@IrO_2$ NPs (navy). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

negligible size differences among the samples implied that their morphologies have no discernible impact on their photo- and electrocatalytic activities.

UV–Vis DRS analysis using Kubelka-Monk plot showed that the band gap values were calculated to be 1.45 eV for IrO_2, 1.52 eV for Cr@IrO_2 and 1.56 eV for Co@IrO_2 NPs, respectively (Fig. 1b). [32] There was no significant change in the band gap values of IrO_2 NPs upon doping with transition metal ions; thus, all three samples were presumably anticipated to display photocatalytic activity in the visible region. Nevertheless, a notable disparity in absorption intensity in the 420 nm region was evidenced among the three samples. For instance, Cr@IrO_2 NPs revealed a substantial increase in the absorption intensity at this wavelength, while IrO_2 and Co@IrO_2 NPs showed minimal absorption. Significant differences in the photocatalytic properties among the three samples could be expected at this specific wavelength. Since the morphology and surface area of the catalysts are important parameters affecting the catalytic efficiency, SEM and surface area analysis were performed on all samples. SEM images of IrO_2, Cr@IrO_2, and Co@IrO_2 NPs depicted in Fig. 1c-1e predominantly displayed a spherical shape with minor protrusions on their surfaces (see Fig. S2). [33] The Brunnauer-Emmett-Teller (BET) method was used to construct adsorption–desorption isotherms using N_2 as the adsorbent. [34] Analysis of the isothermal data presented in Fig. 1f-1 h indicated that the surface areas of IrO_2, Cr@IrO_2, and Co@IrO_2 NPs were 11.9 m^2/g, 19.7 m^2/g and 13.3 m^2/g, respectively, with Cr@IrO_2 NPs possessing the largest surface area. This finding was consistent with the shrinking of the unit cell of IrO_2 NPs upon Cr doping as evidenced by PXRD.

To evaluate their relevance to the photocatalytic properties, the intensities of the V_os determined by core level XPS measurements were compared. Fig. 2a and 2b displayed the Ir 4*f* and O 1 *s* core level spectra for three IrO_2 NPs. Ir 4*f* core level spectra contained two distinct peaks: one at 62.1 eV and another at 60.0 eV for the Ir $4f_{7/2}$ peak, assignable to Ir^{4+} and Ir^{4-x} associated with the IrO_2 and defect-related structures in IrO_2 (i.e., IrO_x), respectively (Fig. 2a). [35,36] Significant increase of the IrO_x peak related to the defect structure was observed in case of Cr@IrO_2 NPs. In Fig. 2b, the O 1 *s* core level spectra featured three prominent peaks at 530.0 eV, 531.8 eV and 532.9 eV, respectively. These peaks were attributed to the IrO_2 matrix, V_os (oxygen vacancies) and –OH, respectively. Considering the intensity change of the V_os peak, Cr@IrO_2 NPs showed considerably pronounced modifications due to the defect structures upon Cr ion doping. [37] Therefore, the XPS analysis of the three IrO_2 NPs revealed that the concentration and intensity of the defect structures (i.e., IrO_x in Ir 4*f* and V_os in O 1 *s*) were consistently greater in Cr@IrO_2 NPs compared to IrO_2 and Co@IrO_2 NPs. Since the amount of defect structure could directly be related to the photocatalytic properties, these features were directly connected to the observed photocatalytic enhancement (vide infra). Therefore, it is vital to determine how the changes in the electronic structure of IrO_2 induced by transition metal doping affected the V_os. [38].

The electronic structures of Cr@IrO_2, and Co@IrO_2 NPs were determined from the Cr and Co *L*-edge XAS spectra (Fig. 2c and 2d). The Cr *L*-edge spectrum showed three distinct peaks at 576.0, 577.3 and 578.7 eV, corresponding to the typical Cr^{3+} L_3 edge structure of Cr_2O_3 (Fig. 2c). [39] Similarly, the Co *L*-edge spectrum displayed three sharp peaks at 777. 5, 778.9 and 781.1 eV correspond to the Co^{2+} L_3 edge structure of CoO (Fig. 2d). [40] These results confirmed that the

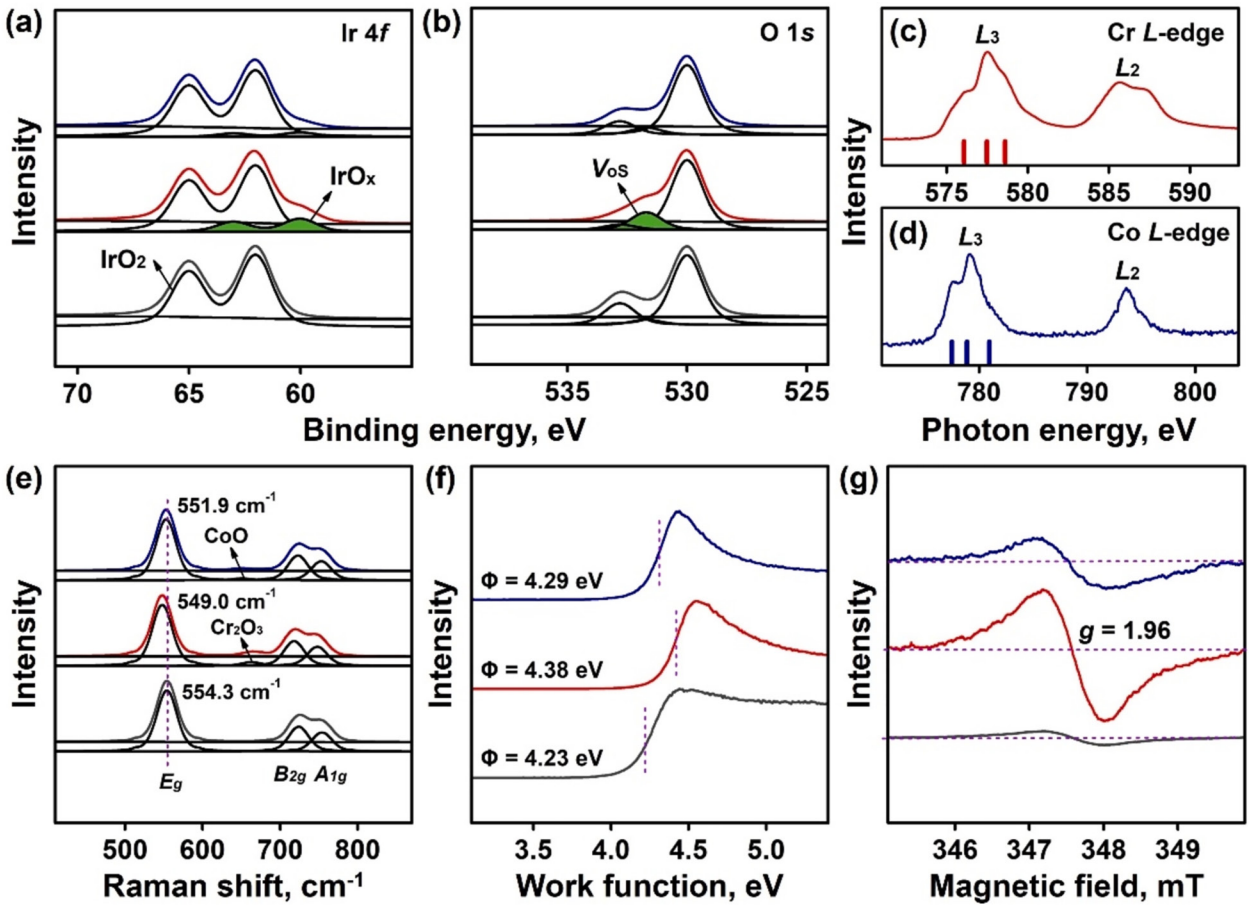


Fig. 2. XPS spectra of (a) Ir 4*f* and (b) O 1 *s* core-level spectra, XAS spectra of (c) Cr *L*-edge of Cr@IrO_2 and (d) Co *L*-edge of Co@IrO_2 NPs, (e) Raman spectra, (f) work function values, and (g) EPR analysis for transition metal ions doped IrO_2 NPs: the pristine IrO_2 (grey), Cr@IrO_2 (red), and Co@IrO_2 NPs (navy). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

electronic states of Cr and Co doped in IrO_2 NPs were consistent with those of Cr_2O_3 and CoO, respectively. Based on XPS and XAS data, we hypothesized that when the doped transition metal possessed a charge state similar to that of Ir in IrO_2 NPs (a + 4 valence state with even number), the charge transfer might be ineffective in generating V_os.

Raman spectra of pristine IrO_2 NPs, Cr@IrO_2, and Co@IrO_2 NPs were shown in Fig. 2e. Three Raman peak shifts at 554.3 (E_g), 726.1 (B_{2g}) and 749.0 cm^{-1} (A_{1g}) were consistent with the typical rutile structure of IrO_2. [41] Compared to pristine IrO_2 NPs, Cr@IrO_2, and Co@IrO_2 NPs exhibited shifts to smaller Raman shift values, indicating a weakening of the Ir-O bonds due to the transition metal doping. Notably, for Cr@IrO_2 NPs, the E_g peak showed a shift of approximately 5.3 cm^{-1} from the pristine IrO_2 NPs. Therefore, we suggested that Cr ion doping was more effective in the formation of defect structures. This trend was supported by the observed changes in the electronic structure of the two core levels identified in the XPS analysis (Fig. 2a and 2b).

The work function measurements generally indicate the potential charge transfer between IrO_2 and the dopants (Fig. 2f). Following doping with 5 wt% Cr and Co ions, the work function was significantly increased to 4.38 eV for Cr@IrO_2 NPs, whereas it remianed relatively unchanged at 4.29 eV for Co@IrO_2 NPs, compared to 4.23 eV for IrO_2 NPs. [42] This clearly demonstrated that Cr@IrO_2 NPs could facilitate carrier transfer by light or electricity at the surface, indicating potential changes in their catalytic properties. In addition, electron paramagnetic resonance (EPR) experiments were carried out to provide evidence for the presence of oxygen vacancies in Cr@IrO_2, and Co@IrO_2 NPs (Fig. 2g). [43] The EPR spectra of the three IrO_2 NPs exhibited a prominent signal centered at a *g*-factor of approximately 1.96, indicating a Zeeman effect caused by a single electron trapped in the oxygen vacancies. In particular, the charge mismatch in Cr@IrO_2 NPs showed a significant increase in the EPR signal indicating the presence of a substantial number of V_os.

Fig. 3 illustrated the concentration–time profile (C_t/C_0) of photocatalytic conversion reaction efficiencies in the presence and absence of IrO_2, Cr@IrO_2, and Co@IrO_2 NPs under visible light irradiation and the analysis of oxidized organic products derived from hydroxymethylfurfural (HMF) and benzyl alcohol (BA). [44–46] Very interestingly, the photocatalytic conversion of both HMF and BA was found to be much more effective with Cr@IrO_2 NPs under visible light irradiation compared to IrO_2 and Co@IrO_2 NPs (Fig. 3a and 3b). This result may align with the increase in the oxygen defect structure described in previous XRD and XPS analyses as well as the enhanced absorption intensity at 420 nm shown by the DRS spectra.

Table 1 summarized the different values of photoconversion activity profiles in the photocatalytic oxidation of HMF and BA in the presence of IrO_2, Cr@IrO_2, and Co@IrO_2 NPs at 420 nm with literature reports for comparison. Fig. 3c showed the variation of the efficiency of FDCA, the 6-electron oxidized product of HMF, over time under 420 nm. The FDCA efficiency follows the same trend as the photoconversion efficiency of HMF shown in Fig. 3a, with the highest FDCA efficiency observed for Cr@IrO_2 NPs. Given that Cr@IrO_2 NPs contained Cr^{3+} (odd number) ionic states, an increase in oxygen vacancies may presumably be expected owing to the occurrence of stronger charge mismatch phenomena at the interface between Cr_2O_3 and IrO_2 NPs. These oxygen vacancies, rich in localized electrons, conferred an electron-rich character to the oxygen-deficient surface to facilitate the activation of various reactive molecules. Since the reaction conditions (e.g. solvent, temperature, presence of base or oxidizing agent) are quite different for each catalyst,

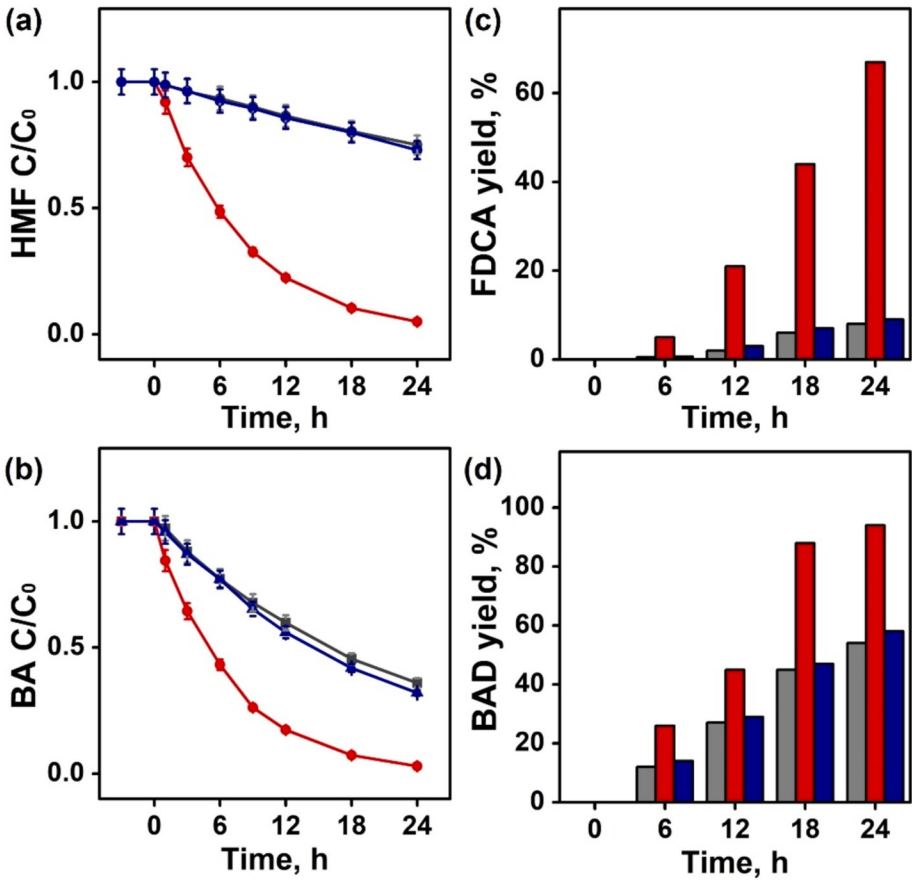


Fig. 3. Photocatalytic degradation of (a) HMF and (b) BA in the presence of transition metal ion doped IrO_2 NPs: pristine IrO_2, Cr@IrO_2, and Co@IrO_2 NPs. The experimental conditions were as follows: [catalyst] = 0.5 g/L, $\lambda \geq 420$ nm, $[HMF]_0 = [BA]_0 = 25$ mM. Yield of end products, (c) FDCA, from HMF and (d) benzaldehyde (BAD), from BA depending on the irradiation time.

Table 1
Photocatalytic oxidation of HMF and BA over IrO_2, Cr@IrO_2, and Co@ IrO_2 NPs and comparison with literature reports in HMF oxidation.

catalyst	solvent	reaction conditions	FDCA yield (%)	HMF conversion (%)	ref
IrO_2	H_2O	air, rt, 24 h	8.0	25	this work
Cr@IrO_2	H_2O	air, rt, 24 h	67	95	this work
Co@IrO_2	H_2O	air, rt, 24 h	9.0	27	this work
2.5 % Au/ZnO	H_2O	air, Na_2CO_3, 5 h	29.4	32.2	47
Ag_1-Au_2/TiO_2	H_2O	O_2, 25 °C, Na_2CO_3, 5 h	41.8	42.6	48
A-TiO_2	H_2O	air, rt, 24 h	~40	>99	49
CoPz/g-C_3N_4	buffer	air flow, rt, $Na_2B_4O_7$, 14 h	99.1	96.1	50

it is difficult to directly compare the activities with reported catalysts. It is, however, noteworthy that Cr@IrO_2 exhibited a relatively high yield in photocatalytic oxidation of HMF under aqueous solution conditions without any additional bases or oxidizing agents. Thus, doping Cr ion would be considered as a facile and effective strategy to enhance photocatalytic activity of IrO_2 NPs. The selective oxidation of the –OH or –CHO groups of HMF for the synthesis of FDCA leads to the formation of two key intermediates: 5-diformyl furan (DFF), and 5-hydroxymethyl-2-furancarboxylic acid (HMFCA). These intermediates served as important raw materials in the synthesis of pharmaceutical intermediates and valuable monomers in polymer synthesis, respectively. Hence, identifying the functional groups involved in the oxidation of HMF by Cr@IrO_2 NPs is essential (See Fig. S3).

In this photocatalytic system, we observed that the –OH group of HMF was selectively oxidized to –CHO, resulting in the exclusive formation of DFF as an intermediate, which, consequently led to the formation of FDCA as the final product. Notably, the other anticipated intermediate, HMFCA, was not detected. Thus, the selective oxidation of HMF using Cr@IrO_2 NPs dispersed in aqueous solution followed a reaction pathway from HMF to DFF, ultimately yielding FDCA as the sole final product under all conditions within 24 h (see Scheme 1).

The proposed mechanism of photocatalytic oxidation of Cr@IrO_2 NPs to FDCA of HMF can be described as follows. Under irradiation, the electrons residing in valence band of Cr(@IrO_2 NPs are promoted to the conduction band and generates electrons and holes separation (eq (2). As the oxygen vacancies formed at the interface of the two heterostructure systems, electrons would be trapped and localized on the oxygen vacancies site, which in turn, could carry out the electron transfer from defect sites to O_2 to promote the formation of superoxide anion ($O_2^{\bullet-}$) (eq (3). On the other hand, the hole created on the valence band can react with H_2O to generate hydroxyl radical species ($HO^{\bullet}$) (eq (4). As a result, HMF is converted to FDCA via photocatalytic oxidation under the action of $O_2^{\bullet-}$ and $HO^{\bullet}$.

$$Cr@IrO_2 + h\nu \rightarrow Cr@IrO_2 + h^+ + e^- \quad (2)$$

$$e^- + O_2 \rightarrow O_2^{\bullet-} \quad (3)$$

$$h^+ + H_2O \rightarrow {}^{\bullet}OH + H^+ \quad (4)$$

To confirm the involvement of reactive oxygen species such as $O_2^{\bullet-}$ in the photocatalytic system, we used 5,5-dimethyl-1-pyrroline N-oxide (DMPO) as a $O_2^{\bullet-}$ radical scavenger, which completely inhibited the photocatalytic activity (Fig. S4). As shown in this fig., the preferential reaction of DMPO with the radicals formed in the photoreaction, which interferes with the Cr@IrO_2 NPs photoreaction, confirms the justification of our proposed mechanism.

The oxidation efficiency from BA to benzyl aldehyde as depicted in Fig. 3d and Table 1, follows a similar trend. This observation suggested that the improved photocatalytic performance could be attributed to the increased formation of defect structures on the surface of Cr@IrO_2 NPs under visible light wavelengths.

To examine electrocatalytic properties of three different IrO_2 NPs, three glassy carbon electrodes (GCE) were prepared with IrO_2-Nafion catalysts. [51] Cyclic Voltammograms were obtained in the presence and absence of IrO_2-Nafion, Cr@IrO_2-Nafion, and Co@IrO_2-Nafion NPs in phosphate buffer solution containing 0.01 M of cysteine (Cys) (Fig. 4a – 4d). In the absence of IrO_2-Nafion NPs, bare GCE exhibited slow electrocatalytic oxidation currents, attributed to the inherently sluggish electrochemical oxidation of Cys on bare GCE. In contrast, GCEs modified with three different IrO_2-Nafion catalysts displayed increased electrocatalytic currents for the oxidation of Cys. The currents were measured to be 2.1 (±0.7) μA for IrO_2, 8.7 (±0.6) μA for Cr@IrO_2, and 2.4 (±0.6) μA Co@IrO_2 NPs. Especially, significant changes were observed only for Cr@IrO_2 NPs, in a good agreement with the photocatalytic conversion efficiency results. This finding is remarkable, as it demonstrates that doping IrO_2 NPs with a small amount of Cr ions can enhance the electro-oxidation reaction. While further optimization of the electrocatalytic conditions is necessary for the selective and sensitive detection of Cys using Cr@IrO_2 NPs, it is noteworthy that we have confirmed the potential of Cr@IrO_2 NPs as a multifunctional catalyst in diverse photo- and electro-chemical oxidation reactions.

Based on the abovementioned experimental results, we proposed that the two doped transition metal ions formed metal oxides (Cr_2O_3 and CoO) possessing different charge states (Cr^{3+} with an odd number and Co^{2+} with an even number) compared to Ir in IrO_2 NPs (+4 valence state with an even number). It is suggested that charge transfer might not effectively generate oxygen vacancies if the doped TM shared the even charge number as Ir in IrO_2 NPs (Eq. 5). However, since Cr@IrO_2 NPs contained Cr in the odd number of + 3 state (Eq. 6), the formation of a greater number of oxygen vacancies could be expected owing to the charge mismatch between Cr_2O_3 and IrO_2.

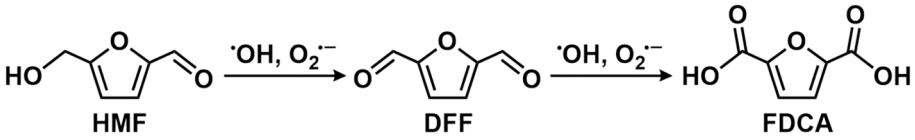
$IrO_2 + Co^{2+} \rightarrow IrO_x + CoO$ x ≒ 2 (5).

$IrO_2 + Cr^{3+} \rightarrow IrO_x + Cr_2O_3$ x < 2 (6).

Consequently, the generation of V_os through charge transfer was more effective in IrO_2 NPs doped with Cr^{3+} ions presumably owing to the charge mismatch between the doped ions and the host ions. More specifically, the charge imbalanced dopant metal ions partially weakened the Ir-O interaction leading to the creation of the V_os states. These V_os states are an important factor in improving photocatalytic and electrocatalytic performance of various oxidation reactions. The number of the V_os increased significantly for Cr@IrO_2 NPs, whereas no significant change was observed for Co@IrO_2 NPs.

4. Conclusion

In summary, our experimental findings reveal the successful synthesis of Cr@IrO_2 and Co@IrO_2 nanoparticles (NPs) derived from pristine IrO_2 NPs. We identified that the doped transition metal ions formed



Scheme 1. The reaction pathway for selective oxidation of HMF to FDCA.

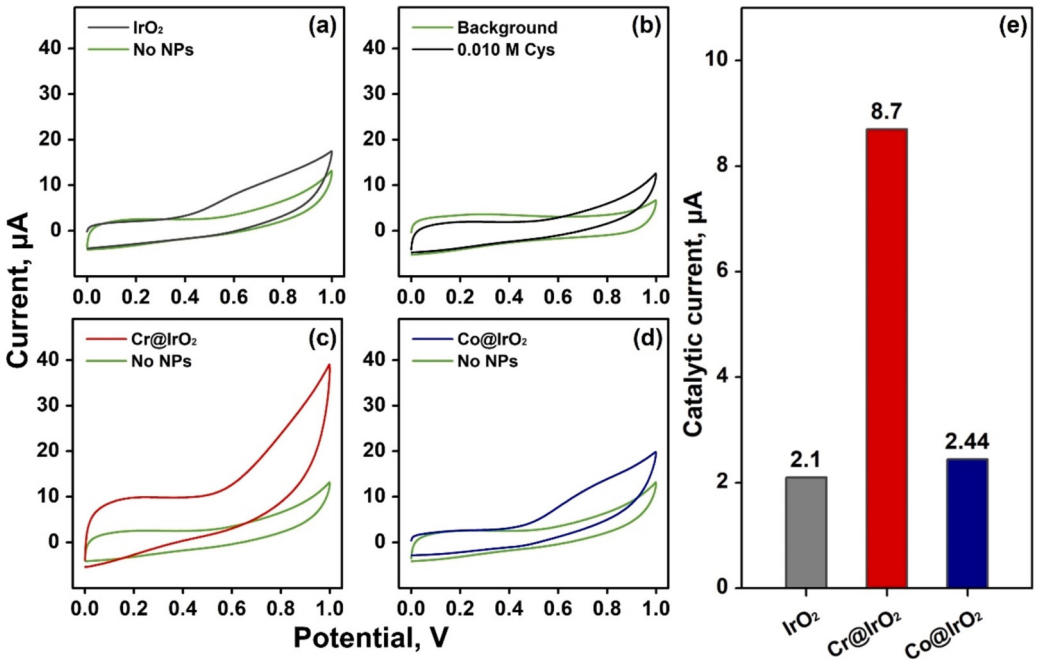


Fig. 4. (a)-(d) CVs (at a scan rate of 50 mV/s) in PBS containing 0.1 M Cys. (a) a sluggish oxidation current at a bare GCE (violet lines) and with (b) IrO_2, (c) $Cr@IrO_2$, and (d) $Co@IrO_2$ NPs. (e) Catalytic currents resulting from the electrochemical oxidation of Cys for the three types of IrO_2 NPs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distinct metal oxides (Cr_2O_3 and CoO) with charge states differing from the host Ir ion in IrO_2 (Cr^{3+} with an odd charge state, Co^{2+} with an even charge state, and Ir^{4+} with an even charge state). Notably, doping with Cr^{3+} ions significantly enhanced the visible light absorption capability of IrO_2 NPs, which are typically inactive in this spectral region.

We demonstrated that the generation of oxygenvacancies via charge transfer was more pronounced in $Cr@IrO_2$ NPs due to the charge mismatch between Cr^{3+} ions and the host Ir^{4+} ions. This charge imbalance partially weakened the Ir-O bonds, leading to a substantial increase in the formation of oxygen vacancies. In contrast, $Co@IrO_2$ NPs showed no significant change in defect concentration. The presence of these oxygen vacancies was found to play a critical role in enhancing the photocatalytic and electrocatalytic performance, particularly in the oxidation reactions of HMF and BA, where improved efficiency and product yields were observed.

Furthermore, our study established the potential of $Cr@IrO_2$ NPs as efficient electrocatalysts for the electrocatalytic oxidation of cysteine. These results demonstrate that doping IrO_2 nanoparticles with Cr ions not only preserves the crystalline structure but also promotes the formation of oxygen vacancies, thereby unlocking pathways to superior catalytic performance.

Given these promising results, several future research directions are underway. These include optimizing doping concentrations, exploring other transition metal dopants, and synthesizing additional renewable fuels. Collectively, this work and subsequent studies are expected to make significant contributions to the fields of photocatalysis and electrocatalysis, paving the way for the development of sustainable multifunctional catalysts for renewable fuel production and green chemical synthesis.

CRediT authorship contribution statement

Hyeri Jeon: Investigation, Formal analysis. **Dung Thanh Hoang:** Investigation, Formal analysis. **Gyuri Kim:** Investigation. **In Young Kim:** Writing – review & editing, Investigation, Formal analysis. **Hangil Lee:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Conceptualization. **Seungwoo Hong:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (RS-2024-00346153 for H. Lee, 2021R1A6A1A10039823 for S. Hong, and and RS-2023-00272031 for H. Jeon) and the Seoul R&BD Program funded by the Seoul Business Agency (SBA) (BT230224 for I. Y. Kim).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apsusc.2024.162088.

Data availability

Data will be made available on request.

References

[1] Q. Zhu, Q. Xu, M. Du, X. Zeng, G. Zhong, B. Qiu, J. Zhang, Recent Progress of Metal Sulfide Photocatalysts for Solar Energy Conversion, Adv. Mater. 34 (2022) 2202929, https://doi.org/10.1002/adma.202202929.
[2] B. Chandran, J.K. Oh, S.W. Lee, D.Y. Um, S.U. Kim, V. Veeramuthu, J.S. Park, H. Shuo, C.R. Lee, Y.H. Ra, Solar-Driven Sustainability: III–V Semiconductor for Green Energy Production Technologies, Nanomicro Lett. 16 (2024) 244, https://doi.org/10.1007/s40820-024-01412-6.
[3] L.I. Granone, F. Sieland, N. Zheng, R. Dillert, D.W. Bahnemann, Photocatalytic Conversion of Biomass into Valuable Products: A Meaningful Approach? Green Chem. 20 (2018) 1169–1192, https://doi.org/10.1039/C7GC03522E.
[4] Z. Su, C. Ye, Y. Xu, B. Wu, D.-H. Kuo, X. Wu, B. Yang, P. Zhang, L. Chen, D. Lu, J. Lin, Z. Yuan, X. Chen, Synergistic vacancy defects and the surface hydrophobic-

to-superhydrophilic wetting engineering in W/S co-doped BiOI for enhanced photocatalytic hydrogen evolution, Chem. Eng. J. 496 (2024) 154282, https://doi.org/10.1016/j.cej.2024.154282.
[5] Z. Su, B. Wu, D.-H. Kuo, L. Chen, P. Zhang, B. Yang, X. Wu, D. Lu, J. Lin, X. Chen, Synergistic hydrazine-driven regulation and Mo/S co-doping to endow BiOBr with heterovalent molybdenum states and abundant oxygen vacancy defects for photocatalytic hydrogen evolution, J. Mater. Chem. A 12 (2024) 28486, https://doi.org/10.1039/d4ta05641h.
[6] X. Chen, Q. Wu, D.-H. Kuo, A.B. Abdeta, H. Zhang, P. Li, T. Huang, O.A. Zelekew, J. Lin, Initiating highly efficient (Bi,Ce)2(O,S)3−x oxysulfide catalysts with rich oxygen vacancies for hydrogen evolution via adjusting valence band configuration, J. Mater. Chem. A 11 (2023) 4126–4141, https://doi.org/10.1039/d2ta09780j.
[7] A.B. Abdeta, Q. Wua, D.-H. Kuo, P. Li, H. Zhang, T. Huang, J. Zhang, J. Lin, X. Chen, Mo-/O-deficient Bi2Mo3(S,O)12 oxysulfide for enhanced visible-light photocatalytic H2 evolution and pollutant reduction via in-situ generated protons: A case of material design in converting an oxidative Bi2Mo3O12 catalyst for the reduction, J. Catal. 413 (2022) 1056–1069, https://doi.org/10.1016/j.jcat.2022.08.017.
[8] X. Chen, H. Sun, D.-H. Kuo, A.B. Abdeta, O.A. Zelekew, Y. Guo, J. Zhang, Z. Yuan, J. Lin, Spherical nanoflower-like bimetallic (Mo,Ni)(S,O)3-x sulfo-oxide catalysts for efficient hydrogen evolution under visible light, Appl. Catal. B-Environ. 287 (2021) 119992, https://doi.org/10.1016/j.apcatb.2021.119992.
[9] X. Chen, T. Huang, D.-H. Kuo, H. Sun, P. Li, O.A. Zelekew, A.B. Abdeta, Q. Wu, J. Zhang, Z. Yuan, J. Lin, Material design with the concept of solid solution-type defect engineering in realizing the conversion of an electrocatalyst of NiS2 into a photocatalyst for hydrogen evolution, Appl. Catal. B-Environ. 298 (2021) 120542, https://doi.org/10.1016/j.apcatb.2021.120542.
[10] L. Han, F. Jing, J. Zhang, X.Z. Luo, Y.L. Zhong, K. Wang, S.H. Zang, D.H. Teng, Y. Liu, J. Chen, C. Yang, Y.T. Zhou, Environment Friendly and Remarkably Efficient Photocatalytic Hydrogen Evolution Based on Metal Organic Framework Derived Hexagonal/Cubic In_2O_3 Phase-Junction, Appl. Catal. B: Environ. 282 (2021) 119602, https://doi.org/10.1016/j.apcatb.2020.119602.
[11] M. Ikram, M. Rashid, A. Haider, S. Naz, J. Haider, A. Raza, M.T. Ansar, M.K. Uddin, N.M. Ali, S.S. Ahmed, M. Imran, S. Dilpazir, Q. Khan, M. Maqbool, A Review of Photocatalytic Characterization, and Environmental Cleaning, of Metal Oxide Nanostructured Materials, Sustain. Mater. Technol. 30 (2021) e00343, https://doi.org/10.1016/j.susmat.2021.e00343.
[12] J. Lim, D. Park, S.S. Jeon, C.W. Roh, J. Choi, D. Yoon, M. Park, H. Jung, H. Lee, Ultrathin IrO_2 Nanoneedles for Electrochemical Water Oxidation, Adv. Funct. Mater. 28 (2018) 1704796, https://doi.org/10.1002/adfm.201704796.
[13] L. Li, B. Wang, G. Zhang, G. Yang, T. Yang, S. Yang, S. Yang, Electrochemically Modifying the Electronic Structure of IrO_2 Nanoparticles for Overall Electrochemical Water Splitting with Extensive Adaptability, Adv. Energy Mater. 10 (2020) 2001600, https://doi.org/10.1002/aenm.202001600.
[14] M.R. Nellist, F.A.L. Laskowski, F. Lin, T.J. Mills, S.W. Boettcher, Semiconductor–Electrocatalyst Interfaces: Theory, Experiment, and Applications in Photoelectrochemical Water Splitting, Acc. Chem. Res. 49 (2016) 733–740, https://doi.org/10.1021/acs.accounts.6b00001.
[15] E. Oakton, D. Lebedev, M. Povia, D.F. Abbott, E. Fabbri, A. Fedorov, M. Nachtegaal, C. Copéret, T.J. Schmidt, IrO_2-TiO_2: A High-Surface-Area, Active, and Stable Electrocatalyst for the Oxygen Evolution Reaction, ACS Catal. 7 (2017) 2346–2352, https://doi.org/10.1021/acscatal.6b03246.
[16] Y. Wang, S. Hao, X. Liu, Q. Wang, Z. Su, L. Lei, X. Zhang, Ce-Doped IrO_2 Electrocatalysts with Enhanced Performance for Water Oxidation in Acidic Media, ACS Appl. Mater. Interfaces 12 (2020) 37006–37012, https://doi.org/10.1021/acsami.0c00389.
[17] A. Pătrua, A. Rabisa, S.E. Temmela, R. Kotza, T.J. Schmidta, Pt/IrO_2–TiO_2 Cathode Catalyst for Low Temperature Polymer Electrolyte Fuel Cell – Application in MEAs, Performance and Stability, Catal. Today 262 (2016) 161–169, https://doi.org/10.1016/j.cattod.2015.09.009.
[18] S. Chen, J. Duan, M. Jaroniec, S.-Z. Qiao, Nitrogen and Oxygen Dual-Doped Carbon Hydrogel Film as a Substrate-Free Electrode for Highly Efficient Oxygen Evolution Reaction, Adv. Mater. 26 (2014) 2925–2930, https://doi.org/10.1002/adma.201305608.
[19] K. Kadakia, M.K. Datta, P.H. Jampani, S.K. Park, P.N. Kumta, Novel F-Doped IrO_2 Oxygen Evolution Electrocatalyst for PEM Based Water Electrolysis, J. Power Sources 222 (2013) 313–331, https://doi.org/10.1016/j.jpowsour.2012.08.051.
[20] H. Lee, J.Y. Kim, S.Y. Lee, J.A. Hong, N. Kim, J. Baik, Y.J. Hwang, Comparative Study of Catalytic Activities among Transition Metal-Doped IrO_2 Nanoparticles, Sci. Rep. 8 (2018) 16777, https://doi.org/10.1038/s41598-018-35116-w.
[21] L.E. Camacho-Forero, F. Godínez-Salomón, G. Ramos-Sánchez, C.P. Rhodes, P. B. Balbuena, Theoretical and Experimental Study of the Effects of Cobalt and Nickel Doping within IrO_2 on the Acidic Oxygen Evolution Reaction, J. Catal. 408 (2022) 64–80, https://doi.org/10.1016/j.jcat.2022.02.016.
[22] W. Jiang, H. Loh, B.Q.L. Low, H. Zhu, J. Low, J.Z.X. Heng, K.Y. Tang, Z. Li, X. J. Loh, E. Ye, Y. Xiong, Role of Oxygen Vacancy in Metal Oxides for Photocatalytic CO_2 Reduction, Appl. Catal. B: Environ. 321 (2023) 122079, https://doi.org/10.1016/j.apcatb.2022.122079.
[23] Y.H. Kim, S. Kim, K.J. Kim, C. Kim, J.H. Jang, Y.M. Kim, H. Lee, Multiscale Probing of the Influence of the Defect-Induced Variation of Oxygen Vacancies on the Photocatalytic Activity of Doped ZnO Nanoparticles, J. Mater. Chem. A 8 (2020) 25345–25354, https://doi.org/10.1039/D0TA08367D.
[24] H. Li, Y. Xu, N. Lv, Q. Zhang, X. Zhang, Z. Wei, Y. Wang, H. Tang, H. Pan, Ti-Doped SnO_2 Supports IrO_2 Electrocatalysts for the Oxygen Evolution Reaction (OER) in PEM Water Electrolysis, ACS Sustain. Chem. Eng. 11 (2023) 1121–1132, https://doi.org/10.1021/acssuschemeng.2c06368.
[25] K.E. Michaux, R.W. Murray, Formation of Iridium (IV) Oxide (IrOx) Films by Electroflocculation, Langmuir 29 (2013) 12254–12258, https://doi.org/10.1021/la4025876.
[26] D. Jin, J. Kang, S. Prabhakaran, Y. Lee, M.H. Kim, D.H. Kim, C. Lee, Chromium-Rich CrxIr1− xO_2 Wire-in-Tube Alloys for Boosted Water Oxidation with Long Standing Electrocatalytic Activity, J. Mater. Chem. A 10 (2022) 13803–13813, https://doi.org/10.1039/D2TA02984G.
[27] S. Siracusano, V. Baglio, S.A. Grigoriev, L. Merlo, V.N. Fateev, A.S.T. Aricò, The Influence of Iridium Chemical Oxidation State on the Performance and Durability of Oxygen Evolution Catalysts in PEM Electrolysis, J. Power Sources 366 (2017) 105–114, https://doi.org/10.1016/j.jpowsour.2017.09.020.
[28] J. Zhu, Z. Deng, F. Chen, J. Zhang, H. Chen, M. Anpo, J. Huang, L. Zhang, Hydrothermal Doping Method for Preparation of Cr^{3+}-TiO_2 Photocatalysts with Concentration Gradient Distribution of Cr^{3+}, Appl. Catal. B: Environ. 62 (2006) 329–335, https://doi.org/10.1016/j.apcatb.2005.08.013.
[29] D. Von Dreifus, A.J.A. De Oliveira, A.V. Do Rosario, E.C. Pereira, Magnetic and Structural Characterization of IrO_2 and Co: IrO_2 Samples Synthesized via Pechini Method, J. Supercond. Nov. Magn. 26 (2013) 2319–2321, https://doi.org/10.1007/s10948-012-1424-5.
[30] F. Ahmed, S. Kumar, N. Arshi, M.S. Anwar, B.H. Koo, C.G. Lee, Doping Effects of Co^{2+} Ions on Structural and Magnetic Properties of ZnO Nanoparticles, Microelectron. Eng. 89 (2012) 129–132, https://doi.org/10.1016/j.mee.2011.03.149.
[31] V.K. Puthiyapura, S. Pasupathi, H. Su, X. Liu, B. Pollet, K. Scott, Investigation of Supported IrO_2 as Electrocatalyst for the Oxygen Evolution Reaction in Proton Exchange Membrane Water Electrolyser, Int. J. Hydrogen Energy 39 (2014) 1905–1913, https://doi.org/10.1016/j.ijhydene.2013.11.056.
[32] J. Feng, Z. Dong, Y. Ji, Y. Li, Accelerating the Discovery of Metastable IrO_2 for the Oxygen Evolution Reaction by the Self-Learning-Input Graph Neural Network, JACS Au 3 (2023) 1131–1140, https://doi.org/10.1021/jacsau.2c00709.
[33] Q. Deng, Y. Sun, J. Wang, S. Chang, M. Ji, Y. Qu, K. Zhang, B. Li, Boosting OER Performance of IrO_2 in Acid via Urchin-Like Hierarchical-Structure Design, Dalton Trans. 50 (2021) 6083–6087, https://doi.org/10.1039/D1DT00329A.
[34] V.K. Puthiyapura, M. Mamlouk, S. Pasupathi, B.G. Pollet, K. Scott, Physical and Electrochemical Evaluation of ATO Supported IrO_2 Catalyst for Proton Exchange Membrane Water Electrolyser, J. Power Sources 269 (2014) 451–460, https://doi.org/10.1016/j.jpowsour.2014.06.078.
[35] H. Lv, Y. Sun, S. Wang, W. Ji, W. Zhou, C. Zhang, N/C Doped Nano-Size IrO_2 Catalyst of High Activity and Stability in Proton Exchange Membrane Water Electrolysis, Int. J. Hydrogen Energy 48 (2023) 16949–16957, https://doi.org/10.1016/j.ijhydene.2023.01.105.
[36] S. Pichaikaran, S. Kotteswaran, M.K. Francis, P.B. Bhargav, W. Bo, N. Ahmed, C. Balaji, IrO_2/Co_3O_4 Supported Mesoporous SBA-16: An Efficient Electro-Catalyst for Oxygen Evolution Reaction, Mol. Catal. 547 (2023) 113383, https://doi.org/10.1016/j.mcat.2023.113383.
[37] C. Cai, S. Han, Y. Tang, Engineering oxygen vacancies on dendrite-like IrO_2 for the oxygen evolution reaction in acidic solution, Sustain, Energy Fuels 4 (2020) 2462–2468, https://doi.org/10.1039/D0SE00007H.
[38] G. Liu, F. Hou, X. Wang, B. Fang, Ir-IrO_2 with Heterogeneous Interfaces and Oxygen Vacancies-Rich Surfaces for Highly Efficient Oxygen Evolution Reaction, Appl. Surf. Sci. 615 (2023) 156333, https://doi.org/10.1016/j.apsusc.2023.156333.
[39] H.Q. Fu, J. Liu, N.M. Bedford, Y. Wang, J.W. Sun, Y. Zou, M. Dong, J. Wright, H. Diao, P. Liu, H.G. Yang, H. Zhao, Synergistic Cr_2O_3@Ag Heterostructure Enhanced Electrocatalytic CO_2 Reduction to CO, Adv. Mater. 34 (2022) 2202854, https://doi.org/10.1002/adma.202202854.
[40] Y. Li, C. Qu, S. Meng, J. Hu, Z. Gao, H. Xu, R. Gao, M. Feng, Revealing Electronic State Evolution of Co (II)/Co (III) in CoO (111) Plane during OER Process through Magnetic Measurement, Chin. Chem. Lett. (2024) 109872, https://doi.org/10.1016/j.cclet.2024.109872.
[41] J. Mougin, N. Rosman, G. Lucazeau, A. Galerie, In Situ Raman Monitoring of Chromium Oxide Scale Growth for Stress Determination, J. Raman Spectrosc. 32 (2001) 739–744, https://doi.org/10.1002/jrs.734.
[42] Y. Wang, M. Zhang, Z. Kang, L. Shi, Y. Shen, B. Tian, Y. Zou, H. Chen, X. Zou, Nano-Metal Diborides-Supported Anode Catalyst with Strongly Coupled TaOx/IrO_2 Catalytic Layer for Low-Iridium-Loading Proton Exchange Membrane Electrolyzer, Nat. Commun. 14 (2023) 5119, https://doi.org/10.1038/s41467-023-40912-8.
[43] J. Chen, X. Wang, L. Zhang, Z. Rui, Strong Metal-Support Interaction Assisted Redispersion Strategy for Obtaining Ultrafine and Stable IrO_2/Ir Active Sites with Exceptional Methane Oxidation Activity, Appl. Catal. B: Environ. 297 (2021) 120410, https://doi.org/10.1016/j.apcatb.2021.120410.
[44] L. Guo, X. Zhang, L. Gan, L. Pan, C. Shi, Z.F. Huang, X. Zhang, J.J. Zou, Advances in Selective Electrochemical Oxidation of 5-Hydroxymethylfurfural to Produce High-Value Chemicals, Adv. Sci. 10 (2023) 2205540, https://doi.org/10.1002/advs.202205540.
[45] V.N. Pham, H. Jeon, S. Hong, H. Lee, Selective Oxidation of Biomass Molecules via ZnO Nanoparticles Modified Using Charge Mismatch of the Doped Co Ions, Inorg. Chem. 61 (2022) 16887–16894, https://doi.org/10.1021/acs.inorgchem.2c02934.
[46] V.N. Pham, H. Jeon, S. Han, S. Hong, H. Lee, Eco-Friendly Transition-Metal-Ion-Doped rGO Nanocomposites with Enhanced Photocatalytic Activities, Adv. Sustain. Syst. 6 (2022) 2200355, https://doi.org/10.1002/adsu.202200355.
[51] A. Banti, C. Zafeiridou, M. Charalampakis, O.-N. Spyridou, J. Georgieva, V. Binas, E. Mitrousi, S. Sotiropoulos, IrO_2 Oxygen Evolution Catalysts Prepared by an Optimized Photodeposition Process on TiO_2 Substrates, Molecules 29 (2024) 2392, https://doi.org/10.3390/molecules29102392.

Further reading

[47] P. Zhu, W. Zhang, Q. Li, H. Xia, Visible-Light-Driven Photocatalytic Oxidation of 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid over Plasmonic Au/ZnO Catalyst, ACS Sustainable Chem. Eng. 10 (2022) 8778–8787, https://doi.org/10.1021/acssuschemeng.2c01143.

[48] W. Zhang, Q. Li, H. Xia, Photocatalytic oxidation of 5-hydroxymethylfurfural to furandicarboxylic acid over the Au-Ag/TiO2 catalysts under visible light irradiation, Appl. Surf. Sci. 613 (2023) 156036, https://doi.org/10.1016/j.apsusc.2022.156036.

[49] T. Ji, C. Liu, X. Lu, J. Zhu, Coupled Chemical and Thermal Drivers in Microwaves toward Ultrafast HMF Oxidation to FDCA, ACS Sustainable Chem. Eng. 6 (2018) 11493–11501, https://doi.org/10.1021/acssuschemeng.8b01630.

[50] S. Xu, P. Zhou, Z. Zhang, C. Yang, B. Zhang, K. Deng, S. Bottle, H. Zhu, Selective Oxidation of 5-Hydroxymethylfurfural to 2,5-Furandicarboxylic Acid Using O2 and a Photocatalyst of Co-thioporphyrazine Bonded to g-C3N4, J. Am. Chem. Soc. 139 (2017) 14775–14782, https://doi.org/10.1021/jacs.7b08861.

RESEARCH INSTITUTE OF NATURAL SCIENCES

숙명여자대학교
자연과학연구소

9 772713 792008 15
ISSN 2713-7929