Supplemental Material for "Simultaneous Estimation and Variable Selection for Linear Transformation Models in the Presence of Functional and Missing Covariates with the Application to Alzheimer's Disease"

Abstract

In the Supplementary Material, we will first describe in Web Appendix A the regularity conditions required for the establishment of the asymptotic properties given in Theorems 1 and 2 of the main paper and then sketch the proofs of the two theorems in Web Appendix B. Web Appendix C contains the additional simulation tables described in the main paper.

Web Appendix A: Regularity Conditions

For the description of the required regularity conditions, we will first define some extra notation. Following the notation used in van der Vaart and Wellner (1996), we denote $\mathbb{P}f = \int_{\mathcal{X}} f(o)dP(o)$ and $\mathbb{P}_n f = n^{-1}\sum_{i=1}^{n} f(O_i)$, the empirical process indexed by the function f evaluated as O_i, the sample points of individual i, with $\mathcal{X}$ denoting the sample space. To avoid confusion, we add a subscript n to all estimators in the following such as $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\boldsymbol{\theta}}_n$, and $\widehat{\Lambda}_n$. Let λ_k be the eigenvalues and $\psi_k(s)$ the eigenfunctions of the covariance operator of $X(s)$. Also let α_{k0} be the true coefficients in the eigenbasis expansion $\gamma_0(s) = \sum_{k=1}^{\infty}\alpha_{k0}\psi_k(s)$, and $\widetilde{R}$ be the truncation number for the FPCA expansion.

We will first describe the conditions concerning the functional predictor $X(\cdot)$ and its FPCA, similar to those in Hall and Hosseini-Nasab (2006), Hall and Horowitz (2007), Hall and Hosseini-Nasab (2009) and Kong et al. (2018), and then the conditions related to the survival data, scalar covariates and model structure, similar to those in Du et al. (2025) and Zhou et al. (2017).

Conditions on Functional Predictor $X(\cdot)$ and Coefficients α_k

Condition 1. *$\lambda_k - \lambda_{k+1} \geq Ck^{-a-1}$ for $k \geq 1$ with $a > 1$.*

Condition 2. $|\alpha_{k0}| \leq Ck^{-b}$ *for* $k > 1$.

Condition 3. *For any* $C > 0$, *there exists an* $\epsilon > 0$ *such that*

$$\sup_{s\in\mathcal{I}}\Big\{\mathbb{E}|X(s)|^C\Big\} < \infty \quad \textit{and} \quad \sup_{s_1,s_2\in\mathcal{I}}\left\{\mathbb{E}\Big[|s_1-s_2|^{-\epsilon}|X(s_1)-X(s_2)|^C\Big]\right\} < \infty\,.$$

Condition 4. *For each integer* $d \geq 1$, $\lambda_k^{-d}\,\mathbb{E}\Big\{\int_{\mathcal{I}} X(s)\psi_k(s)\,ds\Big\}^{2d}$ *is bounded uniformly in* k.

Condition 5. $\mathbb{E}\Big[\exp\big\{C\|X\|\big\}\Big] < \infty$ *for any constant* $C > 0$, *where* $\|X\| = \Big\{\int_{\mathcal{I}} X^2(s)ds\Big\}^{1/2}$.

Condition 6. $X(\cdot)$ *belongs to a Donsker class.*

Condition 7. *The truncation number* $\widetilde{R} = \widetilde{R}_n$ *satisfies* $\widetilde{R}^{4a+4}n^{-1} \to 0$. *The decay rates satisfy* $b > a/2+1$. *The relationship between rates satisfies* $\widetilde{R}^{a/2+2-2b}\log(n) \to 0$.

Conditions on Survival Data and Model Structure

As in the main paper, let $\boldsymbol{Z}$ be the scalar covariates, T the failure time, and $\boldsymbol{V} = (V_1,\ldots,V_J)$ the observation times leading to censoring interval $(L_i, R_i]$. Also let $\Lambda_0(t)$ be the true baseline cumulative hazard function and $G(\cdot)$ the transformation function, and define $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}_{\widetilde{R}}^\top, \Lambda)^\top$.

Condition 8. *The domain of the scalar covariates* $\boldsymbol{Z}$ *is a bounded subset of* $\mathbb{R}^p$. *The matrix* $\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]$ *is positive definite.*

Condition 9. *The failure time* T *and the observation process* $\boldsymbol{V}$ *are independent given the covariates* $\boldsymbol{Z}$ *and* $X(s)$. *Furthermore, there exists a positive number* $\xi > 0$ *such that* $\mathbb{P}(V_{j+1} - V_j \geq \xi \mid \boldsymbol{Z}, X(s)) = 1$, *and the number of examination times* J *satisfies* $\mathbb{E}(J) < \infty$.

Condition 10. *The parameter spaces for* $\boldsymbol{\beta}$ *and* $\boldsymbol{\alpha}_{\widetilde{R}} = (\alpha_{1R},\ldots,\alpha_{\widetilde{R},R})^\top$ *are compact and convex. The true parameters* $\boldsymbol{\beta}_0$ *and* $\boldsymbol{\alpha}_{\widetilde{R},0}$ *are interior points. The parameter space for* Λ *is a collection of non-negative, non-decreasing, continuous functions on* $[u, v]$. *The true function* Λ_0 *is continuously differentiable up to order* r *in* $[u, v]$ *and satisfies* $C^{-1} < \Lambda_0(u) < \Lambda_0(v) < C$ *for some positive constant* C.

Condition 11. *The transformation function $G(\cdot)$ is strictly increasing with $G(0) = 0$ and is three times continuously differentiable on its domain relevant to the model.*

Condition 12. *The Fisher information matrix for the finite-dimensional parameters $\wp = (\boldsymbol{\beta}_a^\top, \boldsymbol{\alpha}_{\widetilde{R}}^\top)^\top$, denoted $\boldsymbol{I}(\wp_0)$, is non-singular.*

Condition 13. *For every $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, $\mathbb{P}\,\ell(\boldsymbol{\theta};\, O) - \mathbb{P}\,\ell(\boldsymbol{\theta}_0;\, O) \preceq -d^2(\boldsymbol{\theta},\, \boldsymbol{\theta}_0)$ and similarly for the log-likelihood $\widetilde{\ell}$ involving the full functional expansion. Here $\preceq$ means 'less than or equal to, up to a positive constant', and $d(\cdot,\cdot)$ is the distance metric.*

Web Appendix B: Proof of the Asymptotic Properties

Web Appendix B(1): Proof of Theorem 1

In this section, we will sketch the proof of Theorem 1 and in particular, the empirical process theory will be used to derive the consistency of the proposed sieve maximum likelihood estimator. Before the proof, we will repeat the definitions of some notation used in the main paper. Let the true model involve the infinite sum $\sum_{k=1}^{\infty} \zeta_{ik}\alpha_{k0}$. The corresponding (hypothetical) log-likelihood contribution is $\widetilde{\ell}(\boldsymbol{\theta}_\infty; O_i)$, where $\boldsymbol{\theta}_\infty$ includes the infinite sequence α_{k0}. The proposed estimation procedure uses the truncated log-likelihood contribution $\ell(\boldsymbol{\theta}; O_i)$ involving $\sum_{k=1}^{\widetilde{R}} \zeta_{ik}\alpha_k$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}_{\widetilde{R}}^\top, \Lambda)^\top$. Let Λ_n denote the Bernstein polynomial approximation of degree m for Λ and $\ell_n(\boldsymbol{\theta}; O) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; O_i)$. The sieve space is $\Theta_n = \{\boldsymbol{\theta} : \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\alpha}_{\widetilde{R}} \in \mathcal{G}_{\widetilde{R}}, \Lambda = \Lambda_n \in \mathcal{M}_n\}$, where $\mathcal{B}$ and $\mathcal{G}_{\widetilde{R}}$ are compact (Condition 10) and $\mathcal{M}_n$ is the space of Bernstein polynomials with bounded coefficients. By definition in Eq.(4) of the main paper, the estimator $\widehat{\boldsymbol{\theta}}_n$ maximizes the penalized objective function $H_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}; O) - \kappa p_{mic}(\boldsymbol{\beta})$, where κ is the tuning parameter and p_{mic} denotes the MIC penalty used in the paper, which is related to the L_0 norm. Note that $\widehat{\boldsymbol{\theta}}_n$ can equivalently be obtained by maximizing the normalized version of $H_n(\boldsymbol{\theta})$

$$\overline{H}_n(\boldsymbol{\theta}) = n^{-1} H_n(\boldsymbol{\theta}) = \mathbb{P}_n \ell(\boldsymbol{\theta}; O) - \kappa_n p_{\text{mic}}(\boldsymbol{\beta})\,,$$

where $\kappa_n = \kappa/n$. Let $\boldsymbol{\theta}_0$ represent the true parameter projected onto the truncated model space, effectively meaning $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_{k0}$ for $k \leq \widetilde{R}$, along with the true Λ_0. The bias

term from truncation is $e_i = \sum_{k=\widetilde{R}+1}^{\infty} \zeta_{ik}\alpha_{k0}$.

Before developing the asymptotic property for the proposed estimator, we first show that it is a sieve maximum likelihood estimator under certain conditions (Liu et al., 2026). For this, denote $\widehat{\boldsymbol{\theta}}_n^*$ as the sieve maximum likelihood estimator obtained from Section 2 by maximizing $\mathbb{P}_n\ell(\boldsymbol{\theta}; O)$. From the definition, we have that $\widehat{\boldsymbol{\theta}}_n$ maximizes $\overline{H}_n(\boldsymbol{\theta})$. Thus, we can get that $\mathbb{P}_n\ell(\widehat{\boldsymbol{\theta}}_n; O) \geq \mathbb{P}_n\ell(\widehat{\boldsymbol{\theta}}_n^*; O) - \epsilon_n$, where

$$\epsilon_n = \kappa_n \left| -p_{mic}(\widehat{\boldsymbol{\beta}}_n) + p_{mic}(\widehat{\boldsymbol{\beta}}_n^*) \right| \to 0$$

when κ_n converges to 0. That is, $\widehat{\boldsymbol{\theta}}_n$ is asymptotically equivalent to a sieve maximum likelihood estimator. Similar arguments can also be applied to $\mathbb{P}_n\widetilde{\ell}(\boldsymbol{\theta}; O) - \kappa_n p_{mic}(\boldsymbol{\beta})$. Furthermore, since $\kappa_n \to 0$ when $n \to \infty$ and the penalty function p_{mic} (related to the number of non-zero elements) is bounded for a fixed p, the penalty term vanishes asymptotically. Thus, the proof for the consistency can be concentrated on the properties of the unpenalized log-likelihood function $\ell(\boldsymbol{\theta}; O)$ instead.

Proof of Consistency (Theorem 1(i))

We follow Theorem 5.7 of van der Vaart (2000) to verify the three conditions for $\mathbb{M}_n(\boldsymbol{\theta}) = \mathbb{P}_n\ell(\boldsymbol{\theta}; O)$ and $\mathbb{M}(\boldsymbol{\theta}) = \mathbb{P}\ell(\boldsymbol{\theta}; O)$.

(i) $\sup_{\boldsymbol{\theta}\in\Theta_n} |\mathbb{M}_n(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{\theta})| \xrightarrow{p} 0$.

We establish this by showing that the function class $\mathcal{F}_n = \{\ell(\boldsymbol{\theta}; O) : \boldsymbol{\theta} \in \Theta_n\}$ is Glivenko-Cantelli. The key is to bound the $L_1(\mathbb{P}_n)$ covering number $N(\epsilon, \mathcal{F}_n, L_1(\mathbb{P}_n))$. For this, we first establish a Lipschitz-like continuity for $\ell(\boldsymbol{\theta}; O)$. Note that the derivatives of $S(\cdot)$ w.r.t $\boldsymbol{x}^\top\boldsymbol{W}$ are bounded based on Conditions 8, 10 and 11. Then it follows from the use of the mean value theorem for the function $F(\boldsymbol{x}) = \log\{S(L \mid \boldsymbol{x}) - S(R \mid \boldsymbol{x})\}$, we have that

$$|\ell(\boldsymbol{\theta}^{(1)}; O) - \ell(\boldsymbol{\theta}^{(2)}; O)| \leq C\, d\,(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$$

for $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\} \in \Theta_n$, where $S(t \mid \boldsymbol{x}) = \exp[-G\{\Lambda(t)\exp(\boldsymbol{x}^\top\boldsymbol{W})\}]$. As shown in the

sketch :

$$d\left(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\right) \approx \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\| + \|\boldsymbol{\alpha}_{\widetilde{R}}^{(1)} - \boldsymbol{\alpha}_{\widetilde{R}}^{(2)}\| + \|\Lambda^{(1)} - \Lambda^{(2)}\|_{\infty}$$
$$\leq \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\| + \|\boldsymbol{\alpha}_{\widetilde{R}}^{(1)} - \boldsymbol{\alpha}_{\widetilde{R}}^{(2)}\| + \max_{0 \leq j \leq m} |\phi^{j(1)} - \phi^{j(2)}| \,.$$

Let d_{Θ} be the sum of Euclidean norms for $\boldsymbol{\beta}, \boldsymbol{\alpha}_{\widetilde{R}}$ and the max norm for $\boldsymbol{\phi}$. The parameter space Θ_n is contained within a product of compact sets (Condition 10). The covering number $N(\epsilon, \Theta_n, d_{\Theta})$ is bounded by the product of covering numbers for each component space. For Euclidean balls of dimension p and $\widetilde{R}$, the covering number grows as $(M/\epsilon)^p$ and $(M/\epsilon)^{\widetilde{R}}$. For the Bernstein coefficients $\boldsymbol{\phi} \in \mathbb{R}^{m+1}$ with $\sum |\phi^j| \leq M_n$, the covering number grows as $(M_n/\epsilon)^{m+1}$ (van de Geer, 2000). Therefore, $N(\epsilon, \Theta_n, d_{\Theta}) \leq C_1 (M/\epsilon)^p (M/\epsilon)^{\widetilde{R}} (M_n/\epsilon)^{m+1}$. Since ℓ is Lipschitz w.r.t d (with potentially random coefficient C), the covering number for $\mathcal{F}_n$ is bounded such that

$$N(\epsilon, \mathcal{F}_n, L_1(\mathbb{P}_n)) \leq N(\epsilon/(\mathbb{E}[C] + 1), \Theta_n, d_{\Theta})$$

by using the standard arguments (e.g., van der Vaart & Wellner Thm 2.7.11). Thus we have that

$$\log N(\epsilon, \mathcal{F}_n, L_1(\mathbb{P}_n)) \leq C + (p + \widetilde{R} + m + 1) \log(1/\epsilon) + (m+1) \log M_n \,.$$

For the uniform convergence, we need that $\log N(\epsilon, \mathcal{F}_n, L_1(\mathbb{P}_n))/n \to 0$ for all $\epsilon > 0$. This requires $\{(p + \widetilde{R} + m) \log n\}/n \to 0$. Since p is fixed (or $p = p_n$ grows slowly), this holds if $\widetilde{R} \log n/n \to 0$ and $m \log n/n \to 0$. Futhermore, Condition 7 implies $\widetilde{R}/n^{1/5} \to 0$, which is sufficient.

(ii) $\sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > \epsilon} \mathbb{M}(\boldsymbol{\theta}) < \mathbb{M}(\boldsymbol{\theta}_0)$.

First note that the Gibbs inequality implies that $\sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > \epsilon} \mathbb{M}(\boldsymbol{\theta}) \leq \mathbb{M}(\boldsymbol{\theta}_0)$. Assume that $\sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > \epsilon} \mathbb{M}(\boldsymbol{\theta}) = \mathbb{M}(\boldsymbol{\theta}_0)$. Then there exists a sequence $\boldsymbol{\theta}_m$ such that

$$\mathbb{M}(\boldsymbol{\theta}_m) \to \sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > \epsilon} \mathbb{M}(\boldsymbol{\theta})$$

and $d(\boldsymbol{\theta}_m, \boldsymbol{\theta}_0) > \epsilon$. By Conditions 10, 11 and 13, there exists a subsequence $\boldsymbol{\theta}_{\widetilde{m}}$

converging to $\boldsymbol{\theta}_{m0}$. Since $\mathbb{M}(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, $\mathbb{M}(\boldsymbol{\theta}_{m0}) = \mathbb{M}(\boldsymbol{\theta}_0)$ and consequently $\boldsymbol{\theta}_{m0} = \boldsymbol{\theta}_0$ according to the identifiability of the proposed model. However, $\boldsymbol{\theta}_{\widetilde{m}}$ does not converge to $\boldsymbol{\theta}_0$ due to the fact $d\left(\boldsymbol{\theta}_{\widetilde{m}}, \boldsymbol{\theta}_0\right) > \epsilon$. This conflicts with the aforementioned results that $\boldsymbol{\theta}_{\widetilde{m}}$ converges to $\boldsymbol{\theta}_{m0}$. Therefore, we have that $\sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > \epsilon} \mathbb{M}(\boldsymbol{\theta}) < \mathbb{M}(\boldsymbol{\theta}_0)$.

(iii) $\mathbb{M}_n(\widehat{\boldsymbol{\theta}}_n) \geq \mathbb{M}_n(\boldsymbol{\theta}_0) - o_p(1)$.

Let $\boldsymbol{\theta}_{0,n} = (\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{\widetilde{R},0}, \Lambda_n)$, where Λ_n approximates Λ_0. By optimality of $\widehat{\boldsymbol{\theta}}_n$ for H_n and noting the penalty term is $o_p(1)$ evaluated at $\boldsymbol{\theta}_{0,n}$ and $\hat{\boldsymbol{\theta}}_n$, we have that $\mathbb{M}_n(\widehat{\boldsymbol{\theta}}_n) \gtrsim \mathbb{M}_n(\boldsymbol{\theta}_{0,n})$. We need that $\mathbb{M}_n(\boldsymbol{\theta}_{0,n}) - \mathbb{M}_n(\boldsymbol{\theta}_0) \geq -o_p(1)$. Decompose $A_n = (\mathbb{P}_n - \mathbb{P})(\ell(\boldsymbol{\theta}_{0,n}) - \ell(\boldsymbol{\theta}_0))$ and $B_n = \mathbb{P}(\ell(\boldsymbol{\theta}_{0,n}) - \ell(\boldsymbol{\theta}_0))$.

For B_n, based on Conditions 10 and 11 and using Taylor expansion around $\boldsymbol{\theta}_0$ and the rate $||\Lambda_n - \Lambda_0||_\infty = O(m^{-r/2})$, we have that $B_n = O(d^2(\boldsymbol{\theta}_{0,n}, \boldsymbol{\theta}_0)) = O(m^{-r}) = O(n^{-r\nu}) = o(1)$. For A_n, define the class $\mathcal{F}_{\widetilde{n}} = \{\ell(\boldsymbol{\theta}; O) - \ell(\boldsymbol{\theta}_{0,\widetilde{n}}; O) : \boldsymbol{\theta} \in \Theta_n, d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_{0,\widetilde{n}}) \leq \delta_n\}$, where $\delta_n \to 0$. We need to show that this class is $\mathbb{P}$-Donsker. This requires bounding the bracketing numbers $N_{[]}(\epsilon, \mathcal{F}_{\widetilde{n}}, L_2(\mathbb{P}))$. The size of the brackets depends on the modulus of continuity of $\ell(\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$. Similar to the covering number argument, this depends on the dimensions $\{p, \widetilde{R}, m\}$ and smoothness (Condition 11). The calculation $J_{[]}(\delta, \mathcal{F}_{\widetilde{n}}, L_2(\mathbb{P})) < \infty$ confirms the Donsker property. By Theorem 19.5 of van der Vaart (2000) and Corollary 2.3.12 of van der Vaart and Wellner (1996) , we have that $A_n = O_p(n^{-1/2}) = o_p(1)$. Thus one can conlude that $\mathbb{M}_n(\widehat{\boldsymbol{\theta}}_n) \geq \mathbb{M}_n(\boldsymbol{\theta}_{0,n}) - o_p(1) \geq \mathbb{M}_n(\boldsymbol{\theta}_0) + B_n - A_n - o_p(1) = \mathbb{M}_n(\boldsymbol{\theta}_0) - o_p(1)$.

Consistency of $\widehat{\gamma}(s)$: Decompose the error $\|\widehat{\alpha} - \alpha_0\|_{L_2}$ into three parts as

(i) Error from estimating coefficients $\sum_{k=1}^{\widetilde{R}} \lambda_k^{1/2} \|\widehat{\alpha}_{kR} - \alpha_{kR,0}\|$. This term is $O_p(\alpha_n \cdot (\sum \lambda_k)^{1/2}) = o_p(1)$ since $\alpha_n \to 0$ and $\sum \lambda_k < \infty$.

(ii) Error from estimating eigenfunctions $\sum_{k=1}^{\widetilde{R}} |\alpha_{k0}| \|\widehat{\boldsymbol{\psi}}_k - \boldsymbol{\psi}_k\|$. Using Lemma 2 from Kong et al. (2018) or Hall and Hosseini-Nasab (2006), $\|\widehat{\boldsymbol{\psi}}_k - \boldsymbol{\psi}_k\| \leq C\delta_k^{-1} \|\widehat{\boldsymbol{K}} - \boldsymbol{K}\|$. Using Lemma 1 from Kong et al. (2018) , $\|\widehat{\boldsymbol{K}} - \boldsymbol{K}\| = O_p(n^{-1/2})$. The sum is

bounded based on Condition 1 ($\delta_k^{-1} \leq Ck^{a+1}$) and Condition 2 ($|\alpha_{k0}| \leq Ck^{-b}$). The sum is $O_p(n^{-1/2}\sum_{k=1}^{\widetilde{R}} k^{a+1-b})$. This converges to 0 if $b > a+2$, which is guaranteed by Condition 7.

(iii) Truncation bias $(\sum_{k=\widetilde{R}+1}^{\infty} \alpha_{k0}^2)^{1/2}$. It follows from Condition 2 that

$$O\{(\sum_{k=\widetilde{R}+1}^{\infty} k^{-2b})^{1/2}\} = O(\widetilde{R}^{-b+1/2}),$$

which goes to 0 as $\widetilde{R} \to \infty$ since $b > 1/2$.

This proves that $\|\widehat{\gamma}(\cdot) - \gamma_0(\cdot)\|_{L_2} \xrightarrow{p} 0$.

Proof of Sparsity (Theorem 1(ii))

In this part, we show that $\widehat{\boldsymbol{\beta}}_b = \mathbf{0}$ with probability tending to one. It suffices to verify that

$$H_n\left(\left[\widehat{\boldsymbol{\beta}}_a, \mathbf{0}, \widehat{\gamma}(\cdot)\right]^\top, \Lambda_0\right) = \max_{\|\boldsymbol{\beta}_b\| \leq Cn^{-1/2}} H_n\left(\left[\widehat{\boldsymbol{\beta}}_a, \boldsymbol{\beta}_b, \widehat{\gamma}(\cdot)\right]^\top, \Lambda_0\right).$$

To this end, consider β_j for $j > s_\beta$ and thus $\beta_{j0} = 0$. We examine the derivative of H_n with respect to β_j at a point $\boldsymbol{\theta}$ in the local neighborhood $\|\boldsymbol{\beta}_b\| \leq Cn^{-1/2}$ with $\beta_j \neq 0$.

Note that

$$\frac{\partial H_n(\boldsymbol{\theta})}{\partial \beta_j} = \frac{\partial \ell_n(\boldsymbol{\theta}; O)}{\partial \beta_j} - \kappa \frac{\partial p_{\text{mic}}(\boldsymbol{\beta})}{\partial \beta_j}.$$

For the likelihood score, a Taylor expansion around $\beta_{j0} = 0$ gives

$$\frac{\partial \ell_n(\boldsymbol{\theta}; O)}{\partial \beta_j} = \frac{\partial \ell_n(\boldsymbol{\theta}_0; O)}{\partial \beta_j} + \frac{\partial^2 \ell_n(\boldsymbol{\theta}^*; O)}{\partial \beta_j^2} \beta_j,$$

and

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0; O)}{\partial \beta_j} = O_p(1), \qquad \frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}^*; O)}{\partial \beta_j^2} = O_p(1),$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Therefore we have that

$$\frac{\partial \ell_n(\boldsymbol{\theta}; O)}{\partial \beta_j} = O_p(\sqrt{n}).$$

On the other hand, for $|\beta_j|$ in the shrinking neighborhood of zero, the MIC penalty

satisfies

$$\frac{\partial p_{\text{mic}}(\boldsymbol{\beta})}{\partial \beta_j} = c\sqrt{n}\ \text{sgn}(\beta_j)\{1 + o(1)\}$$

for some constant $c > 0$. Hence,

$$\kappa\frac{\partial p_{\text{mic}}(\boldsymbol{\beta})}{\partial \beta_j} = c\kappa\sqrt{n}\ \text{sgn}(\beta_j)\{1 + o(1)\}\,.$$

With $\kappa = \log(n_0)$ and $n_0 = O_p(n)$, we have $\kappa \to \infty$. Hence, the penalty derivative dominates the likelihood score in this local neighborhood. Consequently, with probability tending to one,

$$\text{sgn}\left\{\frac{\partial H_n(\boldsymbol{\theta})}{\partial \beta_j}\right\} = -\,\text{sgn}(\beta_j)\,,$$

which implies that

$$\frac{\partial H_n(\boldsymbol{\theta})}{\partial \beta_j} < 0 \quad \text{when } \beta_j > 0\,, \qquad \frac{\partial H_n(\boldsymbol{\theta})}{\partial \beta_j} > 0 \quad \text{when } \beta_j < 0\,.$$

Thus, H_n decreases as $|\beta_j|$ moves away from zero and the maximum is attained at $\boldsymbol{\beta}_b = \mathbf{0}$. This establishes the sparsity property.

Proof of Asymptotic Normality (Theorem 1(iii))

Let $\boldsymbol{\varphi} = (\boldsymbol{\beta}_a^\top, \boldsymbol{\alpha}_{\widetilde{R}}^\top)^\top$ and $\widehat{\boldsymbol{\varphi}}_n$ and $\widetilde{\boldsymbol{\varphi}}_n$ be the penalized and unpenalized (sieve MLE) estimators, respectively.

(a) First we show that $\sqrt{n}(\widehat{\boldsymbol{\varphi}}_n - \widetilde{\boldsymbol{\varphi}}_n) = o_p(1)$.

By expanding the score equation $\nabla_\varphi H_n(\widehat{\boldsymbol{\varphi}}_n, \widehat{\Lambda}_n) = \mathbf{0}$ around $\widetilde{\boldsymbol{\varphi}}_n$, we have that

$$\begin{aligned}\mathbf{0} &= \nabla_\varphi \ell_n(\widehat{\boldsymbol{\varphi}}_n, \widehat{\Lambda}_n) - \kappa\nabla_\varphi p_{\text{mic}}(\widehat{\boldsymbol{\beta}}_n) \\ &\approx \underbrace{\nabla_\varphi \ell_n(\widetilde{\boldsymbol{\varphi}}_n, \widehat{\Lambda}_n)}_{=\mathbf{0}} + \nabla^2_{\varphi\varphi}\ell_n(\boldsymbol{\varphi}^*, \widehat{\Lambda}_n)(\widehat{\boldsymbol{\varphi}}_n - \widetilde{\boldsymbol{\varphi}}_n) - \kappa\nabla_\varphi p_{\text{mic}}(\widehat{\boldsymbol{\beta}}_n)\,.\end{aligned}$$

This gives that

$$\left(-\frac{1}{n}\nabla^2_{\varphi\varphi}\ell_n(\boldsymbol{\varphi}^*, \widehat{\Lambda}_n)\right)\sqrt{n}(\widehat{\boldsymbol{\varphi}}_n - \widetilde{\boldsymbol{\varphi}}_n) \approx \frac{\kappa}{\sqrt{n}}\nabla_\varphi p_{\text{mic}}(\widehat{\boldsymbol{\beta}}_n)\,.$$

By following similar arguments on the penalty derivative of $p_{\text{mic}}(\cdot)$ as before and noting that $\kappa = o(\sqrt{n})$ and the active coefficients are bounded away from zero, the

right-hand side is $o_p(1)$. Based on Conditions 8, 10-12, the Hessian term converges to $\boldsymbol{I}_1(\boldsymbol{\varphi}_0)$. This implies that $\sqrt{n}(\widehat{\boldsymbol{\varphi}}_n - \widetilde{\boldsymbol{\varphi}}_n) = o_p(1)$.

(b) Now we show that $\sqrt{n}(\widetilde{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0) \Rightarrow N(\mathbf{0}, \boldsymbol{\Sigma})$.

This follows from the theory of semiparametric M-estimation on sieve spaces. We use the Hilbert space framework with directional derivatives $\dot{\ell}, \ddot{\ell}$ and Fisher inner product $< \cdot, \cdot >$ defined as in Zhou et al. (2017) and Zhao et al. (2020). The space $\overline{\Gamma}$ is the tangent space. The functional $h(\boldsymbol{\theta}) = \boldsymbol{b}^\top \boldsymbol{\varphi}$ is smooth . By Riesz representation, $\dot{h}(\boldsymbol{\theta}_0)[\boldsymbol{\iota}] =< \boldsymbol{\iota}^*, \boldsymbol{\iota} >$ for some $\boldsymbol{\iota}^* \in \overline{\Gamma}$. General theorems for sieve M-estimators (e.g., Shen (1997); see also application in Chen et al. (2006) or Zhao et al. (2020) appendix) state that under appropriate conditions (smoothness, entropy/Donsker conditions for the class, identification, well-behaved information matrix that relying on Conditions 3-6 and 8-13), we have that

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \approx \mathbb{G}_n(\boldsymbol{\ell}^*) + o_p(1)\,,$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ is the empirical process and $\boldsymbol{\ell}^*$ is related to the efficient score function for $\boldsymbol{\theta}$ in the Hilbert space $\overline{\Gamma}$. Specifically for the parameter of interest $\boldsymbol{\varphi}$:

$$\sqrt{n}(\widetilde{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0) \Rightarrow N(\mathbf{0}, \boldsymbol{\Omega})$$

where $\boldsymbol{\Omega} = [\boldsymbol{I}(\boldsymbol{\varphi}_0)]^{-1}$ is the inverse of the efficient Fisher information matrix for $\boldsymbol{\varphi}$. $\boldsymbol{I}(\boldsymbol{\varphi}_0)$ is obtained by considering the scores for $\boldsymbol{\varphi}$ and projecting out the influence of the infinite-dimensional nuisance parameter Λ . The matrix $\boldsymbol{\Sigma}$ in the theorem statement is the submatrix of $\boldsymbol{\Omega}$ corresponding to the non-zero components $\boldsymbol{\beta}_a$ and $\boldsymbol{\alpha}_{\widetilde{R}}$.

Web Appendix B(2): Proof of Theorem 2

Before the proof, we need to give another condition.

Condition 14. *The parametric model $\pi(\wp)$ for the missingness mechanism is continuously differentiable in $\wp$, and there exists a constant $\delta > 0$ such that $\pi(\wp) \geq \delta$ almost*

surely. The estimator $\widehat{\boldsymbol{\wp}}_n$ *satisfies* $\widehat{\boldsymbol{\wp}}_n \to \boldsymbol{\wp}_0$ *and*

$$\sqrt{n}(\widehat{\boldsymbol{\wp}}_n - \boldsymbol{\wp}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varsigma_i + o_p(1)\,,$$

where $\boldsymbol{\wp}_0$ *is the true value of* $\boldsymbol{\wp}$*,* $\mathbb{E}(\varsigma_i) = 0$*, and* $\mathbb{E}\|\varsigma_i\|^2 < \infty$*.*

Proof of Consistency under IPW (Theorem 2(i))

We first define the reweighted log-likelihood function

$$\ell^{[1]}_{\boldsymbol{\wp}}(\boldsymbol{\theta};\, O) = \varpi(\boldsymbol{\wp})\, \ell(\boldsymbol{\theta};\, O)$$

for a generic observation $O \in \mathcal{X}$, where $\ell(\boldsymbol{\theta};\, O)$ is defined as before, and $\varpi(\boldsymbol{\wp}) = \chi/\pi(\boldsymbol{\wp})$ is the inverse probability weight. The IPW estimator $\widehat{\boldsymbol{\theta}}^{[1]}_n$ is obtained by maximizing the penalized objective function

$$H^{\text{IPW}}_n(\boldsymbol{\theta}) = \ell^{[1]}_n(\boldsymbol{\theta}; O) - \kappa p_{mic}(\boldsymbol{\beta})\,, \qquad \text{where} \qquad \ell^{[1]}_n(\boldsymbol{\theta}; O) = \sum_{i=1}^{n} \ell^{[1]}_{\boldsymbol{\wp},i}(\boldsymbol{\theta}; O_i)\,.$$

Following the same argument as in Theorem 1, the consistency of $\widehat{\boldsymbol{\theta}}^{[1]}_n$ is implied by the properties of the unpenalized weighted log-likelihood $\ell^{[1]}_{\boldsymbol{\wp}}(\boldsymbol{\theta})$.

Let $\mathbb{M}^{[1]}_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\wp}}_n) = \mathbb{P}_n \ell^{[1]}_{\widehat{\boldsymbol{\wp}}_n}(\boldsymbol{\theta}; O)$, $\mathbb{M}_n(\boldsymbol{\theta}) = \mathbb{P}_n \ell(\boldsymbol{\theta}; O)$ and $\mathbb{M}(\boldsymbol{\theta}) = \mathbb{P}\ell(\boldsymbol{\theta}; O)$. We verify the three conditions of Theorem 5.7 of van der Vaart (2000).

(i) $\sup_{\boldsymbol{\theta}\in\Theta_n} |\mathbb{M}^{[1]}_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\wp}}_n) - \mathbb{M}(\boldsymbol{\theta})| \xrightarrow{p} 0$.

We decompose this difference as

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta_n} |\mathbb{M}^{[1]}_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\wp}}_n) - \mathbb{M}(\boldsymbol{\theta})| &\le \|\mathbb{M}^{[1]}_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\wp}}_n) - \mathbb{M}^{[1]}_n(\boldsymbol{\theta}, \boldsymbol{\wp}_0)\|_{\Theta_n} \\ &\quad + \|\mathbb{M}^{[1]}_n(\boldsymbol{\theta}, \boldsymbol{\wp}_0) - \mathbb{M}_n(\boldsymbol{\theta})\|_{\Theta_n} + \|\mathbb{M}_n(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{\theta})\|_{\Theta_n} \\ &=: I_{1,n} + I_{2,n} + I_{3,n}\,.\end{aligned}$$

By Condition 14 and a first-order Taylor expansion, we obtain that

$$I_{1,n} \le \|\{\varpi(\widehat{\boldsymbol{\wp}}_n) - \varpi(\boldsymbol{\wp}_0)\}\, \ell(\boldsymbol{\theta}; O)\| \le C\, |\varpi(\widehat{\boldsymbol{\wp}}_n) - \varpi(\boldsymbol{\wp}_0)| \to 0\,.$$

For $I_{2,n}$, since

$$I_{2,n} \leq \|\{\varpi(\wp_0) - 1\}\, \ell(\boldsymbol{\theta}; O)\| \leq C \cdot |\mathbb{E}\{\varpi(\wp_0) - 1\}| + o_p(1)\,,$$

and

$$\mathbb{E}\{\varpi(\wp_0) - 1\} = \mathbb{E}\left\{\frac{\chi_0 - \pi(\wp_0)}{\pi(\wp_0)}\right\} = \mathbb{E}\left\{\frac{\mathbb{E}(\chi_0 \mid O) - \pi(\wp_0)}{\pi(\wp_0)}\right\},$$

it follows that $I_{2,n} \to 0$ if $\pi(\wp_0)$ is correctly specified. The term $I_{3,n}$ is identical to that in Theorem 1 and thus converges to zero as shown earlier. Therefore, condition (i) holds.

(ii) $\sup_{\boldsymbol{\theta}: d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > \epsilon} \mathbb{M}(\boldsymbol{\theta}) < \mathbb{M}(\boldsymbol{\theta}_0)$.

This condition is identical to that of Theorem 1 and has already been verified. We omit it here.

(iii) $\mathbb{M}_n^{[1]}(\widehat{\boldsymbol{\theta}}_n, \widehat{\wp}_n) \geq \mathbb{M}_n(\boldsymbol{\theta}_0) - o_p(1)$.

Using a Taylor expansion and Conditions 8 and 14, we have that

$$\mathbb{M}_n^{[1]}(\widehat{\boldsymbol{\theta}}_n, \widehat{\wp}_n) = \mathbb{M}_n^{[1]}(\widehat{\boldsymbol{\theta}}_n, \wp_0) + o_p(1)\,.$$

Then we can decompose

$$\begin{aligned}\mathbb{M}_n^{[1]}(\widehat{\boldsymbol{\theta}}_n, \wp_0) - \mathbb{M}_n^{[1]}(\boldsymbol{\theta}_0, \wp_0) &= \left\{\mathbb{M}_n^{[1]}(\widehat{\boldsymbol{\theta}}_n, \wp_0) - \mathbb{M}_n^{[1]}(\boldsymbol{\theta}_{0,n}, \wp_0)\right\} \\ &\quad + \left\{\mathbb{M}_n^{[1]}(\boldsymbol{\theta}_{0,n}, \wp_0) - \mathbb{M}_n^{[1]}(\boldsymbol{\theta}_0, \wp_0)\right\} \\ &\geq \mathbb{M}_n^{[1]}(\boldsymbol{\theta}_{0,n}, \wp_0) - \mathbb{M}_n^{[1]}(\boldsymbol{\theta}_0, \wp_0)\,,\end{aligned}$$

and express

$$\begin{aligned}\mathbb{M}_n^{[1]}(\boldsymbol{\theta}_{0,n}, \wp_0) - \mathbb{M}_n^{[1]}(\boldsymbol{\theta}_0, \wp_0) &= \{\mathbb{P}_n - \mathbb{P}\}\left[\ell^{[1]}(\boldsymbol{\theta}_{0,n}, \wp_0) - \ell^{[1]}(\boldsymbol{\theta}_0, \wp_0)\right] \\ &\quad + \mathbb{P}\left[\ell^{[1]}(\boldsymbol{\theta}_{0,n}, \wp_0) - \ell^{[1]}(\boldsymbol{\theta}_0, \wp_0)\right].\end{aligned}$$

The two terms on the right-hand side are analogous to the empirical fluctuation term A_n and the expected change B_n in the proof of Theorem 1(iii). The same arguments apply here, and thus condition (iii) holds.

Therefore, all three conditions of Theorem 5.7 of van der Vaart (2000) are satisfied, and the consistency of $\widehat{\gamma}(s)$ follows by the same argument as before. This establishes the consistency of $\widehat{\boldsymbol{\theta}}_n^{[1]}$.

Proof of Sparsity under IPW (Theorem 2(ii))

For the proof, we consider the penalized IPW estimator $\widehat{\boldsymbol{\theta}}_n^{[1]}$ obtained by maximizing $H_n^{\mathrm{IPW}}(\boldsymbol{\theta})$. For a noise variable $j > s_\beta$ with true coefficient $\beta_{j0} = 0$, we examine the derivative of the objective function

$$\frac{\partial H_n^{\mathrm{IPW}}(\boldsymbol{\theta})}{\partial \beta_j} = \frac{\partial \ell_n^{[1]}(\boldsymbol{\theta}; O)}{\partial \beta_j} - \kappa \frac{\partial p_{\mathrm{mic}}(\boldsymbol{\beta})}{\partial \beta_j} .$$

Then following similar arguments from the sparsity proof of Theorem 1 and using the fact that the estimated weights ϖ_i's satisfy a first-order expansion under Condition 14, we have that

$$\frac{\partial \ell_n^{[1]}(\boldsymbol{\theta}; O)}{\partial \beta_j} = O_p(\sqrt{n}) .$$

Note that the penalty contributes $\kappa\sqrt{n}\,\mathrm{sgn}(\beta_j)$. Therefore for sufficiently small $|\beta_j| \neq 0$, the penalty dominates the likelihood component and drives the derivative in the opposite direction of β_j, indicating that the penalized objective $H_n^{\mathrm{IPW}}(\boldsymbol{\theta})$ is maximized at $\beta_j = 0$. Hence the IPW estimator $\widehat{\boldsymbol{\theta}}_n^{[1]}$ satisfies the sparsity property.

Proof of Asymptotic Normality under IPW (Theorem 2(iii))

Let $\boldsymbol{\varphi} = (\boldsymbol{\beta}_a^\top, \boldsymbol{\alpha}_{\widetilde{R}}^\top)^\top$ denote the vector of active coefficients and $\widehat{\boldsymbol{\varphi}}_n$ and $\widetilde{\boldsymbol{\varphi}}_n$ the penalized and unpenalized (sieve MLE) estimators under IPW, respectively. As shown in Theorem 1(iii), the penalty has a negligible asymptotic impact such that

$$\sqrt{n}(\widehat{\boldsymbol{\varphi}}_n - \widetilde{\boldsymbol{\varphi}}_n) = o_p(1) ,$$

and hence it suffices to establish the asymptotic normality of $\widetilde{\boldsymbol{\varphi}}_n$.

We apply Theorem 5.23 of van der Vaart (2000) with the identification $Z = O$ and $m_{\boldsymbol{\theta}}(Z) = \ell^{[1]}(\boldsymbol{\theta}; O)$. Under the regularity conditions, the map $r \mapsto m_{\boldsymbol{\theta}}(r)$ is measurable, and $\boldsymbol{\theta} \mapsto m_{\boldsymbol{\theta}}(r)$ is differentiable for all r. Since the weight function $\varpi(\boldsymbol{\wp}_0)$ is bounded,

and both $\boldsymbol{\theta}$ and O lie in compact spaces, each element of the gradient $\dot{m}_{\boldsymbol{\theta}}(r) = \partial m_{\boldsymbol{\theta}}(r)/\partial \boldsymbol{\theta}$ is uniformly bounded by some constant C. It then follows from the mean value theorem and the Cauchy-Schwarz inequality that

$$|m_{\boldsymbol{\theta}_1}(r) - m_{\boldsymbol{\theta}_2}(r)| = |\dot{m}_{\boldsymbol{\theta}^*}(r)^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)| \leq C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\,,$$

where $\boldsymbol{\theta}^*$ lies on the line segment between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Thus $\mathbb{E}\,\dot{m}^2(r) < \infty$ and all elements of $\partial m_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}$ and $\partial^2 m_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top$ are dominated by integrable functions. Therefore, the map $\boldsymbol{\theta} \mapsto \mathbb{E}\, m_{\boldsymbol{\theta}}$ admits a second-order Taylor expansion.

From consistency and regularity conditions, it follows that

$$\sqrt{n}(\widetilde{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0) = I^{-1}(\boldsymbol{\varphi}_0) \cdot n^{-1/2} \sum_{i=1}^{n} \varpi_{0,i} \cdot \frac{\partial \ell}{\partial \boldsymbol{\varphi}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(1)\,,$$

where $I(\boldsymbol{\varphi}_0)$ is the Fisher information matrix under the IPW likelihood, and $\varpi_{0,i} = \varpi_i(\boldsymbol{\wp}_0)$. Moreover, since $\widehat{\boldsymbol{\wp}}_n$ is an asymptotically linear estimator for $\boldsymbol{\wp}_0$, the joint asymptotic distribution satisfies

$$\sqrt{n}\begin{bmatrix} \widehat{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0 \\ \widehat{\boldsymbol{\wp}}_n - \boldsymbol{\wp}_0 \end{bmatrix} \longrightarrow \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma}^* & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right),$$

where $\boldsymbol{\Sigma}_{22}$ is the asymptotic variance of $\widehat{\boldsymbol{\wp}}_n$ and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top$ is the cross-covariance between $\widehat{\boldsymbol{\varphi}}_n$ and $\widehat{\boldsymbol{\wp}}_n$.

By the first-order condition of the penalized likelihood, we have that

$$n^{-1/2} \sum_{i=1}^{n} \varpi_i(\widehat{\boldsymbol{\wp}}_n) \cdot \frac{\partial \ell}{\partial \boldsymbol{\varphi}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} = 0\,.$$

A Taylor expansion of this score function around $\boldsymbol{\theta}_0$ yields that

$$n^{-1/2} \sum_{i=1}^{n} \varpi_i(\widehat{\boldsymbol{\wp}}_n) \cdot \frac{\partial \ell}{\partial \boldsymbol{\varphi}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + A_n \cdot \sqrt{n}(\widehat{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0) = 0\,,$$

where

$$A_n = \frac{1}{n} \sum_{i=1}^{n} \varpi_i(\widehat{\boldsymbol{\wp}}_n) \cdot \frac{\partial^2 \ell}{\partial \boldsymbol{\varphi}\partial \boldsymbol{\varphi}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(1)\,.$$

Since $\partial^2\ell/\partial\boldsymbol{\varphi}\partial\boldsymbol{\varphi}^\top$ is uniformly bounded, we have that

$$A_n = \frac{1}{n}\sum_{i=1}^{n} \varpi_i(\boldsymbol{\wp}_0)\cdot \frac{\partial^2\ell}{\partial\boldsymbol{\varphi}\partial\boldsymbol{\varphi}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(1) = I(\boldsymbol{\varphi}_0) + o_p(1)\,.$$

Hence

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0) = -I^{-1}(\boldsymbol{\varphi}_0)\cdot n^{-1/2}\sum_{i=1}^{n}\varpi_i(\boldsymbol{\wp}_0)\cdot\frac{\partial\ell}{\partial\boldsymbol{\varphi}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\\ - I^{-1}(\boldsymbol{\varphi}_0)\cdot B(\boldsymbol{\varphi}_0,\boldsymbol{\wp}_0)\cdot\sqrt{n}(\widehat{\boldsymbol{\wp}}_n - \boldsymbol{\wp}_0) + o_p(1)\,,\end{aligned}$$

where

$$B(\boldsymbol{\varphi},\boldsymbol{\wp}) = \mathbb{E}\left[\frac{\partial\ell}{\partial\boldsymbol{\varphi}}\cdot\left(\frac{\partial\,\varpi_i(\boldsymbol{\wp})}{\partial\boldsymbol{\wp}}\right)^\top\right].$$

Thus from the joint asymptotic linearity and by applying Theorem 1 of Pierce (1982), we conclude that

$$\sqrt{n}(\widehat{\boldsymbol{\varphi}}_n - \boldsymbol{\varphi}_0) \xrightarrow{d} \mathcal{N}(0, \widetilde{\boldsymbol{\Sigma}})\,,$$

where $\widetilde{\boldsymbol{\Sigma}}$ denotes the corresponding asymptotic covariance matrix. Therefore the asymptotic normality holds.

Web Appendix C: Some Additional Simulation Results

This part contains four simulation tables, S1, S2, S3 and S4, and the first three have been described and discussed in Section 5 of the main paper, which will not be repeated here.

[Table S1 about here.]

[Table S2 about here.]

[Table S3 about here.]

To assess the effect of the number of bootstrap samples on standard error estimation, we repeated the study in Table S3 under $\varepsilon = 0.5$ and a 50% censoring rate but used 100 bootstrap samples instead of 30 bootstrap samples. The results are presented in Table S4 and are similar to those given in Table S3, suggesting that the use of 30 bootstrap samples seems to be sufficient under the current setting.

[Table S4 about here.]

References

Chen, X., Fan, Y., and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101(475):1228–1240.

Du, M., Lou, Y., and Sun, J. (2025). Estimation and variable selection for interval-censored failure time data with random change point and application to breast cancer study. *Journal of the American Statistical Association*, 0(0):1–12.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35:70–91.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):109–126.

Hall, P. and Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 146, pages 225–256. Cambridge University Press.

Kong, D., Ibrahim, J. G., Lee, E., and Zhu, H. (2018). Flcrm: Functional linear cox regression model. *Biometrics*, 74(1):109–117.

Liu, Z., Wang, H., Wang, C., and Song, X. (2026). Simultaneous variable selection and estimation of survival model with informative censoring. *Statistica Sinica*, 36(2).

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics*, pages 475–478.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics*, 25(6):2555–2591.

van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Zhao, H., Wu, Q., Li, G., and Sun, J. (2020). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association*, 115(529):204–216.

Zhou, Q., Hu, T., and Sun, J. (2017). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, 112(518):664–672.

Table S1: Simulation results on parameter estimation with $\varepsilon = 0.1$.

				$n = 200$				$n = 400$			
Model	p	CR	Parameters	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
PH	10	35%	$\beta_1 = 0.7$	0.0179	0.1100	0.1130	0.960	0.0102	0.0650	0.0764	0.970
			$\beta_2 = 0.7$	0.0119	0.1191	0.1191	0.970	0.0087	0.0738	0.0802	0.965
			$\beta_3 = 0.7$	0.0227	0.1205	0.1203	0.945	0.0153	0.0676	0.0812	0.970
			$\alpha_1 = 0.5$	0.0148	0.0534	0.0490	0.925	0.0054	0.0343	0.0329	0.950
			$\alpha_2 = 0.8$	0.0115	0.1172	0.1101	0.950	0.0134	0.0712	0.0751	0.965
		50%	$\beta_1 = 0.7$	0.0099	0.1177	0.1239	0.955	0.0118	0.0693	0.0836	0.975
			$\beta_2 = 0.7$	0.0195	0.1169	0.1312	0.985	0.0092	0.0822	0.0879	0.955
			$\beta_3 = 0.7$	0.0244	0.1319	0.1322	0.940	0.0159	0.0782	0.0890	0.965
			$\alpha_1 = 0.5$	0.0159	0.0553	0.0529	0.955	0.0058	0.0376	0.0355	0.945
			$\alpha_2 = 0.8$	0.0161	0.1250	0.1205	0.950	0.0148	0.0839	0.0821	0.970
	30	35%	$\beta_1 = 0.7$	0.0249	0.1104	0.1261	0.985	0.0140	0.0812	0.0805	0.940
			$\beta_2 = 0.7$	0.0225	0.1221	0.1335	0.960	0.0061	0.0841	0.0847	0.950
			$\beta_3 = 0.7$	0.0366	0.1128	0.1333	0.975	0.0143	0.0711	0.0854	0.975
			$\alpha_1 = 0.5$	0.0268	0.0593	0.0542	0.910	0.0119	0.0378	0.0344	0.925
			$\alpha_2 = 0.8$	0.0313	0.1261	0.1236	0.940	0.0175	0.0749	0.0793	0.955
		50%	$\beta_1 = 0.7$	0.0302	0.1183	0.1394	0.985	0.0144	0.0887	0.0886	0.935
			$\beta_2 = 0.7$	0.0238	0.1350	0.1471	0.975	0.0066	0.0907	0.0932	0.970
			$\beta_3 = 0.7$	0.0419	0.1266	0.1469	0.970	0.0169	0.0793	0.0941	0.975
			$\alpha_1 = 0.5$	0.0312	0.0639	0.0589	0.910	0.0128	0.0406	0.0374	0.910
			$\alpha_2 = 0.8$	0.0376	0.1372	0.1359	0.945	0.0224	0.0830	0.0869	0.950
PO	10	35%	$\beta_1 = 0.7$	-0.0059	0.2015	0.1663	0.920	0.0125	0.1082	0.1149	0.965
			$\beta_2 = 0.7$	0.0045	0.2017	0.1760	0.955	0.0043	0.1147	0.1215	0.950
			$\beta_3 = 0.7$	0.0154	0.1804	0.1766	0.960	0.0137	0.1032	0.1219	0.970
			$\alpha_1 = 0.5$	0.0124	0.0665	0.0635	0.945	0.0031	0.0475	0.0434	0.940
			$\alpha_2 = 0.8$	-0.0043	0.1654	0.1583	0.945	0.0075	0.1059	0.1097	0.955
		50%	$\beta_1 = 0.7$	-0.0174	0.2355	0.1773	0.915	0.0137	0.1163	0.1226	0.970
			$\beta_2 = 0.7$	0.0046	0.2371	0.1885	0.915	-0.0001	0.1229	0.1296	0.945
			$\beta_3 = 0.7$	0.0208	0.2017	0.1885	0.960	0.0216	0.1161	0.1303	0.970
			$\alpha_1 = 0.5$	0.0129	0.0693	0.0676	0.945	0.0031	0.0512	0.0465	0.925
			$\alpha_2 = 0.8$	0.0010	0.1810	0.1690	0.930	0.0106	0.1181	0.1172	0.960
	30	35%	$\beta_1 = 0.7$	0.0298	0.1634	0.1797	0.960	0.0172	0.1109	0.1188	0.970
			$\beta_2 = 0.7$	0.0186	0.1873	0.1898	0.960	-0.0030	0.1298	0.1255	0.950
			$\beta_3 = 0.7$	0.0281	0.1609	0.1898	0.985	0.0104	0.1104	0.1260	0.985
			$\alpha_1 = 0.5$	0.0234	0.0733	0.0673	0.940	0.0150	0.0459	0.0448	0.935
			$\alpha_2 = 0.8$	0.0296	0.1683	0.1704	0.965	0.0255	0.1058	0.1136	0.955
		50%	$\beta_1 = 0.7$	0.0311	0.1680	0.1936	0.970	0.0122	0.1220	0.1269	0.970
			$\beta_2 = 0.7$	0.0238	0.1974	0.2043	0.965	-0.0002	0.1442	0.1342	0.930
			$\beta_3 = 0.7$	0.0363	0.1813	0.2042	0.980	0.0160	0.1197	0.1350	0.985
			$\alpha_1 = 0.5$	0.0292	0.0781	0.0730	0.930	0.0157	0.0508	0.0481	0.925
			$\alpha_2 = 0.8$	0.0356	0.1757	0.1837	0.960	0.0268	0.1203	0.1218	0.960

Table S2: Simulation results on variable selection with $\varepsilon = 0.5$.

				$n = 200$				$n = 400$			
Model	p	CR	Method	TP	FP	MMSE	STD	TP	FP	MMSE	STD
PH	10	35%	Oracle	3	0	0.0260	0.0371	3	0	0.0112	0.0125
			Proposed	3	0.045	0.0279	0.0416	3	0.030	0.0113	0.0158
			MLE	—	—	0.1003	0.0717	—	—	0.0412	0.0265
		50%	Oracle	3	0	0.0297	0.0393	3	0	0.0130	0.0159
			Proposed	3	0.035	0.0328	0.0414	3	0.035	0.0133	0.0185
			MLE	—	—	0.1231	0.0837	—	—	0.0475	0.0325
	30	35%	Oracle	3	0	0.0263	0.0524	3	0	0.0131	0.0187
			Proposed	3	0.050	0.0269	0.0605	3	0.015	0.0132	0.0203
			MLE	—	—	0.4669	0.2574	—	—	0.1560	0.0681
		50%	Oracle	3	0	0.0315	0.0705	3	0	0.0171	0.0227
			Proposed	3	0.065	0.0338	0.0767	3	0.015	0.0174	0.0248
			MLE	—	—	0.5849	0.3507	—	—	0.1985	0.0845
PO	10	35%	Oracle	3	0	0.0614	0.0700	3	0	0.0239	0.0283
			Proposed	2.940	0.030	0.0614	0.1165	3	0.040	0.0243	0.0337
			MLE	—	—	0.2227	0.1230	—	—	0.0964	0.0543
		50%	Oracle	3	0	0.0634	0.0677	3	0	0.0287	0.0368
			Proposed	2.890	0.035	0.0673	0.1339	3	0.040	0.0315	0.0411
			MLE	—	—	0.2507	0.1402	—	—	0.1087	0.0648
	30	35%	Oracle	3	0	0.0588	0.0760	3	0	0.0284	0.0321
			Proposed	2.990	0.005	0.0585	0.0832	3	0.005	0.0284	0.0327
			MLE	—	—	0.8473	0.3929	—	—	0.3412	0.1259
		50%	Oracle	3	0	0.0633	0.0913	3	0	0.0362	0.0359
			Proposed	2.980	0.055	0.0685	0.1136	3	0.015	0.0370	0.0378
			MLE	—	—	1.0199	0.5317	—	—	0.3994	0.1425

Table S3: Simulation results on parameter estimation with missing covariates.

				Complete-Case Analysis				IPW Analysis			
Model	ε	CR	Parameters	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
PH	0.1	35%	$\beta_1 = 0.7$	0.0013	0.0860	0.0826	0.925	0.0313	0.1122	0.1029	0.925
			$\beta_2 = 0.7$	-0.1329	0.0913	0.0895	0.635	-0.0064	0.1278	0.1145	0.910
			$\beta_3 = 0.7$	-0.0863	0.0841	0.0870	0.845	0.0027	0.1128	0.1078	0.935
			$\alpha_1 = 0.5$	-0.0439	0.0408	0.0356	0.670	0.0127	0.0514	0.0439	0.910
			$\alpha_2 = 0.8$	-0.0633	0.0802	0.0811	0.875	0.0283	0.1082	0.1026	0.935
		50%	$\beta_1 = 0.7$	0.0276	0.1033	0.0974	0.945	0.0405	0.1302	0.1206	0.905
			$\beta_2 = 0.7$	-0.0909	0.0917	0.1032	0.885	0.0037	0.1157	0.1283	0.965
			$\beta_3 = 0.7$	-0.0458	0.0956	0.0991	0.925	0.0151	0.1187	0.1182	0.950
			$\alpha_1 = 0.5$	-0.0172	0.0443	0.0405	0.890	0.0210	0.0533	0.0483	0.910
			$\alpha_2 = 0.8$	-0.0301	0.0938	0.0926	0.920	0.0355	0.1053	0.1086	0.950
	0.5	35%	$\beta_1 = 0.7$	0.0012	0.0859	0.0825	0.925	0.0324	0.1132	0.1029	0.935
			$\beta_2 = 0.7$	-0.1330	0.0911	0.0895	0.630	-0.0068	0.1281	0.1162	0.940
			$\beta_3 = 0.7$	-0.0870	0.0840	0.0870	0.845	0.0055	0.1130	0.1108	0.945
			$\alpha_1 = 0.5$	-0.0440	0.0408	0.0356	0.675	0.0126	0.0516	0.0452	0.920
			$\alpha_2 = 0.8$	-0.0641	0.0800	0.0811	0.880	0.0262	0.1082	0.1012	0.905
		50%	$\beta_1 = 0.7$	0.0285	0.1037	0.0974	0.940	0.0401	0.1294	0.1189	0.920
			$\beta_2 = 0.7$	-0.0931	0.1005	0.1032	0.885	0.0039	0.1147	0.1280	0.965
			$\beta_3 = 0.7$	-0.0460	0.0957	0.0990	0.925	0.0163	0.1191	0.1195	0.945
			$\alpha_1 = 0.5$	-0.0173	0.0449	0.0405	0.885	0.0210	0.0534	0.0494	0.940
			$\alpha_2 = 0.8$	-0.0312	0.0933	0.0925	0.915	0.0338	0.1045	0.1085	0.960
PO	0.1	35%	$\beta_1 = 0.7$	0.1397	0.1446	0.1341	0.825	0.0664	0.1785	0.1716	0.930
			$\beta_2 = 0.7$	-0.2080	0.1720	0.1430	0.695	-0.0247	0.1958	0.2031	0.975
			$\beta_3 = 0.7$	-0.0838	0.1590	0.1382	0.895	-0.0043	0.1952	0.1930	0.960
			$\alpha_1 = 0.5$	-0.0136	0.0558	0.0496	0.925	0.0227	0.0682	0.0636	0.920
			$\alpha_2 = 0.8$	-0.0347	0.1396	0.1254	0.915	0.0324	0.1667	0.1567	0.915
		50%	$\beta_1 = 0.7$	0.1195	0.1593	0.1484	0.860	0.0252	0.2005	0.1911	0.955
			$\beta_2 = 0.7$	-0.1384	0.1770	0.1564	0.875	0.0086	0.1965	0.2062	0.970
			$\beta_3 = 0.7$	-0.0402	0.1754	0.1503	0.930	0.0203	0.1929	0.1929	0.945
			$\alpha_1 = 0.5$	0.0055	0.0570	0.0545	0.955	0.0257	0.0647	0.0650	0.915
			$\alpha_2 = 0.8$	0.0004	0.1439	0.1369	0.935	0.0366	0.1580	0.1628	0.945
	0.5	35%	$\beta_1 = 0.7$	0.1393	0.1431	0.1341	0.820	0.0636	0.1784	0.1759	0.920
			$\beta_2 = 0.7$	-0.2070	0.1709	0.1427	0.710	-0.0237	0.1930	0.1986	0.965
			$\beta_3 = 0.7$	-0.0833	0.1592	0.1382	0.895	-0.0043	0.1959	0.1931	0.950
			$\alpha_1 = 0.5$	-0.0136	0.0560	0.0496	0.920	0.0232	0.0686	0.0633	0.920
			$\alpha_2 = 0.8$	-0.0353	0.1402	0.1254	0.910	0.0331	0.1684	0.1593	0.930
		50%	$\beta_1 = 0.7$	0.1202	0.1607	0.1484	0.845	0.0303	0.1950	0.1909	0.950
			$\beta_2 = 0.7$	-0.1400	0.1799	0.1564	0.875	0.0065	0.1965	0.2054	0.960
			$\beta_3 = 0.7$	-0.0425	0.1778	0.1503	0.925	0.0181	0.1933	0.1909	0.925
			$\alpha_1 = 0.5$	0.0053	0.0574	0.0545	0.950	0.0254	0.0646	0.0646	0.950
			$\alpha_2 = 0.8$	-0.0003	0.1432	0.1368	0.940	0.0354	0.1578	0.1623	0.935

Table S4: Simulation results for variance estimation under $\varepsilon = 0.5$ and 50% CR with varying bootstrap sample sizes.

# of Bootstrap samples		β_1	β_2	β_3	α_1	α_2
				PH Model		
30	SSE	0.1294	0.1147	0.1191	0.0534	0.1045
	SEE	0.1189	0.1280	0.1195	0.0494	0.1085
100	SSE	0.1152	0.1338	0.1194	0.0549	0.1072
	SEE	0.1194	0.1291	0.1180	0.0496	0.1095
				PO Model		
30	SSE	0.1950	0.1965	0.1933	0.0646	0.1578
	SEE	0.1909	0.2054	0.1909	0.0646	0.1623
100	SSE	0.1952	0.1849	0.1884	0.0646	0.1569
	SEE	0.1886	0.2058	0.1888	0.0657	0.1653