Biometrics 000, 000–000 DOI: 000
000 0000

Support Information: Simultaneous Selection and Inference for Varying Coefficients With Zero Regions: A Soft-Thresholding Approach

Yuan Yang[1,$], Ziyang Pan[2,$], Jian Kang[2,*], Chad Brummett[3], and Yi Li[2]

[1] Parexel, Waltham, MA, U.S.A.

[2] Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.

[3] Department of Anesthesiology, University of Michigan, Ann Arbor, MI, U.S.A.

[$] Authors contributed equally

[*]*email:* jiankang@umich.edu

This paper has been submitted for consideration for publication in *Biometrics*

We first introduce some common notation that will be used throughout the Appendix. Let a_{1n} and a_{2n} be two sequences of real numbers indexed by positive integers and a_{2n} is positive for all n. For a real number a_1, say a_{1n} tends to a limit a_1 in symbols: $a_{1n} \to a_1$ as $n \to \infty$. We say $a_{1n} = O(a_{2n})$ if there exist an $M > 1$ and a finite $N > 0$ such that $M^{-1} < |a_{1n}/a_{2n}| < M$ when $n > N$. We say $a_{1n} = o(a_{2n})$ if $|a_{1n}/a_{2n}| \to 0$ as $n \to \infty$. For a sequence of random variables Z_n, we say $Z_n = O_p(a_{1n})$ if for any $\delta > 0$, there exist a finite $M > 0$ and a finite $N > 0$ such that $\Pr(|Z_n/a_{1n}| > M) < \delta$ when $n > N$; and $Z_n = o_p(a_{1n})$ if for any $\delta > 0$, $\Pr(|Z_n/a_{1n}| > \delta) \to 0$ as $n \to \infty$. The convergence of Z_n in distribution to a random variable Z is denoted by $Z_n \to_d Z$, which implies that $\lim F_n(z) = F(z)$ as $n \to \infty$ for every z at which F is continuous, where F_n and F are the cumulative distribution functions of random variables Z_n and Z, respectively. Let $\mathrm{E}_n f(\cdot) = n^{-1}\sum_{i=1}^n f(\cdot)$ be the empirical mean of f, and $\mathrm{E}f$ the theoretical mean of f. Let $\otimes$ denote the Kronecker product. Let f' and f'' denote the first and second derivatives of f function, respectively. Let $N(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2. Let $\mathcal{I}(\mathcal{A})$ be an event indicator function, where $\mathcal{I}(\mathcal{A}) = 1$ if event $\mathcal{A}$ is true and $\mathcal{I}(\mathcal{A}) = 0$ otherwise. Let I_d be a $d \times d$ identity matrix. For a real valued function θ on $\mathbb{D}$, $||\theta||_\infty = \sup_{w\in\mathbb{D}} |\theta(w)|$ denotes its supreme norm and $||\theta||_2 = \{\int_{w\in\mathbb{D}} |\theta(w)|^2 dw\}^{1/2}$ denotes its $\mathcal{L}_2$ norm. For a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, let $||\boldsymbol{\theta}||_2 = \big\{\sum_j ||\theta_j||_2^2\big\}^{1/2}$ and $||\boldsymbol{\theta}||_\infty = \max_{1\leqslant j\leqslant p} ||\theta_j||_\infty$.

S1. Regularity conditions

We make the following technical conditions to ensure the theoretical properties as outlined in Theorems 1, 2, and 3.

(C1) The covariates $\boldsymbol{X}$ take values in a bounded subset of $\mathbb{R}^p$. That is, there exist finite real numbers C_1 and C_2 such that $\Pr(C_1 < X_j < C_2, \text{for all } j = 1, \ldots, p) = 1$.

(C2) The eigenvalues $\lambda_1 \leqslant \ldots \leqslant \lambda_p$ of $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T \mid W)$ are bounded away from zero and infinity almost surely, that is, there are positive constants M_1 and M_2 such that $\Pr(M_1 \leqslant \lambda_1 \leqslant \ldots \leqslant$

$\lambda_p \leqslant M_2) = 1$. Consequently, the eigenvalues of $\mathrm{E}(\boldsymbol{V}_n \boldsymbol{V}_n^T)$ are bounded away from zero and infinity almost surely.

(**C3**) $\lim_{\lambda \to \infty} \mathrm{E}\{\epsilon^2 \mathcal{I}(|\epsilon| > \lambda)\} = 0$ and $\mathrm{E}\{\exp(t\epsilon)\} \leqslant \exp(\sigma^2 t^2/2)$ for any $t \in \mathbb{R}$.

(**C4**) $l_n''(\boldsymbol{\gamma})$ is bounded and has a bounded inverse around $\tilde{\boldsymbol{\gamma}}$; that is, $\mathrm{E}\{\boldsymbol{U}(\tilde{\boldsymbol{\gamma}}; \boldsymbol{X}, W)\boldsymbol{X}^T\}$ is invertible.

(**C5**) The distribution of W is absolutely continuous with a density bounded away from zero and infinity on $\mathbb{D}$.

(**C6**) For $\nu \in (0, 1/2)$ and $m > 1/2$, $q = O(n^\nu)$, $\tilde{p} = o(\min\{n/q, q^{2m}\})$, $\rho = o(\tilde{p}^{1/2} q^{-m}) = o(\min\{n^{1/2-\nu/2-m\nu}, 1\})$ and $\nabla(\eta) = o(q^{-m}) = o(n^{-m\nu})$.

(**C7**) The true varying coefficients β_{0j} $(j = 1, \ldots, p)$ are bounded.

Conditions (**C1**),(**C2**) and (**C4**) are mild regularity conditions used in the existing literature (Fan and Zhang, 1999; Huang et al., 2002). Condition (**C3**) essentially assumes the error distribution is sub-Gaussian, which has been assumed for varying-coefficient models (Wei et al., 2011). Condition (**C5**) guarantees that observations are randomly scattered (Huang et al., 2004). Condition (**C6**) is a technical assumption that controls convergence rate, estimation bias, and model sparsity. Related conditions have been discussed by (Huang et al., 2002, 2004). Condition (**C7**) is reasonable for a wide range of applications. Similar assumptions have been made by (Huang and Shen, 2004) for other varying coefficient models.

S2. TECHNICAL DERIVATIONS

S2.1 *Properties of $H_\eta(\theta,\alpha)$*

For any $\eta > 0$, $\alpha > 0$, and a real function θ, we have

$$\begin{aligned}
&\left|\zeta_{(\theta,\alpha)} - H_\eta(\theta,\alpha)\right| \\
=&\Bigg|(\theta-\alpha)\mathcal{I}(\theta>\alpha) + (\theta+\alpha)\mathcal{I}(\theta<-\alpha) - \frac{1}{2}\left\{1+\frac{2}{\pi}\text{arctan}\left(\frac{\theta-\alpha}{\eta}\right)\right\}(\theta-\alpha) - \\
&\frac{1}{2}\left\{1-\frac{2}{\pi}\text{arctan}\left(\frac{\theta+\alpha}{\eta}\right)\right\}(\theta+\alpha)\Bigg| \\
=&\Bigg|(\theta-\alpha)\left[\mathcal{I}(\theta>\alpha) - \frac{1}{2}\left\{1+\frac{2}{\pi}\text{arctan}\left(\frac{\theta-\alpha}{\eta}\right)\right\}\right] + \\
&(\theta+\alpha)\left[\mathcal{I}(\theta<-\alpha) - \frac{1}{2}\left\{1+\frac{2}{\pi}\text{arctan}\left(\frac{\theta+\alpha}{\eta}\right)\right\}\right]\Bigg| \\
\leqslant&\left|(\theta-\alpha)\left[\mathcal{I}(\theta>\alpha) - \frac{1}{2}\left\{1+\text{sign}(\theta-\alpha)+\frac{\eta}{\theta-\alpha}+O(\eta^3)\right\}\right]\right| + \\
&\left|(\theta+\alpha)\left[\mathcal{I}(\theta<-\alpha) - \frac{1}{2}\left\{1+\text{sign}(\theta+\alpha)+\frac{\eta}{\theta+\alpha}+O(\eta^3)\right\}\right]\right| \\
=&\eta + O(\eta^3).
\end{aligned}$$

Therefore, the bias due to approximation is bounded by $\eta + O(\eta^3)$.

When α and η are fixed, the first derivative of h function in terms of θ is

$$\begin{aligned}
H'_\eta(\theta,\alpha) = &\frac{1}{\pi}\cdot\frac{(\theta-\alpha)/\eta}{1+(\theta-\alpha)^2/\eta^2} + \frac{1}{2}\left\{1+\frac{2}{\pi}\text{ arctan}\left(\frac{\theta-\alpha}{\eta}\right)\right\} - \frac{1}{\pi}\cdot\frac{(\theta-\alpha)/\eta}{1+(\theta-\alpha)^2/\eta^2} \\
&+\frac{1}{2}\left\{1-\frac{2}{\pi}\text{ arctan}\left(\frac{\theta+\alpha}{\eta}\right)\right\},
\end{aligned}$$

and the second derivative is

$$H''_\eta(\theta,\alpha) = \frac{2}{\pi}\cdot\frac{(\eta-\theta+\alpha)/\eta^2}{1+(\theta-\alpha)^2/\eta^2} - \frac{2}{\pi}\cdot\frac{(\eta-\theta-\alpha)/\eta^2}{1+(\theta+\alpha)^2/\eta^2}.$$

To facilitate the ensuing proofs, we also provide the approximation of H' here. For $-\alpha < \theta < \alpha$, by the Taylor expansion of H' around $\eta = 0$, we have

$$H'_\eta(\theta,\alpha) = \frac{1}{\pi}\left\{\frac{2(\theta-\alpha)^2-8}{(\theta-\alpha)^5} - \frac{2(\theta+\alpha)^2-8}{(\theta+\alpha)^5}\right\}\eta^3 + o(\eta^3).$$

S3. TECHNICAL PROOFS

Let $M_n(\boldsymbol{\theta}) = -\mathrm{E}_n l^s(\boldsymbol{\theta})$ and $M_0(\boldsymbol{\theta}) = -\mathrm{E} l^s(\boldsymbol{\theta})$ be the empirical and theoretical mean of $l^s(\boldsymbol{\theta})$. Let $|v|$ denote the Euclidean norm of a real valued vector v. For a real valued function θ on $\mathbb{D}$, $||\theta||_\infty = \sup_{w\in\mathbb{D}} |\theta(w)|$ denotes its supreme norm and $||\theta||_2 = \{\int_{w\in\mathbb{D}} |\theta(w)|^2\}^{1/2}$ denotes its $\mathcal{L}_2$ norm. For a vector valued function $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, let $||\boldsymbol{\theta}||_2 = \left\{ \sum_j ||\theta_j||_2^2 \right\}^{1/2}$ and $||\boldsymbol{\theta}||_\infty = \max_{1\leqslant j\leqslant p} ||\theta_j||_\infty$. Let $N_{[]}(\delta, \mathbb{S}, \mathcal{L}_p)$ be the δ-bracketing number for $\mathbb{S}$ under norm $\mathcal{L}_p$ and $\mathrm{E}^*(g)$ denote the outer expectation of process g. For two sequences a_n and b_n, we say $a_n \simeq b_n$ if $a_n/b_n = O(1)$. The convergence of Z_n in distribution to a random variable Z is denoted by $Z_n \rightarrow_d Z$, which implies that $\lim F_n(z) = F(z)$ as $n \rightarrow \infty$ for every z at which F is continuous, where F_n and F are the cumulative distribution functions of random variables Z_n and Z, respectively. The convergence of Z_n in probability to a random variable Z is denoted by $Z_n \rightarrow_p Z$, which implies that $\lim \Pr(|Z_n - Z| > \epsilon) = 0$ as $n \rightarrow \infty$ for all $\epsilon > 0$. A sequence of random vectors or matrices converge to a random vector or matrix if and only if each component of random vectors or matrices converges in probability to each component of the vector or matrix.

Lemma 1: *For any function $\beta(w) \in \mathbb{H}$ and any $\alpha > 0$, there exists at least one $\theta(w) \in \mathbb{F}_0$ such that $\beta(w) = \zeta_{\{\theta,\alpha\}}(w)$.*

Proof of Lemma 1: When the zero region is empty, then $\theta(w) = \alpha + \beta(w)$ if $\beta(w) > 0$ and $\theta(w) = \beta(w) - \alpha$ if $\beta(w) < 0$. We show that Lemma 1 is valid when $\beta(w)$ has only one zero region (w_0, w_1), where $w_0, w_1 \in (0, 1)$. The proof can be easily extended to more general settings. Without loss of generality, we further assume $\beta(w) < 0$ on $[0, w_0)$ and $\beta(w) > 0$ on $(w_1, 1]$. The definition of $\beta(w)$ implies that $\beta^{(j)}$ exists on $[0, w_0]$ and $[w_1, 1]$, and that there exists a constant $M > 0$ such that $|\beta^{(j)}(w_k)| < M$ for $j = 1, \ldots, d$ and $k = 0, 1$.

In the following, we construct a θ satisfying: (i) $\theta(w) = b(w) - \alpha$ on $[0, w_0]$ and $\theta(w) =$

$b(w)+\alpha$ on $[w_1,1]$; (ii) for $j=1,\ldots,d$, $\theta^{(j)}(w_0)=\beta^{(j)}(w_0)$, and $\theta^{(j)}(w_1)=\beta^{(j)}(w_1)$; (iii) $|\theta(w)|<\alpha$ on (w_0,w_1); and (iv) $|\theta^{(d)}(s)-\theta^{(d)}(w)|\leqslant C|s-w|^t$ for s, w in $[0,1]$ and some constant C, where $0<t\leqslant 1$.

Let $f(w)=e^{-1/w}\mathcal{I}(w>0)$. It follows that $f(w)\in[0,1)$ and $f^{(d)}(0)=0$ for any $d\geqslant 1$. Define $f_0(w,a_0)=f(-w+a_0)/\{f(-w+a_0)+f(-w_0+w)\}$ and $f_1(w,a_1)=f(w-a_1)/\{f(w-a_1)+f(w_1-w)\}$, where $a_0\in(w_0,(w_0+w_1)/2)$ and $a_1\in((w_0+w_1)/2,w_1)$. As $f(w)$ is infinitely differentiable over the real line, so is $f_k(w)$ for $k=0,1$. It is easy to verify that $f_k(w,a_1)$ satisfies that $f_k(w_k,a_k)=1$, $f_k(a_k,a_k)=0$, $f_k^{(j)}(w,a_k)=0$ when $w=a_k$ or w_k, and $0\leqslant f_k(w,a_k)\leqslant 1$ for $k=0,1$ and $j\geqslant 1$.

Let $\theta_0^*(w)=-\alpha+\sum_{j=1}^d\frac{\beta^{(j)}(w_0)}{j!}(w-w_0)^j$ and $\theta_1^*(w)=\alpha+\sum_{j=1}^d\frac{\beta^{(j)}(w_1)}{j!}(w-w_1)^j$. We define

$$\theta(w)=\begin{cases} b(w)-\alpha, & w\in[0,w_0] \\ \theta_0^*(w)*f_0(w,a_0), & w\in(w_0,a_0] \\ 0, & w\in(a_0,a_1) \\ \theta_1^*(w)*f_1(w,a_1), & w\in[a_1,w_1) \\ b(w)+\alpha, & w\in[w_1,1] \end{cases},$$

and show that there exist a_0 and a_1 which ensure the above $\theta(w)$ satisfies conditions (i)-(iv).

It is obvious that $\theta(w)$ satisfies (i) and $\theta(w)$ is continuous. Since $f_k(w_k,a_k)=1$ and $f_k^{(j)}(w_k,a_k)=0$ for $j\geqslant 1$, we have that $\theta^{(j)}(w_k)=\theta_k^{*(j)}(w_k)=\beta^{(j)}(w_k)$ for $j=1,\ldots,d$, where $k=0,1$. Therefore, condition (ii) is satisfied.

Since $\theta_0^*(w)$ and $f_0(w,a_0)$ are infinitely differentiable over (w_0,a_0), so is $\theta(w)$ over (w_0,a_0). Similarly, $\theta(w)$ is also infinitely differentiable over (a_1,w_1). Because $f_k^{(j)}(a_k,a_k)=0$ for $j\geqslant 0$ and $k=0,1$, we have that $\theta^{(j)}(a_k)=0$ for $j\geqslant 0$ and $k=0,1$. Therefore, $\theta(w)$ is infinitely differentiable over (w_0,w_1), which implies $\theta(w)$ also satisfies condition (iv) over (w_0,w_1).

Apparently, condition (iv) is satisfied when w and s are in the same region (zero or non-zero region) by taking $t = 1$. We only verify that condition (iv) is valid when $w \in [0, w_0)$ and $s \in [w_0, w_1]$. The other situations can be verified similarly.

To proceed, we notice

$$\begin{aligned}|\theta^{(d)}(w) - \theta^{(d)}(s)| &= |\theta^{(d)}(w) - \theta^{(d)}(w_0) + \theta^{(d)}(w_0) - \theta^{(d)}(s)| \\ &\leqslant |\theta^{(d)}(w) - \theta^{(d)}(w_0)| + |\theta^{(d)}(w_0) - \theta^{(d)}(s)| \\ &\leqslant C_1|w - w_0| + C_2|w_0 - s| \\ &\leqslant \max\{C_1, C_2\}|w - s|.\end{aligned}$$

Hence, condition (iv) is valid for $t = 1$.

To prove condition (iii), we just need to find a_0 and a_1 such that $\theta^{'}(w) \geqslant 0$ over $[w_0, w_1]$. By the construction of $\theta(w)$, we have $\theta^{'}(w) = 0$ over $[a_0, a_1]$. When $w \in (a_1, w_1)$, we let $r_1 = w_1 - a_1$ and show

$$\begin{aligned}|\theta_1^{*'}(w)| &= \left|\sum_{j=1}^{d} \frac{b^{(j)}(w_1)}{(j-1)!}(w - w_1)^{j-1}\right| \\ &\leqslant \sum_{j=1}^{d} \left|\frac{b^{(j)}(w_1)}{(j-1)!}(w - w_1)^{j-1}\right| \leqslant M \sum_{j=1}^{d} r_1^{j-1} \leqslant \frac{M}{1 - r_1}, \\ \theta_1^{*}(w) &\geqslant \alpha - \left|\sum_{j=1}^{d} \frac{\beta^{(j)}(w_1)}{j!}(w - w_1)^{j}\right| \\ &\geqslant \alpha - \sum_{j=1}^{d} \left|\frac{\beta^{(j)}(w_1)}{j!}(w - w_1)^{j}\right| \geqslant \alpha - \frac{Mr_1}{1 - r_1}, \\ \text{and } f_1^{'}(w, a_1) &= \frac{e^{-1/(w-a_1)-1/(w_1-w)}\left\{1/(w - a_1)^2 + 1/(w_1 - w)^2\right\}}{\{e^{-1/(w-a_1)} + e^{-1/(w_1-w)}\}^2} \\ &\geqslant \frac{1/(w - a_1)^2 + 1/(w_1 - w)^2}{2^2} \\ &\geqslant \frac{1}{2r_1^2}.\end{aligned}$$

Then

$$\begin{aligned}\theta^{'}(w) &= \theta_1^{*'}(w) f_1(w, a_1) + \theta_1^{*}(w) f_1^{'}(w, a_1) \\ &\geqslant \theta_1^{*}(w) f_1^{'}(w, a_1) - |\theta_1^{*'}(w) f_1(w, a_1)| \\ &\geqslant \left(\alpha - \frac{Mr_1}{1 - r_1}\right)\frac{1}{2r_1^2} - \frac{M}{1 - r_1}.\end{aligned}$$

Let

$$g(r) = \left(\alpha - \frac{Mr}{1-r}\right)\frac{1}{2r^2} - \frac{M}{1-r},$$

then when $0 < r < 1$,

$$g^{'}(r) = -\frac{\alpha}{r^3} - \frac{M}{2r^2(1-r)^2} - \frac{M}{(1-r)^2} < 0.$$

Therefore, $g(r)$ is strictly decreasing on $(0,1)$. As $\lim_{r\downarrow 0} g(r) = \infty$ and $\lim_{r\uparrow 1} g(r) = -\infty$, there exists a unique $r^* \in (0,1)$ such that $g(r^*) = 0$. Therefore, $g(r) > 0$ over $(0, r^*)$. Let $r_1 = \min\{r^*, (w_1 - w_0)/2\}$, and we have $\theta^{'}(w) > 0$ over $(w_1 - r_1, w_1)$. Thus, we find an $a_1 = w_1 - r_1$ such that $|\theta(w)| \leqslant \alpha$ over (a_1, w_1). Similarly, we can find an a_0 such that $|\theta(w)| \leqslant \alpha$ over (w_0, a_0). Therefore, condition (iii) is satisfied.

Combining all the results, we have found a $\theta \in \mathbb{F}_0$ such that $\zeta_{(\theta,\alpha)}(w) = \beta(w)$, which completes the proof. ■

Lemma 2: *For any smooth zero-crossing function $\beta^*(w) \in \mathbb{F}_0$, there exists $\beta(w) \in \mathbb{H}$ such that $\beta^*(w) = \beta(w)$ on any set of finite grid points $\{w_1, w_2, \ldots, w_F\}$.*

Proof of Lemma 2: We assume $\beta^*(w)$ has only one zero point w_0. Extension to multiple zero points case is straightforward. Then w_0 must fall into (w_i, W_{i+1}) for some i. Since $\beta^*(w)$ is smooth, there exists $a, b \in \mathbb{R}$ and $w_i < a < b < w_{i+1}$ such that $\beta^*(w)$ is increasing or decreasing in $[a,b]$ and $\beta^*(a) * \beta^*(b) < 0$. For simplicity, we consider the increasing case. Then we have $\beta^*(w)$ increasing on $[a,b]$, and $\beta^*(a) < 0$ and $\beta^*(b) > 0$.

Let $\theta_0^*(w) = \beta^*(a) + \sum_{j=1}^{d} \frac{\beta^{(j)}(a)}{j!}(w-a)^j$ and $\theta_1^*(w) = \beta^*(b) + \sum_{j=1}^{d} \frac{\beta^{(j)}(b)}{j!}(w-b)^j$. With

$f_0(w,x)$ and $f_1(w,x)$ defined in the proof of Lemma 1 and $a_0 < a_1$, we construct

$$\beta(w) = \begin{cases} \beta^*(w), & w < a \\ \theta_0^*(w) * f_0(w, a_0), & w \in (a, a_0] \\ 0, & w \in (a_0, a_1) \cdot \\ \theta_1^*(w) * f_1(w, a_1), & w \in [a_1, b) \\ \beta^*(w), & w > b \end{cases}$$

According to the proof of Lemma 1, the constructed $\beta(w)$ belongs to $\mathbb{H}$. By the construction of $\beta(w)$, we have $\beta(w) = \beta^*(w)$ for any $w < a$ and $w > b$. Therefore, $\beta^*(w_j) = \beta(w_j)$ for $j = 1, ..., F$. ■

LEMMA 3: *Under Conditions **(C1)**, **(C5)**, and **(C7)**, if $\beta_j \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$ with q and α_j the same as in the penalized likelihood, then $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho)^{1/2})$; if $\beta_j \notin \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, we have $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho + \tilde{p}q^{-2m})^{1/2})$, where m is the smoothness parameter as in Definition 1.*

Proof of Lemma 3:

Let $l_0(\beta; \boldsymbol{X}, Y, W) = \left[Y - \sum_{j=1}^p X_j b_j(W)\right]^2$. By model assumption, we have $\mathrm{E}_{Y|\boldsymbol{X},W} Y = \sum_{j=1}^p X_j b_{0j}(W)$, then the true parameter $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^T = \arg\min_{\beta \in \mathbb{H}^p} \mathrm{E} l_0(\beta; \boldsymbol{X}, Y, W)$.

By definition, we have $l(\theta; \boldsymbol{X}, Y, W) = \left[Y - \sum_{j=1}^p X_j \zeta_{\{\theta_j, \alpha_j\}}(W)\right]^2 + \rho \sum_{j=1}^p \{\theta_j(W)\}^2$ and $\tilde{\theta} = (\boldsymbol{B}^T \tilde{\gamma}_1, \ldots, \boldsymbol{B}^T \tilde{\boldsymbol{\gamma}}_p)^T = \arg\min_{\theta \in \mathbb{F}^p} \mathrm{E} l(\theta; \boldsymbol{X}, Y, W)$. Since $\beta_{0j} = 0$ for $j > \tilde{p}$, we can infer that $\tilde{\theta}_j = 0$ for $j > \tilde{p}$, and thus $\tilde{\beta}_j = 0$ for $j > \tilde{p}$.

Then by calculation,

$$
\begin{aligned}
&\mathrm{E}l_0(\beta_0;\boldsymbol{X},Y,W)-\mathrm{E}l(\tilde{\theta};\boldsymbol{X},Y,W)\\
=&\mathrm{E}\left[Y-\sum_{j=1}^{p}X_jb_{0j}(W)\right]^2-\mathrm{E}\left[Y-\sum_{j=1}^{p}X_j\tilde{\beta}_j(W)\right]^2-\rho\mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2\\
=&\mathrm{E}\left[\sum_{j=1}^{p}X_j\left\{\tilde{\beta}_j(W)-b_{0j}(W)\right\}\right]\left[2Y-\sum_{j=1}^{p}X_jb_{0j}(W)-\sum_{j=1}^{p}X_j\tilde{\beta}_j(W)\right]-\rho\mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2\\
=&-\mathrm{E}\left[\sum_{j=1}^{p}X_j\left\{\tilde{\beta}_j(W)-b_{0j}(W)\right\}\right]^2-\rho\mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2.
\end{aligned}
\tag{S3.1}
$$

According to Lemma 1, for $j=1,\ldots,\tilde{p}$, there exists $\theta_j\in\mathbb{F}_0$ such that $\zeta_{\{\theta_j,\alpha_j\}}=\beta_{0j}$. If $\theta_j\notin\mathbb{F}$, then we can find $\theta_j^*\in\mathbb{F}$ such that $||\theta_j-\theta_j^*||_2=O(q^{-m})$. When $j>\tilde{p}$, let $\theta_j^*=0$, then we have $\zeta_{\{\theta_j^*,\alpha_j\}}=0=\beta_{0j}$. Let $\beta^*(w)=(\beta_1^*,\ldots,\beta_p^*)^T$, where $\beta_j^*=\zeta_{\left\{\theta_j^*,\alpha_j\right\}}(w)$. Then by Condition (**C1**) and (**C5**), we have

$$
\begin{aligned}
&\mathrm{E}l_0(\beta^*;\boldsymbol{X},Y,W)-\mathrm{E}l_0(\beta_0;\boldsymbol{X},Y,W)=\mathrm{E}\left[\sum_{j=1}^{p}X_j\left\{\beta_j^*(W)-b_{0j}(W)\right\}\right]^2\\
=&\mathrm{E}\left[\sum_{j,k}X_jX_k\left\{\beta_j^*(W)-b_{0j}(W)\right\}\left\{\beta_k^*(W)-b_k(W)\right\}\right]\\
=&\mathrm{E}\left[\left\{\beta_1^*(W)-b_{01}(W),\ldots,\beta_p^*(W)-b_{0p}(W)\right\}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T\mid W)\left\{\beta_1^*(W)-b_{01}(W),\ldots,\beta_p^*(W)-b_{0p}(W)\right\}^T\right.\\
\leqslant&\lambda_p\mathrm{E}\sum_{j=1}^{p}(\beta_j^*(W)-b_{0j}(W))^2=\lambda_p\sum_{j=1}^{\tilde{p}}\mathrm{E}||\beta_j^*(W)-b_{0j}(W)||_2^2\\
=&O(\tilde{p}q^{-2m}).
\end{aligned}
\tag{S3.2}
$$

If for $j=1,\ldots,\tilde{p}$, $\theta_j\in\mathbb{F}$, let $\theta_j^*=\theta_j$, then we have $\beta^*=\beta$ and $\mathrm{E}l_0(\beta^*;\boldsymbol{X},Y,W)-\mathrm{E}l_0(\beta_0;\boldsymbol{X},Y,W)=0$. Here, we assume all β_{0j} $(j=1,\ldots,\tilde{p})$ have the same smoothness, either $\beta_j\in\mathbb{S}_{q,\alpha_j}$ for $j=1,\ldots,\tilde{p}$, or $\beta_j\notin\mathbb{S}_{q,\alpha_j}$ for $j=1,\ldots,\tilde{p}$.

By definition of $\tilde{\theta}$, we have $\mathrm{E}l(\tilde{\theta})\leqslant\mathrm{E}l(\theta^*)=\mathrm{E}l_0(\beta^*)+\rho\mathrm{E}\sum_{j=1}^{p}\left\{\theta_j^*(W)\right\}^2$. Therefore, $\mathrm{E}l(\tilde{\theta})-\mathrm{E}l_0(\beta^*)\leqslant\rho\mathrm{E}\sum_{j=1}^{p}\left\{\theta_j^*(W)\right\}^2$. If $\theta_j\notin\mathbb{F}$ for all $j\leqslant\tilde{p}$, based on equation (S3.1), (S3.2) and Condition (**C7**), we have

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2 &= \mathrm{E}l(\tilde{\theta}) - \mathrm{E}l_0(\beta_0) - \rho \mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2 \\
&\leqslant \mathrm{E}l(\tilde{\theta}) - \mathrm{E}l_0(\beta^*) + \mathrm{E}l_0(\beta^*) - \mathrm{E}l_0(\beta) - \rho \mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2 \\
&\leqslant \rho \mathrm{E}\sum_{j=1}^{p}\left\{\theta_j^*(W)\right\}^2 - \rho \mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2 + \mathrm{E}l_0(\beta^*) - \mathrm{E}l_0(\beta) \\
&= O(\tilde{p}\rho + \tilde{p}q^{-2m}).
\end{aligned}$$

If $\theta_j \in \mathbb{F}$ for all j, then $\mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2 = O(\tilde{p}\rho)$.

By Conditions (**C1**) and (**C5**), we also have

$$\mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2 = \mathrm{E}\left[\sum_{j,k} X_j X_k \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\left\{\tilde{\beta}_k(W) - b_{0k}(W)\right\}\right]$$

$$=\mathrm{E}\Bigg[\left\{\tilde{\beta}_1(W) - b_{01}(W), \ldots, \tilde{\beta}_p(W) - b_{0p}(W)\right\}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T \mid W)\left\{\tilde{\beta}_1(W) - b_{01}(W), \ldots, \tilde{\beta}_p(W) - b_{0p}(W)\right\}$$

$$\geqslant \lambda_1 \mathrm{E}\sum_{j=1}^{p}(\tilde{\beta}_j(W) - b_{0j}(W))^2 = \lambda_1 \sum_{j=1}^{p} ||\tilde{\beta}_j(W) - b_{0j}(W)||_2^2.$$

Therefore, $\max_{1\leqslant j\leqslant p} ||\tilde{\beta}_j - b_{0j}||_2^2 = O(\mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2)$. In addition, $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = \max_{1\leqslant j\leqslant p} ||\tilde{\beta}_j - b_{0j}||_\infty \leqslant \max_{1\leqslant j\leqslant p} ||\tilde{\beta}_j - b_{0j}||_2$. Combining all above results, we conclude: if $\beta_j \notin \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, we have $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho + \tilde{p}q^{-2m})^{1/2})$; if $\beta_j \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho)^{1/2})$. ■

We introduce two important lemmas in order to prove our main theorems. Lemma 4 is a variation of the Lyapunov central limit theorem and will be used in the proof of Lemma 6, and Lemma 5 is used in the proof of Theorem 1.

LEMMA 4: *Suppose ϵ_i are independent with mean 0 and variance 1, and ϵ_i satisfy Condition (**C3**). If $\max_i a_i^2/(\sum_i a_i^2) \to 0$, then*

$$\frac{\sum_i a_i \epsilon_i}{\sqrt{(\sum_i a_i^2)}} \to_d N(0,1).$$

LEMMA 5 (Consistency): *Under Conditions (**C1**), (**C2**), (**C4**), (**C6**) and (**C7**),*

$$||\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}||_2^2 = o_p(\tilde{p}q^{-1}),$$

where $\tilde{\boldsymbol{\theta}} = \boldsymbol{B}\tilde{\boldsymbol{\gamma}}$.

Proof of Lemma 5:

Let $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_p^*)^T$. We choose $\theta_j^* \in \mathbb{F}$ such that $||\boldsymbol{\theta}_j^*||_2^2 = O(q^{-1})$ for $j = 1, \ldots, p$. Let $T_n(a) = M_n(\tilde{\boldsymbol{\theta}} + a\boldsymbol{\theta}^*)$. The derivative of T_n with respect to a is

$$T_n'(a) = -2\mathrm{E}_n\left[\left\{Y - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j + a\theta_j^*)\right\} \sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a\theta_j^*)\theta_j^* - \rho\sum_{j=1}^{p}(\tilde{\theta}_j + a\theta_j^*)\theta_j^*\right]. \quad \text{(S3.3)}$$

When a is sufficiently small, T_n is convex. Thus, T_n' is non-decreasing. Therefore, we only need to show that for any small $a_0 > 0$, $-T_n'(a_0) < 0$ and $-T_n'(-a_0) > 0$. Then, $||\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}||_2 \leqslant a_0||\boldsymbol{\theta}^*||_2$. Since $\tilde{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \mathrm{E}l(\boldsymbol{\gamma}; \boldsymbol{X}, Y, W)$, then $\tilde{\theta}_j = \boldsymbol{B}\tilde{\boldsymbol{\gamma}}_j \equiv 0$ for $l > \tilde{p}$. By Condition (**C6**), $\alpha_j > ||\theta_j^*||_2$ for $l > \tilde{p}$. Thus, $h_j(\tilde{\theta}_j + a\theta_j^*) \equiv 0$ for $l > \tilde{p}$. Then, we have $\sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j + a\theta_j^*) = \sum_{j=1}^{\tilde{p}} X_j h_j(\tilde{\theta}_j + a\theta_j^*)$.

From (S3.3), we have

$$\begin{aligned}
-\frac{1}{2}T_n'(a_0) =& \mathrm{E}_n\left[\left\{Y - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j + a_0\theta_j^*)\right\} \cdot \left\{\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right\} - \rho\sum_{j=1}^{p}(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right] \\
=& \mathrm{E}_n\left\{Y - \sum_{j=1}^{p} X_j\tilde{\beta}_j\right\} \cdot \left\{\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right\} + \\
& \mathrm{E}_n\left\{\sum_{j=1}^{p} X_j\tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j)\right\} \cdot \left\{\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right\} + \\
& \mathrm{E}_n\left\{\sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j) - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j + a_0\theta_j^*)\right\} \cdot \left\{\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right\} - \\
& \rho\mathrm{E}_n\sum_{j=1}^{p}(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^* \\
=& A_1 + A_2 + A_3 + A_4,
\end{aligned}$$

where $\tilde{\beta}_j = \zeta_{(\tilde{\theta}_j, \alpha_j)}$.

By the definition of h_j, we have that $|h_j'(\tilde{\theta}_j + a_0\theta_j^*)| \leqslant 1$ for $j = 1, \ldots, \tilde{p}$ and $|h_j'(\tilde{\theta}_j + a_0\theta_j^*)| \equiv$

0 for $j = \tilde{p}+1, \ldots, p$. Let $h_n = \sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*$. Then $\mathrm{E}(h_n^2) = O(\sum_{j=1}^{\tilde{p}} ||\boldsymbol{\theta}_j^*||_2^2) = O(\tilde{p}q^{-1})$ by Condition (**C2**). Since $Y - \sum_{j=1}^{p} X_j\tilde{\beta}_j = \epsilon$, by Chebyshev's inequality, we have

$$\Pr(|A_1| > 1/\sqrt{n}) \leqslant \frac{\mathrm{E}(\mathrm{E}_n h_n\epsilon)^2}{1/n} \leqslant \frac{\mathrm{E}\{(\mathrm{E}_n h_n)^2(\mathrm{E}_n\epsilon)^2\}}{1/n} = \frac{O(\mathrm{E}h_n^2)\mathrm{E}\epsilon^2/n}{1/n} = O(\tilde{p}q^{-1})\sigma^2.$$

Therefore, $|A_1| = o_p(n^{-1/2}) = o_p(\tilde{p}q^{-1})$.

By the definition of $\tilde{\boldsymbol{\gamma}}$, it satisfies the score equation

$$0 = \mathrm{E}l^{s'} = -2\mathrm{E}\left\{(Y - \boldsymbol{X}^T\tilde{\boldsymbol{h}}) \cdot \tilde{\boldsymbol{U}} \otimes \boldsymbol{B}(W) - \rho\tilde{\boldsymbol{\theta}} \otimes \boldsymbol{B}(W)\right\}, \qquad \text{(S3.4)}$$

where $\tilde{\boldsymbol{h}}$, $\tilde{\boldsymbol{U}}$, $\tilde{\boldsymbol{\theta}}$ are $\boldsymbol{h}$, $\boldsymbol{U}$, $\boldsymbol{\theta}$ with $\boldsymbol{\gamma}$ replaced by $\tilde{\boldsymbol{\gamma}}$ respectively. Since $\boldsymbol{B}(W) \neq 0$ for any $W \in \mathbb{D}$, equation (S3.4) becomes $\mathrm{E}\left\{(Y - \boldsymbol{X}^T\tilde{\boldsymbol{h}}) \cdot \tilde{\boldsymbol{U}} - \rho\tilde{\boldsymbol{\theta}}\right\} = 0$. We then have $\mathrm{E}\big[\tilde{\boldsymbol{U}}\boldsymbol{X}^T\{\tilde{\boldsymbol{\beta}} - \boldsymbol{h}(\tilde{\boldsymbol{\gamma}})\} - \rho\tilde{\boldsymbol{\theta}}\big] = 0$, because $Y - \boldsymbol{X}^T\tilde{\boldsymbol{\beta}} = \epsilon$. Note that $\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)$ is invertible according to Condition (**C4**), then we have $(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{h}}) = \rho\{\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)\}^{-1}\tilde{\boldsymbol{\theta}}$. By the Cauchy-Schwarz inequality and Condition (**C1**), (**C2**) and (**C6**),

$$\begin{aligned}|A_2|^2 &\leqslant \left(\frac{1}{n}\sum_{i=1}^{n} h_n^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p} X_j\left\{\tilde{\beta}_j - h_j(\tilde{\theta}_j)\right\}\right]^2\right) = O_p(q^{-1})O_p\left(\mathrm{E}\left\{\boldsymbol{X}^T(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{h}})\right\}^2\right)\\ &= O_p(q^{-1})O_p\left(\mathrm{E}\left[\rho\boldsymbol{X}^T\left\{\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)\right\}^{-1}\tilde{\boldsymbol{\theta}}\right]^2\right) = O_p(\rho^2\tilde{p}q^{-1}).\end{aligned}$$

Hence, $A_2 = o_p(\tilde{p}q^{-1})$.

Moreover, we have $A_3 = O\Big(-\mathrm{E}_n\big\{\sum_{j=1}^{\tilde{p}} X_j\theta_j^*\big\}^2\Big) = -a_0 O_p(\tilde{p}q^{-1})$ and $A_4 = -O_p(\rho\tilde{p} + \rho a_0 p q^{-1}) = o_p(\tilde{p}q^{-1})$ by Condition (**C6**).

Therefore, we have

$$-\frac{1}{2}T_n'(a_0) = o_p(\tilde{p}q^{-1}) + o_p(\tilde{p}q^{-1}) - a_0 O_p(\tilde{p}q^{-1}) + o_p(\tilde{p}q^{-1}) = -a_0 O_p(\tilde{p}q^{-1}) < 0,$$

if $a_0 > 0$ and $H_n'(a_0) > 0$, if $a_0 < 0$. Thus, $||\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}||_2^2 = o_p(\tilde{p}q^{-1})$. The proof is completed. ■

Proof of Theorem 1:

By the definitions of M_n and M_0, we have

$$
\begin{aligned}
&(\mathrm{M}_n - \mathrm{M}_0)(\boldsymbol{\theta}) \\
=&(\mathrm{E}_n - \mathrm{E})\Big[- \Big\{ Y - \sum_{j=1}^{p} X_j h_j(\theta_j) \Big\}^2 - \rho \sum_{j=1}^{p} \theta_j^2 \Big] \\
=&(\mathrm{E}_n - \mathrm{E})\Big[- \Big(Y - \sum_{j=1}^{p} X_j \tilde{\beta}_j \Big)^2 - \Big\{ \sum_{j=1}^{p} X_j \tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\theta_j) \Big\}^2 - \\
&2(Y - \sum_{j=1}^{p} X_j \tilde{\beta}_j)\Big\{ \sum_{j=1}^{p} X_j \tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\theta_j) \Big\} - \rho \sum_{j=1}^{p} \theta_j^2 \Big] \\
=&(\mathrm{E}_n - \mathrm{E})\Big[- \epsilon^2 - \Big\{ \sum_{j=1}^{p} X_j \tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\theta_j) \Big\}^2 - 2\Big\{ \sum_{j=1}^{p} X_j \tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\theta_j) \Big\}\epsilon - \rho \sum_{j=1}^{p} \theta_j^2 \Big].
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&(\mathrm{M}_n - \mathrm{M}_0)(\boldsymbol{\theta}) - (\mathrm{M}_n - \mathrm{M}_0)(\tilde{\boldsymbol{\theta}}) \\
=&2\mathrm{E}_n\Big[\Big\{ \sum_{j=1}^{p} X_j h_j(\theta_j) - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j) \Big\}\epsilon\Big] - (\mathrm{E}_n - \mathrm{E})\Big\{ \Big[\sum_{j=1}^{p} X_j \big\{ h_j(\theta_j) - h_j(\tilde{\theta}_j) \big\} \Big]^2 \Big\} + \\
&2(\mathrm{E}_n - \mathrm{E})\Big[\sum_{j=1}^{p} X_j \Big\{ h_j(\theta_j) - h_j(\tilde{\theta}_j) \Big\} \Big] \Big[\sum_{j=1}^{p} X_j \Big\{ \tilde{\beta}_j - h_j(\tilde{\theta}_j) \Big\} \Big] - \\
&\rho(\mathrm{E}_n - \mathrm{E}) \left\{ \sum_{j=1}^{p} (\theta_j - \tilde{\theta}_j)(\theta_j + \tilde{\theta}_j) \right\} \\
=&B_1 + B_2 + B_3 + B_4
\end{aligned}
$$

For $j = 1, \ldots, p$, let

$$
\begin{aligned}
G_j &= \Big\{ \theta_j : ||\theta_j - \tilde{\theta}_j||_2 \leqslant \delta, 0 < \delta < 1, \theta_j \in \mathbb{F} \Big\}, \\
H_j &= \Big\{ h_j(\theta_j) : ||\theta_j - \tilde{\theta}_j||_2 \leqslant \delta, 0 < \delta < 1, \theta_j \in \mathbb{F} \Big\}, \\
S_j &= \Big\{ X_j h_j(\theta_j) : ||\theta_j - \tilde{\theta}_j||_2 \leqslant \delta, 0 < \delta < 1, \theta_j \in \mathbb{F} \Big\},
\end{aligned}
$$

and

$$
S = \left\{ \sum_{j=1}^{p} X_j h_j(\theta_j) : ||\theta_j - \tilde{\theta}_j||_2 \leqslant \delta, 0 < \delta < 1, \theta_j \in \mathbb{F}, j = 1, \ldots p \right\}, \tag{S3.5}
$$

where $\mathbb{F} = \big\{ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T : \theta_j \in \mathbb{F}, j = 1, \ldots, p \big\}$.

Since $|h_j(\theta_j) - h_j(\tilde\theta_j)| \leqslant |\theta_j - \tilde\theta_j|$, we have $N_{[\,]}\{\delta_1, H_j, \mathcal{L}_2(\mathbb{D})\} \simeq N_{[\,]}\{\delta_1, G_j, \mathcal{L}_2(\mathbb{D})\}$. By Condition (**C1**), we further have $N_{[\,]}\{(C_2 - C_1)\delta_1, S_j, \mathcal{L}_2(\mathbb{D})\} \simeq N_{[\,]}\{\delta_1, G_j, \mathcal{L}_2(\mathbb{D})\}$.

By Condition (**C6**), we have $\alpha_j > \delta$ for $j = 1, \ldots, p$. Then by the definition of $\tilde\theta_j$, we have

$$S = \left\{ \sum_{j=1}^{\tilde p} X_j h_j(\theta_j) : ||\theta_j - \tilde\theta_j||_2 \leqslant \delta, 0 < \delta < 1, \theta_j \in \mathbb{F}, j = 1, \ldots \tilde p \right\}.$$

According to the construction of S, we have that

$$N_{[\,]}\big(\tilde p(C_2 - C_1)\delta_1, S, \mathcal{L}_2(\mathbb{D})\big) \simeq \big\{N_{[\,]}((C_2 - C_1)\delta_1, S_j, \mathcal{L}_2(\mathbb{D}))\big\}^{\tilde p} \simeq \big\{N_{[\,]}(\delta_1, G_j, \mathcal{L}_2(\mathbb{D}))\big\}^{\tilde p},$$

since the bracket numbers are the same over j for S_j as well as G_j.

From the calculation by Shen and Wong (1994), $\log N_{[\,]}\{\delta_1, G_j, \mathcal{L}_2(\mathbb{D})\} = c_1 q \log(\delta/\delta_1)$, we have $\log N_{[\,]}\{\tilde p(C_2 - C_1)\delta_1, S, \mathcal{L}_2(\mathbb{D})\} \simeq c_1 \tilde p q \log(\delta/\delta_1)$.

By Condition (**C3**), the stochastic process $\Big\{\sqrt{n}\mathrm{E}_n\big[\{\sum_{j=1}^{\tilde p} X_j h_j(\theta_j) - \sum_{j=1}^{\tilde p} X_j h_j(\tilde\theta_j)\}\epsilon\big], \theta_j \in \mathbb{F}, j = 1, \ldots, \tilde p\Big\}$ is sub-Gaussian for the $\mathcal{L}_2(\mathbb{D})$-semimetric on S. According to Corollary 2.2.8 of Van Der Vaart et al. (1996), we have

$$\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2 < \delta, \boldsymbol{\theta} \in \mathbb{F}_n^{Ln}} \sqrt{n}|B_1|\right\} \simeq \int_0^\delta \sqrt{\log N_{[\,]}\{\tilde p\delta_1, S, \mathcal{L}_2(\mathbb{D})\}}\, \mathrm{d}(\tilde p\delta_1) \simeq (\tilde p q)^{1/2}\delta.$$

With the similar calculation of the bracketing number and Lemma 3.4.2 of Van Der Vaart and Wellner (1996) Van Der Vaart et al. (1996), we have

$$\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2 < \delta, \boldsymbol{\theta} \in \mathbb{F}_n^{Ln}} \sqrt{n}|B_2|\right\} \simeq (\tilde p q)^{1/2}\delta.$$

Since $\boldsymbol{X}^T(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{h}}) = \rho\boldsymbol{X}^T\{\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)\}^{-1}\tilde{\boldsymbol{\theta}} = O_p(\rho||\tilde{\boldsymbol{\theta}}||)$ is bounded, we can also have

$$\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2 < \delta, \boldsymbol{\theta} \in \mathbb{F}_n^{Ln}} \sqrt{n}|B_3|\right\} \simeq \mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2 < \delta, \boldsymbol{\theta} \in \mathbb{F}_n^{Ln}} \sqrt{n}|B_1|\right\} \simeq (\tilde p q)^{1/2}\delta.$$

By Condition (**C7**), $|\theta_j + \tilde\theta_j|$ is bounded, then

$$\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2 < \delta, \boldsymbol{\theta} \in \mathbb{F}_n^{Ln}} \sqrt{n}|B_4|\right\} \simeq \mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2 < \delta, \boldsymbol{\theta} \in \mathbb{F}_n^{Ln}} \sqrt{n}|B_1|\right\} \simeq (\tilde p q)^{1/2}\delta.$$

According to Theorem 3.4.1 of Van Der Vaart and Wellner (1996) Van Der Vaart et al. (1996), the key function $\phi(\delta)$ takes the form of $\phi_n(\delta) = (\tilde p q)^{1/2}\delta$. Therefore, $||\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}||_2 = O_p\big((\tilde p q/n)^{1/2}\big)$.

By Lemma 3 and Condition (**C6**), If $\beta_j \notin \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, then

$$\begin{aligned}
||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 &= ||\zeta_{(\hat{\boldsymbol{\theta}},\alpha)} - \boldsymbol{\beta}_0||_2 \\
&\leqslant ||\zeta_{(\hat{\boldsymbol{\theta}},\alpha)} - h(\hat{\boldsymbol{\theta}})||_2 + ||h(\hat{\boldsymbol{\theta}}) - h(\tilde{\boldsymbol{\theta}})||_2 + ||h(\tilde{\boldsymbol{\theta}}) - \tilde{\boldsymbol{\beta}}||_2 + ||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 \\
&= O_p(\tilde{p}^{1/2}\nabla(\eta)) + O_p(||\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}||_2) + O(\tilde{p}^{1/2}\nabla(\eta)) + O((\tilde{p}\rho + \tilde{p}q^{-2m})^{1/2}) \\
&= O_p\left(\tilde{p}^{1/2}\nabla(\eta) + (\tilde{p}q/n)^{1/2} + \tilde{p}^{1/2}\nabla(\eta) + (\tilde{p}\rho + \tilde{p}q^{-2m})^{1/2}\right) \\
&= O_p\left((\tilde{p}q/n)^{1/2} + \tilde{p}^{1/2}q^{-m}\right);
\end{aligned}$$

if $\beta_j \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_p\left(\tilde{p}^{1/2}\nabla(\eta) + (\tilde{p}q/n)^{1/2} + (\tilde{p}\rho)^{1/2}\right)$
$= O_p\left(\tilde{p}^{1/2}r(\eta) + (\tilde{p}q/n)^{1/2}\right)$. The proof is completed. ■

LEMMA 6 (Normality): *Under Conditions (**C1**)–(**C7**), for $j = 1, \ldots, p$, and any $w \in \mathbb{D}$,*

$$\{\sigma^2_{nj}(w)\}^{-1/2}\left\{\hat{\theta}_j(w) - \tilde{\theta}_j(w)\right\} \to_d N(0, 1),$$

where $\sigma^2_{nj}(w) = \sigma^2[n^2\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\}^T\{l''_n(\tilde{\boldsymbol{\gamma}})\}^{-1}\left\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\right\}\{l''_n(\tilde{\boldsymbol{\gamma}})\}^{-1}\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\}]^{-1}$.

Proof of Lemma 6:

By the Mean Value Theorem, there exists a $\boldsymbol{\gamma}^*$ between $\tilde{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}}$, such that

$$0 = l'_n(\hat{\boldsymbol{\gamma}}) = l'_n(\tilde{\boldsymbol{\gamma}}) + l''_n(\boldsymbol{\gamma}^*)(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}). \tag{S3.6}$$

According to the previous calculation,

$$\begin{aligned}
l'_n(\boldsymbol{\gamma}) &= -2\mathrm{E}_n\left\{(Y - \boldsymbol{X}^T\boldsymbol{h}) \cdot \boldsymbol{U} \otimes \boldsymbol{B}(W) - \rho\boldsymbol{\theta} \otimes \boldsymbol{B}(W)\right\} \\
&= -2\mathrm{E}_n\left\{\boldsymbol{U} \otimes \boldsymbol{B}(W)\epsilon + \boldsymbol{U} \otimes \boldsymbol{B}(W) \cdot \boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{h}) - \rho\boldsymbol{\theta} \otimes \boldsymbol{B}(W)\right\} \qquad \text{(S3.7)} \\
&= -2\mathrm{E}_n\left\{\boldsymbol{v}\epsilon + \boldsymbol{v} \cdot \boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{h}) - \rho\boldsymbol{\theta} \otimes \boldsymbol{B}(W)\right\}.
\end{aligned}$$

Since $l''_n(\boldsymbol{\gamma}^*)$ is invertible, then we have $\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}} = -\{l''_n(\boldsymbol{\gamma}^*)\}^{-1}l'_n(\tilde{\boldsymbol{\gamma}})$. To prove the theorem, it suffices to show that for any $\boldsymbol{c}_n \in \mathbb{R}_{q*p}$ whose components are not all zero and $\boldsymbol{c}_n^T\boldsymbol{c}_n = O_p(q)$, $\boldsymbol{c}_n^T(\hat{\gamma} - \tilde{\gamma})/\mathrm{SD}\left\{\boldsymbol{c}_n^T(\hat{\gamma} - \tilde{\gamma})\right\} \to_d N(0, 1)$, where

$$\mathrm{SD}\left\{\boldsymbol{c}_n^T(\hat{\gamma} - \tilde{\gamma})\right\} = \sqrt{(1/n^2)\boldsymbol{c}_n^T\{l''_n(\tilde{\boldsymbol{\gamma}})\}^{-1}\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\}\{l''_n(\tilde{\boldsymbol{\gamma}})\}^{-1}\boldsymbol{c}_n\sigma^2}.$$

By some algebra, we have

$$\begin{aligned}\boldsymbol{c}_n^T(\hat{\gamma}-\tilde{\gamma}) =& -\boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}l_n'(\tilde{\gamma})\\ =&\sum_{i=1}^n a_i\epsilon_i^* + \boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\mathrm{E}_n\big\{\boldsymbol{v}(\tilde{\gamma})\cdot\boldsymbol{X}^T(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{h}})-\rho\tilde{\boldsymbol{\theta}}\otimes\boldsymbol{B}(W)\big\}\\ =&A_1+A_2,\end{aligned}$$

where $a_i = \boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\boldsymbol{v}_i(\tilde{\gamma})\sigma/n$ and ϵ_i^* are independent with mean zero and variance one conditioning on $\{\theta_i, W_i, i=1,\dots,n\}$.

Since $\mathrm{E}_n\big\{\tilde{\boldsymbol{U}}\otimes\boldsymbol{B}(W)\boldsymbol{X}^T(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{h}})-\rho\tilde{\boldsymbol{\theta}}\otimes\boldsymbol{B}(W)\big\} = \rho\mathrm{E}_n\bigg\{\Big[\tilde{\boldsymbol{U}}\boldsymbol{X}^T\big\{\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)\big\}^{-1}\tilde{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}\Big]\otimes\boldsymbol{B}\bigg\}$, we have $A_2 = o_p(\rho q^{1/2})$. Moreover,

$$\begin{aligned}\sum_{i=1}^n a_i^2 =& \frac{\sigma^2}{n^2}\sum_{i=1}^n \boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\boldsymbol{v}_i(\tilde{\gamma})\boldsymbol{v}_i^T(\tilde{\gamma})\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\boldsymbol{c}_n\\ =& \frac{\sigma^2}{n}\boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\frac{1}{n}\sum_{i=1}^n\boldsymbol{v}_i(\tilde{\gamma})\boldsymbol{v}_i^T(\tilde{\gamma})\cdot\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\boldsymbol{c}_n\\ =& O_p(\boldsymbol{c}_n^T\boldsymbol{c}_n/n) = O_p(q/n),\end{aligned}$$

thus we have $A_2/\sqrt{(\sum a_i^2)} = o_p(n\rho/q) = o_p(1)$ by Condition (**C6**).

By Slutsky's Theorem, we then only need to prove $A_1/\sqrt{(\sum a_i^2)}$ follows a Normal distribution. By Condition (**C3**) and Lemma 4, we only need to verify that $\max_i a_i^2/\sum_{i=1}^n a_i^2 \to_p 0$. With some calculations, we have

$$\begin{aligned}\max_{1\leqslant i\leqslant n} a_i^2 =& \frac{\sigma^2}{n^2}\max_{1\leqslant i\leqslant n}\Big[\boldsymbol{c}_n^T\left\{-l_n''(\boldsymbol{\gamma}^*)\right\}^{-1}\left\{\boldsymbol{V}_n^T(\tilde{\gamma})\boldsymbol{V}_n(\tilde{\gamma})\right\}^{1/2}\left\{\boldsymbol{V}_n^T(\tilde{\gamma})\boldsymbol{V}_n(\tilde{\gamma})\right\}^{-1/2}\boldsymbol{v}_i(\tilde{\gamma})\Big]^2\\ \leqslant& \frac{\sigma^2}{n^2}\boldsymbol{c}_n^T\left\{l_n''(\boldsymbol{\gamma}^*)\right\}^{-1}\cdot\boldsymbol{V}_n^T(\tilde{\gamma})\boldsymbol{V}_n(\tilde{\gamma})\cdot\left\{l_n''(\boldsymbol{\gamma}^*)\right\}^{-1}\boldsymbol{c}_n\cdot\\ &\max_{1\leqslant i\leqslant n}\boldsymbol{v}_i^T(\tilde{\gamma})\left\{\boldsymbol{V}_n^T(\tilde{\gamma})\boldsymbol{V}_n(\tilde{\gamma})\right\}^{-1}\boldsymbol{v}_i(\tilde{\gamma}).\end{aligned}$$

According to Condition (**C2**), we have

$$\frac{\max_i a_i^2}{\sum_{i=1}^n a_i^2} = \max_{1\leqslant i\leqslant n}\boldsymbol{v}_i^T(\boldsymbol{V}_n^T\boldsymbol{V}_n)^{-1}\boldsymbol{v}_i \to_p 0,$$

as $n\to\infty$.

Because $\hat{\boldsymbol{\gamma}}\to_p\tilde{\boldsymbol{\gamma}}$, we have $\boldsymbol{\gamma}^*\to_p\tilde{\boldsymbol{\gamma}}$. Since for any $w\in\mathbb{D}$, $\hat{\theta}_j(w) = (\boldsymbol{e}_j\otimes\boldsymbol{B}(w))^T\hat{\boldsymbol{\gamma}}$, then let $\boldsymbol{c}_n = \boldsymbol{e}_j \otimes \boldsymbol{B}(w)$, we have

$$\{\sigma_{nj}^2(w)\}^{-1/2}\left\{\hat{\theta}_j(w) - \tilde{\theta}_j(w)\right\} \rightarrow_d N(0,1),$$

where $\sigma_{nj}^2(w) = \sigma^2/n^2\left\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\right\}^T\{l_n''(\tilde{\gamma})\}^{-1}\left\{\boldsymbol{V}_n^T(\tilde{\gamma})\boldsymbol{V}_n(\tilde{\gamma})\right\}\{l_n''(\tilde{\gamma})\}^{-1}\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\}$. The proof is completed. ■

Proof of Theorem 2:

It is straightforward to show that if $(Z-\mu)/\sigma \sim N(0,1)$, then

$$\Pr\left\{\zeta_{(Z,\alpha)} < x\right\} = \Phi\left(\frac{x+\alpha-\mu}{\sigma}\right)\mathcal{I}(x \geqslant 0) + \Phi\left(\frac{x-\alpha-\mu}{\sigma}\right)\mathcal{I}(x < 0).$$

Under regularity conditions and by Lemma 6, for $1 \leqslant j \leqslant p$ and any $w \in \mathbb{D}$, we have $\lim_{n\to\infty}\Pr\left(\sigma_{nj}^{-1}\hat{\theta}_j(w) - \sigma_{nj}^{-1}\tilde{\theta}_j(w) < x\right) = \Phi(x)$. Note that $\sigma_{nj}^{-1}\zeta_{\left\{\hat{\theta}_j,\alpha_j\right\}}(w) = \zeta_{\left\{\sigma_{nj}^{-1}\hat{\theta}_j,\sigma_{nj}^{-1}\alpha_j\right\}}(w)$, then we have

$$\begin{aligned}
&\lim_{n\to\infty}\left|\Pr\left[\zeta_{\left\{\hat{\theta}_j,\alpha_j\right\}}(w) \leqslant x\right] - \Phi\left(\frac{x+\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}}\right)\mathcal{I}(x \geqslant 0) - \right.\\
&\qquad\left.\Phi\left(\frac{x-\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}}\right)\mathcal{I}(x < 0)\right|\\
&= \lim_{n\to\infty}\left|\Pr\left[\zeta_{\left\{\sigma_{nj}^{-1}\hat{\theta}_j,\sigma_{nj}^{-1}\alpha_j\right\}}(w) \leqslant \sigma_{nj}^{-1}x\right] - \Phi\left\{\frac{x+\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}}\right\}\mathcal{I}(x \geqslant 0) - \right.\\
&\qquad\left.\Phi\left\{\frac{x-\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}}\right\}\mathcal{I}(x < 0)\right|\\
&= 0.
\end{aligned}$$

■

Proof of Theorem 3:

Let $u_{nj}^* = \hat{\theta}_j - \hat{\sigma}_{nj}z_{\xi/2}$ and $v_{nj}^* = \hat{\theta}_j + \hat{\sigma}_{nj}z_{\xi/2}$.

(a). When $P_+ > \xi/2$ and $P_- > \xi/2$, or $P_- < \xi/2$ and $P_+ > 1-\xi/2$, or $P_+ < \xi/2$ and $P_- > 1-\xi/2$, then $\zeta_{(u_{nj}^*,\alpha_j)} \neq 0$ and $\zeta_{(v_{nj}^*,\alpha_j)} \neq 0$. Therefore $\hat{\theta}_j - \hat{\sigma}_{nj}z_{\xi/2} \leqslant \tilde{\theta}_j \leqslant \hat{\theta}_j + \hat{\sigma}_{nj}z_{\xi/2}$

is equivalent to $\zeta(\hat{\theta}_j, \alpha_j) - \hat{\sigma}_{nj} z_{\xi/2} \leqslant \zeta_{(\tilde{\theta}_j,\alpha_j)} \leqslant \zeta(\hat{\theta}_j, \alpha_j) + \hat{\sigma}_{nj} z_{\xi/2}$. Therefore,

$$\begin{aligned}
&\lim_{n\to\infty} \Pr\left\{\zeta_{(\hat{\theta}_j,\alpha_j)} - \hat{\sigma}_{nj} z_{\xi/2} \leqslant \zeta_{(\tilde{\theta}_j,\alpha_j)} \leqslant \zeta_{(\hat{\theta}_j,\alpha_j)} + \hat{\sigma}_{nj} z_{\xi/2}\right\}\\
=&\lim_{n\to\infty} \Pr\left(\hat{\theta}_j - \hat{\sigma}_{nj} z_{\xi/2} \leqslant \tilde{\theta}_j \leqslant \hat{\theta}_j + \hat{\sigma}_{nj} z_{\xi/2}\right)\\
=&\lim_{n\to\infty} \Pr\left(\tilde{\theta}_j - \hat{\sigma}_{nj} z_{\xi/2} \leqslant \hat{\sigma}_{nj}^{-1}\hat{\theta}_j \leqslant \tilde{\theta}_j + \hat{\sigma}_{nj} z_{\xi/2}\right)\\
=&1 - \xi.
\end{aligned}$$

That is, $\left[\zeta_{(\hat{\theta}_j,\alpha_j)} - \hat{\sigma}_{nj} z_{\xi/2}, \zeta_{(\hat{\theta}_j,\alpha_j)} + \hat{\sigma}_{nj} z_{\xi/2}\right]$ is the $1 - \xi$ confidence interval for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$.

(b). When $P_+ < \xi/2$ and $\xi - P_+ < P_- < 1 - \xi/2$, then $\zeta_{(u^*_{nj},\alpha_j)} \neq 0$ and $\zeta_{(v^*_{nj},\alpha_j)} = 0$. Let $A = \hat{\sigma}_{nj}^{-1}\alpha_j + \delta_0 - \hat{\sigma}_{nj}^{-1}\hat{\theta}_j$ and B satisfy $\Pr(z < -A) + \Pr(z > B) = \xi$, where $z \sim N(0,1)$ and $\delta_0 > 0$ is small enough such that $\hat{\sigma}_{nj}^{-1}\hat{\theta} - B < -\hat{\sigma}_{nj}^{-1}\alpha_j$. Then, $\lim_{n\to\infty} \Pr\left\{-A \leqslant \hat{\sigma}_{nj}^{-1}(\hat{\theta} - \tilde{\theta}_j) \leqslant B\right\} = 1 - \xi$, i.e. $\lim_{n\to\infty} \Pr(\hat{\theta}_j - \hat{\sigma}_{nj} B \leqslant \tilde{\theta}_j \leqslant \hat{\theta}_j + \hat{\sigma}_{nj} A) = 1 - \xi$. By the definitions of A and B, we have $\zeta_{(\hat{\theta}_j + \hat{\sigma}_{nj} A,\alpha_j)} > 0$ and $\zeta_{(\hat{\theta}_j - \hat{\sigma}_{nj} B)} < 0$. Therefore, similar to part (a), we have,

$$\begin{aligned}
&\lim_{n\to\infty} \Pr\left\{\zeta_{(\hat{\theta},\alpha_j)} - \hat{\sigma}_{nj} B \leqslant \zeta_{(\tilde{\theta}_j,\alpha_j)} \leqslant \hat{\sigma}_{nj}\delta_0\right\}\\
=&\lim_{n\to\infty} \Pr\left(\hat{\theta}_j - \hat{\sigma}_{nj} B \leqslant \tilde{\theta}_j \leqslant \hat{\theta}_j + \hat{\sigma}_{nj} A\right)\\
=&1 - \xi.
\end{aligned}$$

Then, $\left[\zeta_{(\hat{\theta},\alpha_j)} - \hat{\sigma}_{nj} B, \hat{\sigma}_{nj}\delta_0\right]$ is the $1 - \xi$ confidence interval for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$, where
$B = \Phi^{-1}\left\{1 - \xi + \Phi(-\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j + \delta_0)\right\}$.

(c). When $P_- < \xi/2$ and $\xi - P_- < P_+ < 1 - \xi/2$, then $\zeta_{(u^*_{nj},\alpha_j)} = 0$ and $\zeta_{(v^*_{nj},\alpha_j)} \neq 0$. Let $B = \hat{\sigma}_{nj}^{-1}\alpha_j + \delta_0 + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j$ and A satisfy $\Pr(z < -A) + \Pr(z > B) = \xi$, where $z \sim N(0,1)$ and $\delta_0(\delta_0 > 0)$ is small enough such that $\hat{\sigma}_{nj}^{-1}\hat{\theta} + A > \hat{\sigma}_{nj}^{-1}\alpha_j$. Similar to part (b), we have

$$\begin{aligned}
&\lim_{n\to\infty} \Pr\left\{-\hat{\sigma}_{nj}\delta_0 \leqslant \zeta_{(\tilde{\theta}_j,\alpha_j)} \leqslant \zeta_{(\hat{\theta},\alpha_j)} + \hat{\sigma}_{nj} A\right\}\\
=&\lim_{n\to\infty} \Pr\left(\hat{\sigma}_{nj}^{-1}\tilde{\theta} - A \leqslant \hat{\sigma}_{nj}^{-1}\hat{\theta}_j \leqslant \hat{\sigma}_{nj}^{-1}\tilde{\theta} + B\right)\\
=&1 - \xi.
\end{aligned}$$

Then, $\left[-\hat{\sigma}_{nj}\delta_0, \zeta_{(\hat{\theta},\alpha_j)} + \hat{\sigma}_{nj} A\right]$ is the $1 - \xi$ confidence interval for $\zeta_{(\tilde{\theta}_j),\alpha_j}$, where
$A = -\Phi^{-1}\left\{\xi - 1 + \Phi(\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j + \delta_0)\right\}$.

(d). When $P_+ + P_- < \xi$, then $\zeta_{(u^*_{nj},\alpha_j)} = 0$ and $\zeta_{(v^*_{nj},\alpha_j)} = 0$. Therefore, $\hat{\sigma}_{nj}^{-1}\hat{\theta}_j - z_{\xi/2} \leqslant \hat{\sigma}_{nj}^{-1}\tilde{\theta}_j \leqslant \hat{\sigma}_{nj}^{-1}\hat{\theta}_j + z_{\xi/2}$ implies that $0 = \zeta_{(\hat{\sigma}_{nj}^{-1}\hat{\theta}_j - z_{\xi/2}, \hat{\sigma}_{nj}^{-1}\alpha_j)} \leqslant \zeta_{(\hat{\sigma}_{nj}^{-1}\tilde{\theta}_j, \hat{\sigma}_{nj}^{-1}\alpha_j)} \leqslant \zeta_{(\hat{\sigma}_{nj}^{-1}\hat{\theta}_j + z_{\xi/2}, \hat{\sigma}_{nj}^{-1}\alpha_j)} = 0$. Therefore, $\Pr\left\{\zeta_{(\tilde{\theta}_j,\alpha_j)} = 0\right\} \geqslant \lim_{n\to\infty} \Pr(\hat{\sigma}_{nj}^{-1}\tilde{\theta}_j - z_{\xi/2} \leqslant \hat{\sigma}_{nj}^{-1}\hat{\theta}_j \leqslant \hat{\sigma}_{nj}^{-1}\tilde{\theta}_j + z_{\xi/2}) = \lim_{n\to\infty} \Pr(\hat{\sigma}_{nj}^{-1}\tilde{\theta}_j - z_{\xi/2} \leqslant \hat{\sigma}_{nj}^{-1}\hat{\theta}_j \leqslant \hat{\sigma}_{nj}^{-1}\tilde{\theta}_j + z_{\xi/2}) = 1 - \xi$. Then $[0, 0]$ is a confidence interval for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$ with at least $1 - \xi$ coverage probability.

As δ_0 in (b) and (c) can be arbitrarily small, the results remain valid when δ_0 goes to 0. Let $\delta_0 \to 0$, then the confidence interval for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$ with at least $1 - \xi$ coverage probability is

$$
\begin{aligned}
&[u_{nj}(w), v_{nj}(w)] \\
&= \begin{cases}
\left[\hat{\beta}_j(w) - \hat{\sigma}_{nj} z_{\xi/2}, \hat{\beta}_j(w) + \hat{\sigma}_{nj} z_{\xi/2}\right], & P_+ > \xi/2 \text{ and } P_- > \xi/2, \\
 & \text{or } P_- < \xi/2 \text{ and } P_+ > 1 - \xi/2, \\
 & \text{or } P_+ < \xi/2 \text{ and } P_- > 1 - \xi/2 \\
\left[\hat{\beta}_j(w) - \hat{\sigma}_{nj}\hat{B}, 0\right], & P_+ < \xi/2 \text{ and } \xi - P_+ < P_- < 1 - \xi/2 \\
\left[0, \hat{\beta}_j(w) + \hat{\sigma}_{nj}\hat{A}\right], & P_- < \xi/2 \text{ and } \xi - P_- < P_+ < 1 - \xi/2 \\
[0, 0], & P_+ + P_- < \xi
\end{cases},
\end{aligned}
\tag{S3.8}
$$

where $\hat{A} = -\Phi^{-1}\left\{\xi - 1 + \Phi(\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j)\right\}$ and $\hat{B} = \Phi^{-1}\left\{1 - \xi + \Phi(-\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j)\right\}$.

Since the bias $\beta_j - \zeta_{(\tilde{\theta}_j,\alpha_j)}$ is asymptotically negligible relative to the variance of $\hat{\theta}_j$, and $\hat{P}_+ \to P_+$ and $\hat{P}_- \to P_-$ as $n \to \infty$, the asymptotic $1 - \xi$ confidence interval (S3.8) for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$ is also an asymptotic $1 - \xi$ confidence interval for β_j with P_+ and P_- replaced by $\hat{P}_+$ and $\hat{P}_-$.

When $\beta_j(w) \neq 0$, the boundary points will not be zero as we defined in (a) and the limiting coverage probability is $1 - \epsilon$. When $\beta_j(w) = 0$, since $\hat{\beta}_j(w) \to \beta_j(w)$ as $n \to \infty$. Therefore, there exists $N > 0$ such that when $n > N$, $P_+ < \epsilon/2$ or $P_- < \epsilon/2$ and $P_+ + P_- < 1 - \epsilon/2$ by their definition. Then $u_{nj}(w) = 0$ and (or) $v_{nj}(w) = 0$. We have

$$\begin{aligned}
\Pr(u_{nj} = 0 \text{ or } v_{nj} = 0) &= \Pr\left\{\zeta_{(u^*_{nj},\alpha_j)} = \zeta_{(v^*_{nj},\alpha_j)} = 0\right\}\\
&= \Pr\left\{|\hat{\theta}_j - \hat{\sigma}_{nj} z_{\xi/2}| \leqslant \alpha_j \text{ or } |\hat{\theta}_j + \hat{\sigma}_{nj} z_{\xi/2}| \leqslant \alpha_j\right\}\\
&= \Pr\big\{-\alpha_j + \hat{\sigma}_{nj} z_{\xi/2} \leqslant \hat{\theta}_j \leqslant \alpha_j + \hat{\sigma}_{nj} z_{\xi/2} \text{ or}\\
&\qquad -\alpha_j - \hat{\sigma}_{nj} z_{\xi/2} \leqslant \hat{\theta}_j \leqslant \alpha_j - \hat{\sigma}_{nj} z_{\xi/2}\big\}\\
&\geqslant \Pr\big\{-\alpha_j + \hat{\sigma}_{nj} z_{\xi/2} \leqslant \hat{\theta}_j \leqslant \alpha_j + \hat{\sigma}_{nj} z_{\xi/2}\big\}\\
&> 0.
\end{aligned}$$

Therefore, $[u_{nj}, v_{nj}]$ is a sparse confidence interval for β_j. ■

S4. Detailed implementation of a local FDR control-based bootstrap procedure to infer turning points

We estimate the turning points of varying coefficient functions based on our STV model and further construct the confidence intervals using a bootstrap method. To ensure the confidence intervals have proper coverage and eliminate the influence of potential outliers, we adopt the percentile-t method (Hall, 1992), in conjunction with a local false discovery rate (FDR) control method (Efron et al., 2015). The detailed steps are as follows.

(1) We fit the STV model to each original dataset and estimate the left turning point (e_1) and right turning point (e_2). Specifically, the turning points are identified, respectively, by finding the value of w where the first change occurs in the sign of $\beta(w)$, transitioning from greater than 0 to smaller than or equal to 0, and the value of w where the last change occurs in the sign of $\beta(w)$, transitioning from smaller than or equal to 0 to greater than 0.

(2) We generate 200 bootstrap datasets for each original dataset by sampling with replacement. For each bootstrap dataset, we fit the model and calculate the left and right turning points (e_1 and e_2) using the same procedure as in Step (1).

(3) We use the “locfdr” R package (Efron et al., 2015) to remove the potential outlier cases in the

bootstrap estimates by setting the local FDR to be 0.1, and estimate the null distributions for e_1 and e_2 separately. We compute the means and standard deviations of the null distributions.

(4) With the the means and standard deviations computed from Step (3), we standardize the 200 bootstrap estimates of e_1 and e_2, and then apply the percentile-t method (Hall, 1992) to compute their 95% confidence intervals.

S5. Additional simulation analysis for low dimensional covariates

Let $|A|$ denote the cardinality of set A. To compare zero-effect region detection, we define two quantities, estimation-based true positive ratio and estimation-based true negative ratio:

$$\mathrm{ETPR}(\beta) = \frac{|\{w : \hat{\beta}(w) \neq 0 \text{ and } \beta(w) \neq 0\}|}{|\{w : \beta(w) \neq 0\}|},$$

$$\mathrm{ETNR}(\beta) = \frac{|\{w : \hat{\beta}(w) = 0 \text{ and } \beta(w) = 0\}|}{|\{w : \beta(w) = 0\}|}.$$

Since the B-spline and local polynomial methods do not yield exactly zero estimates, the above definitions are not applicable. Instead, we introduce inference-based true positive ratio and true negative ratio:

$$\mathrm{ITPR}(\beta) = \frac{|\{w : 0 \notin \mathrm{CI}\{\hat{\beta}(w)\} \text{ and } \beta(w) \neq 0\}|}{|\{w : \beta(w) \neq 0\}|},$$

$$\mathrm{ITNR}(\beta) = \frac{|\{w : 0 \in \mathrm{CI}\{\hat{\beta}(w)\} \text{ and } \beta(w) = 0\}|}{|\{w : \beta(w) = 0\}|},$$

where $\mathrm{CI}\{\hat{\beta}(w)\}$ is the 95% confidence interval of $\beta(w)$.

We choose 100 grid points on $[0, 3]$ and count the number of W in each set as its cardinality. The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) is adopted to control the false discovery rate in the calculation of the inference-based true positive ratio and the inference-based true negative ratio. Table 1 shows that the soft-thresholded varying coefficient model has higher inference-based true negative ratios than the B-spline varying coefficient model and the local polynomial varying coefficient model, and the performance of our method is improving as n becomes larger. We also compare the non-zero-effect region selection accuracy between our estimation-based method and our inference-based method in

Table 2. The estimation-based true positive ratio is slightly higher than the inference-based true positive ratio, but both of them quickly approach to 1 as n increases. Of note, the estimation-based method is much faster than the inference-based method.

Figure 1 shows the coverage probability of β_1 at each grid point for all of the three methods when $n = 500$. The soft-thresholded varying coefficient model makes more accurate inference on zero-effect regions and non-zero-effect regions, as the coverage probabilities are closer to 95% on average compared to the others. At the transitions between zero- and nonzero-effect regions, all the methods draw less accurate inference, but our method still outperforms the competing methods. Specifically, the B-spline varying coefficient model and the local polynomial varying coefficient model have considerably small coverage probabilities around 50% to 60%, while our method can still achieve a coverage probability of at least 80%.

S6. Comparison of performance with misspecified models, i.e., with zero-crossing varying coefficients

We have conducted a simulation study to compare the performance of our proposed method with the regular B-spline varying coefficient model when the varying coefficients are zero-crossing. The simulation settings are the same as in Section 4.1 in the main text, except that the true coefficient functions are

$\beta_1(w) = -w^2/2 + 3, \beta_2(w) = 2\log(w + 0.1)$, and $\beta_3(w) = -6/(w + 1) + 2$, which are all zero-crossing smooth functions. The following Table 3 shows the comparison of estimation accuracy between regular varying coefficient model and STV model when we choose $n = 200, 500$, and 1000. Using the integrated squared errors and the average integrated squared errors as the criteria, STV performs as well as the regular B-spline varying coefficient model in most cases, possibly because any smooth varying coefficients that cross zero can be well approximated by functions in our specified H functional space as shown by Lemma 2.

S7. Additional results for preoperative opioid study

Additional plots for real data application are provided in this section.

References

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 57*(1), 289–300.

Efron, B., B. Turnbull, and B. Narasimhan (2015). *locfdr: Computes Local False Discovery Rates*. R package version 1.1-8.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics 27*(5), 1491–1518.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *The Annals of Statistics 20*(2), 675 – 694.

Huang, J. Z. and H. Shen (2004). Functional coefficient regression models for non-linear time series: a polynomial spline approach. *Scandinavian journal of statistics 31*(4), 515–534.

Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*(1), 111–128.

Huang, J. Z., C. O. Wu, and L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica 14*(3), 763–788.

Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics 22*, 580–615.

Van Der Vaart, A. W., J. A. Wellner, A. W. van der Vaart, and J. A. Wellner (1996). *Weak convergence*. Springer.

Wei, F., J. Huang, and H. Li (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica 21*(4), 1515.

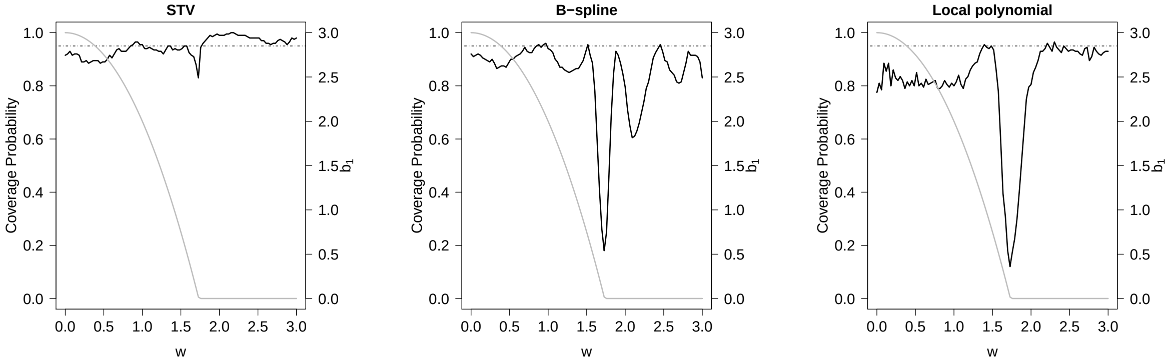

Figure 1: Empirical coverage probabilities (black curves) of the soft-thresholding varying coefficient model (STV), the regular B-spline varying coefficient model (B-spline) and the local polynomial varying coefficient model (local polynomial) in low dimensional covariates simulations. The grey curves are the true values of varying coefficients. The horizontal lines indicate the target coverage probability of 0·95.

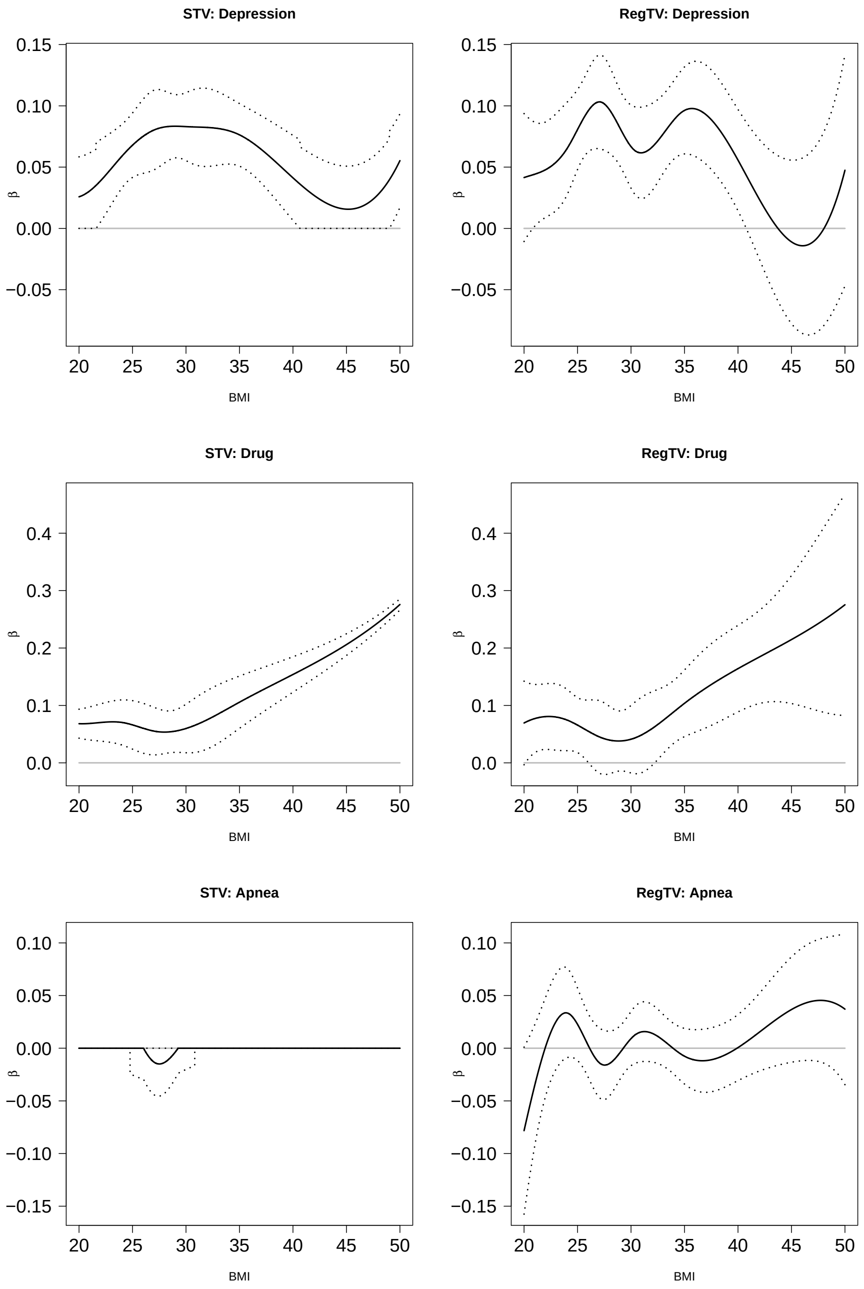

Figure 2: Estimation results (II) for the preoperative opioid use data using the B-spline method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

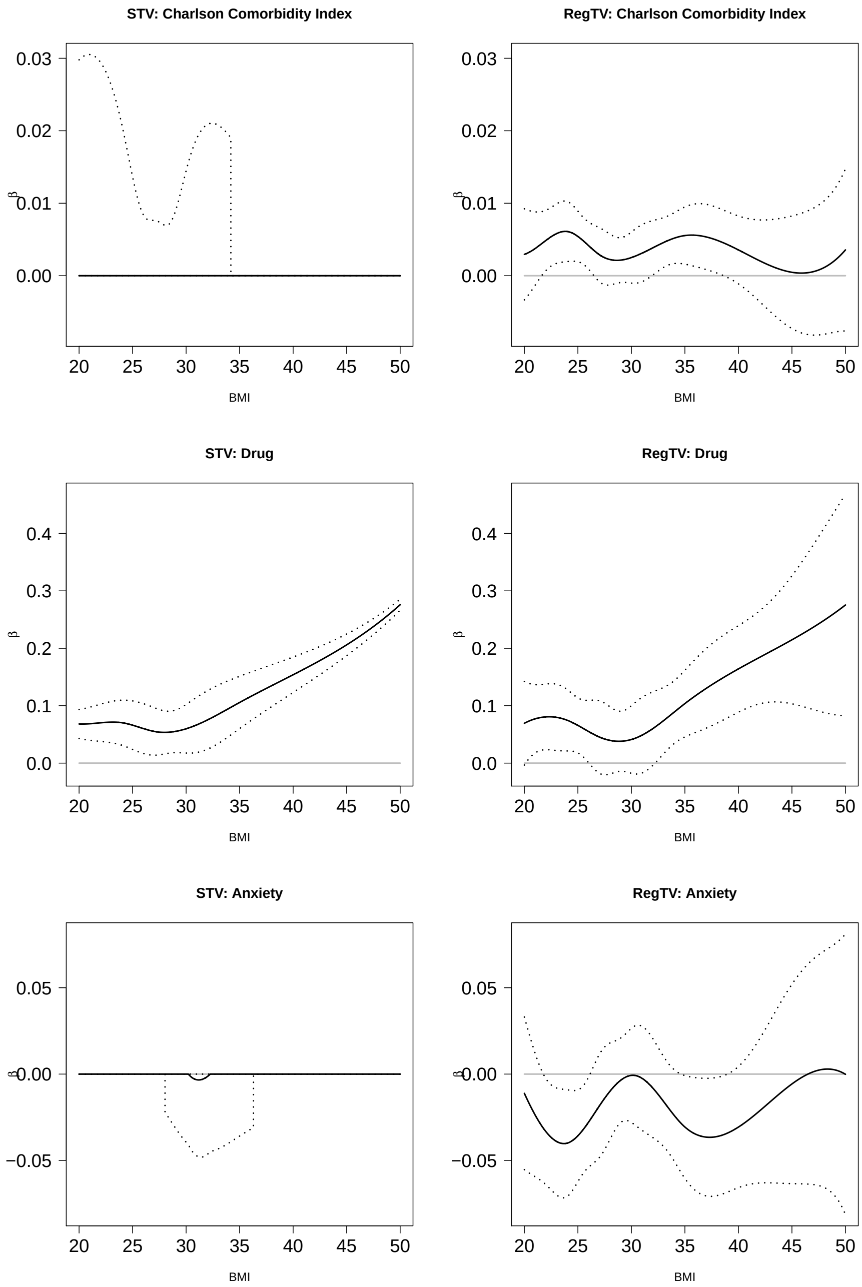

Figure 3: Estimation results (III) for the preoperative opioid use data using the B-spline method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

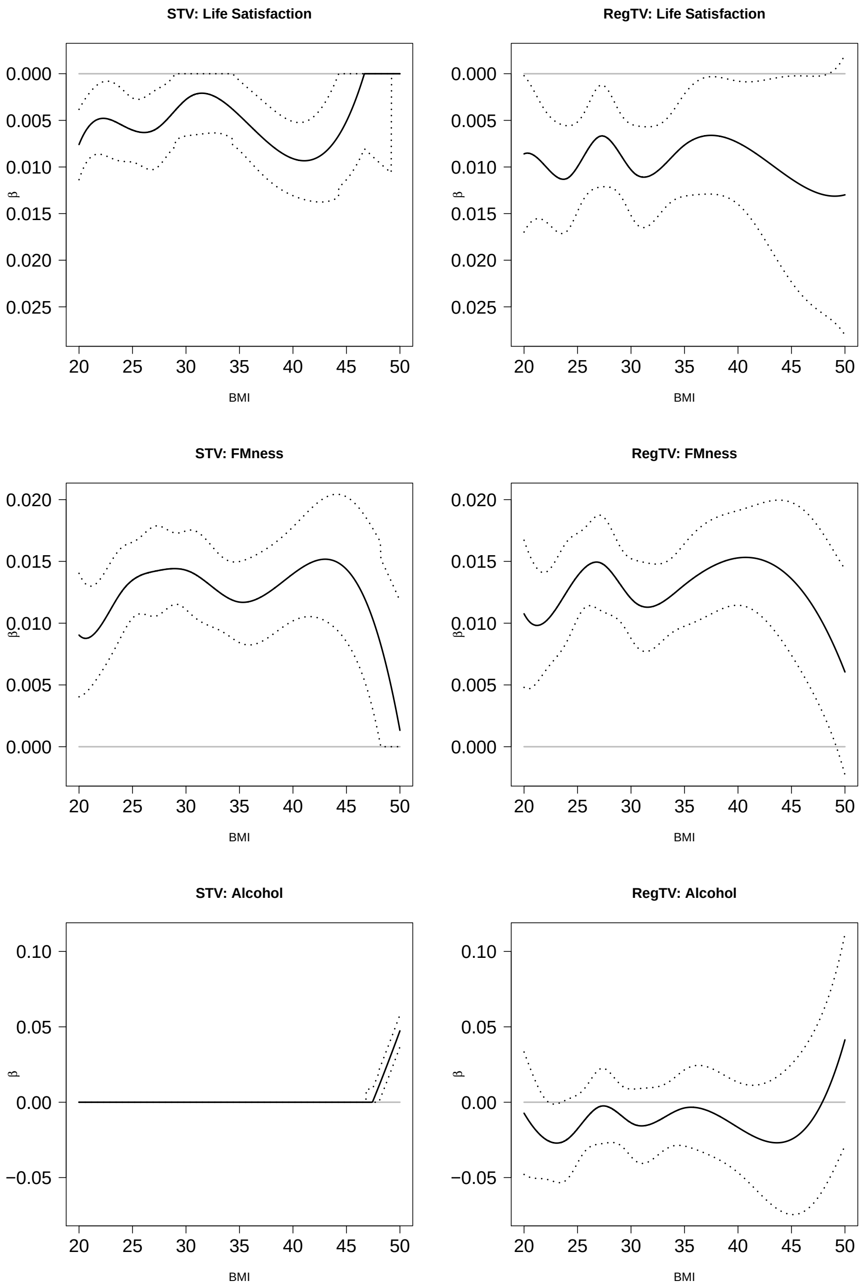

Figure 4: Estimation results (IV) for the preoperative opioid use data using the B-spline method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

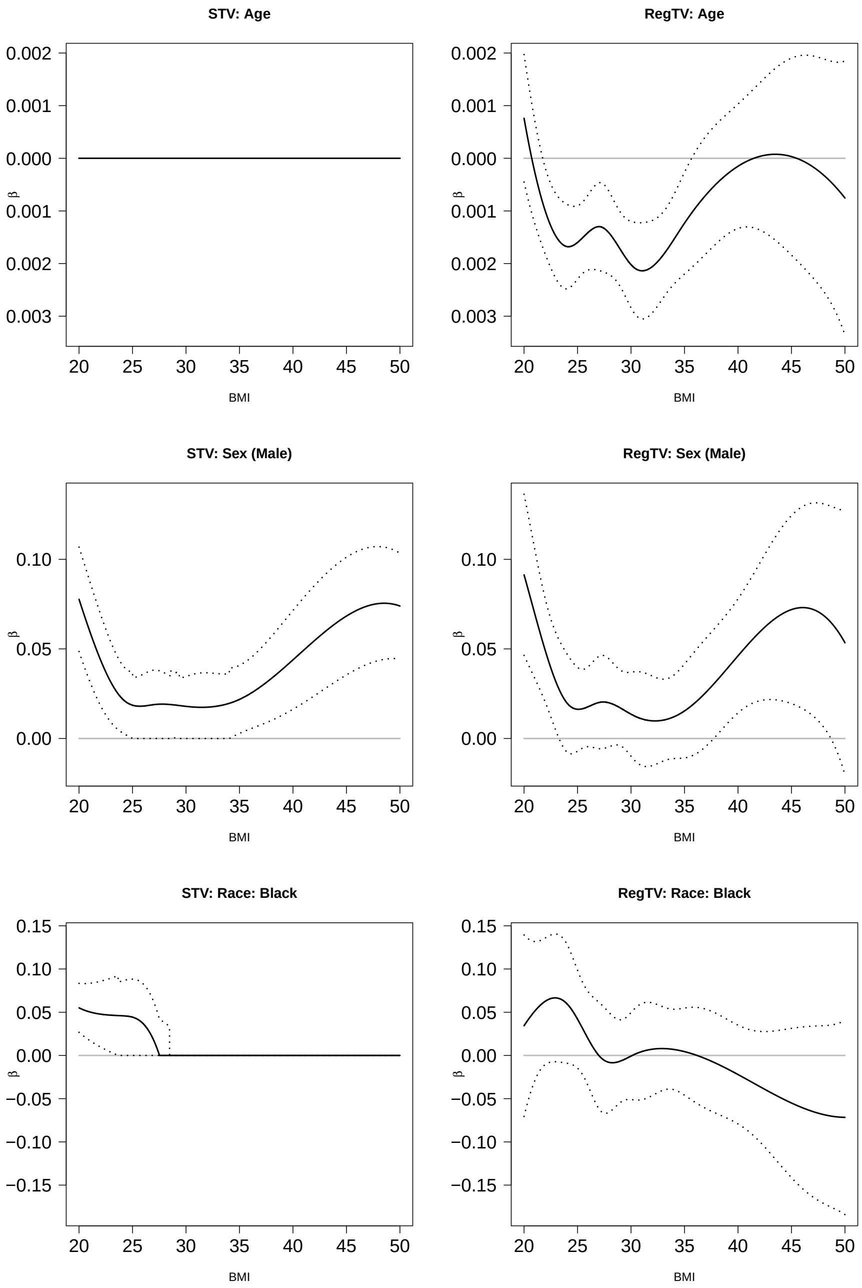

Figure 5: Estimation results (V) for the preoperative opioid use data using the B-spline method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

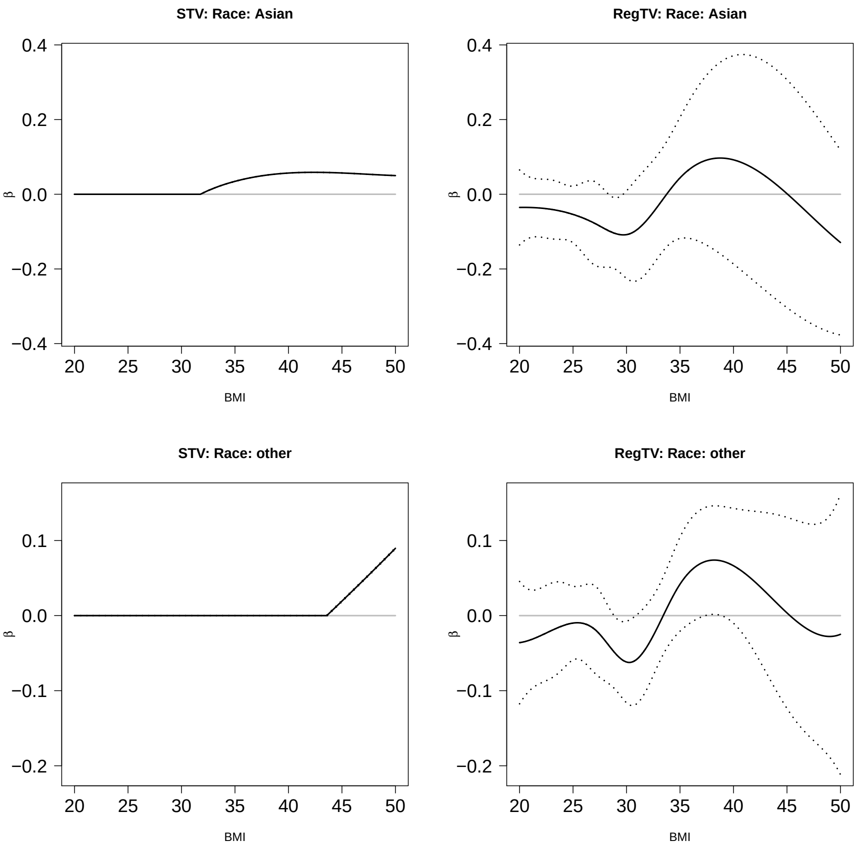

Figure 6: Estimation results (VI) for the preoperative opioid use data using the B-spline method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

Table 1: Comparisons of true positive ratios and true negative ratios among three methods for non-zero-effect region detection

n	Method	ITPR(β_1)	ITPR(β_2)	ITPR(β_3)	ITNR(β_1)	ITNR(β_2)	ITNR(β_3)
200	STV	936 (44)	919 (54)	816 (83)	987 (44)	967 (104)	976 (76)
	B-spline	977 (30)	930 (49)	833 (71)	928 (105)	952 (118)	969 (100)
	local polynomial	992 (23)	974 (38)	891 (78)	854 (141)	870 (161)	930 (127)
500	STV	962 (26)	949 (37)	883 (62)	990 (37)	980 (75)	985 (53)
	B-spline	993 (17)	970 (35)	897 (57)	876 (124)	954 (95)	967 (103)
	local polynomial	996 (12)	984 (24)	933 (54)	858 (112)	863 (123)	926 (133)
1000	STV	974 (18)	963 (25)	911 (48)	992 (24)	985 (45)	981 (69)
	B-spline	997 (9)	991 (15)	929 (43)	772 (152)	907 (129)	961 (90)
	local polynomial	996 (11)	989 (19)	951 (45)	857 (122)	836 (139)	921 (102)

ITPR: the inference-based true positive ratio; ITNR: the inference-based true negative ratio. Values are generated from 200 replications and multiplied by 10^3.

Table 2: Comparisons of true positive ratios and true negative ratios between the estimation-based method and the inference-based method using the soft-thresholded varying coefficient model for non-zero-effect region detection

		200	500	1000	2000	5000	10000
	ETPR	997 (7)	998 (5)	997 (7)	997 (7)	999 (4)	1000 (2)
β_1	ITPR	977 (14)	980 (12)	977 (14)	977 (14)	985 (10)	989 (9)
	ETNR	853 (125)	880 (104)	853 (125)	853 (125)	892 (100)	915 (84)
	ITNR	992 (30)	996 (18)	992 (30)	992 (30)	992 (27)	992 (28)
	ETPR	989 (15)	989 (16)	989 (15)	989 (15)	992 (11)	993 (10)
β_2	ITPR	962 (20)	963 (23)	962 (20)	962 (21)	972 (14)	975 (11)
	ETNR	900 (149)	872 (157)	900 (149)	900 (149)	955 (91)	981 (57)
	ITNR	991 (41)	990 (29)	991 (41)	991 (41)	994 (29)	999 (12)
	ETPR	981 (30)	978 (33)	981 (30)	981 (30)	989 (20)	991 (16)
β_3	ITPR	933 (42)	920 (40)	933 (42)	933 (42)	958 (31)	970 (24)
	ETNR	713 (282)	694 (267)	713 (282)	713 (282)	777 (266)	829 (265)
	ITNR	984 (51)	980 (64)	984 (51)	984 (51)	980 (55)	980 (60)

ETPR: the estimation-based true positive ratio; ITPR: the inference-based true positive ratio; ETNR: the estimation-based true negative ratio; ETPR: the inference-based true negative ratio. Values are multiplied by 10^3.

Table 3: Simulation for Misspecified Model

		Regular B-spline Model				STV			
n	cov(X)	ISE1	ISE2	ISE3	AISE	ISE1	ISE2	ISE3	AISE
	InD	55 (37)	41 (26)	41 (30)	46 (22)	31 (24)	35 (22)	31 (24)	32 (16)
200	CS	66 (44)	51 (37)	52 (38)	56 (30)	36 (28)	40 (26)	36 (28)	37 (20)
	AR1	65 (42)	64 (45)	50 (34)	59 (30)	37 (29)	51 (34)	38 (29)	42 (23)
	InD	23 (14)	18 (8)	17 (10)	19 (7)	11 (8)	16 (7)	11 (7)	13 (5)
500	CS	27 (17)	22 (11)	22 (12)	23 (9)	14 (10)	18 (8)	14 (9)	15 (7)
	AR1	27 (18)	27 (14)	21 (12)	25 (11)	14 (10)	22 (11)	14 (9)	17 (8)
	InD	12 (8)	10 (5)	9 (5)	11 (4)	6 (4)	10 (4)	6 (4)	7 (2)
1000	CS	12 (6)	11 (5)	12 (6)	12 (4)	7 (5)	10 (5)	7 (5)	8 (3)
	AR1	13 (6)	15 (7)	12 (6)	13 (5)	7 (5)	12 (6)	8 (5)	9 (4)

ISE: the integrated squared errors; AISE: the average integrated squared errors. Values are means and standard deviations from 200 replications and multiplied by 10^3.

Table 4: Patient Characteristics by Preoperative Opioid Use

Characteristics		No Preoperative Opioid Use ($n = 21,005$)	Preoperative Opioid Use ($n = 6,362$)
Age		52.72 (16.45)	52.74 (15.00)
BMI		29.69 (7.00)	30.77 (7.79)
Pain severity		2.53 (2.56)	5.39 (2.62)
Fibromyalgia survey score		4.61 (4.02)	8.32 (5.24)
Life satisfaction		7.34 (2.46)	6.03 (2.62)
Charlson comorbidity index		1.74 (3.31)	1.64 (3.30)
Male		9,804(46.7%)	2,876 (45.2%)
Depression		3,138 (14.9%)	2,223 (34.9%)
Race	White	19,418 (92.4%)	5,745 (90.3%)
	Black	381 (0.6%)	315 (6.0%)
	Asian	315 (1.5%)	30 (0.5%)
	Other	891 (4.2%)	272 (4.3%)
Anxiety		3,746 (34.7%)	1,523 (51.1%)
Alcohol		9,754 (46.4%)	2,611 (41.0%)
Apnea		4,720 (22.5%)	1,843 (29.0%)
Illicit drug use		674 (3.2%)	478 (7.5%)
Tobacco use		8,093 (38.5%)	3,435 (54.0%)
ASA score	< 3	7,225 (66.9%)	1,535 (51.5%)
	$\geqslant 3$	6,821 (32.5%)	2,963 (46.6%)

Continuous variables are presented in mean (standard deviation), and categorical variables in count (percentage).