CONCORDANCE MEASURE-BASED FEATURE SCREENING AND VARIABLE SELECTION

Yunbei MA, Yi Li, Huazhen Lin

Center of Statistical Research, School of Statistics

Southwestern University of Finance and Economics, Chengdu, China

and Yi Li

Department of Biostatistics, University of Michigan, USA

Supplementary Material

The supplementary materials consist of: (i) some details of iterative screening-SCAD procedure; (ii) further simulation studies; (iii) some technical lemmas used in the proofs of Theorem 1 and 2; (iv) the proofs of Theorems 1 and 2; (v) the conditions and the proof for the oracle property.

Lin, Ma and Li's research were partially supported by National Natural Science Foundation of China (No.11571282, 11528102 and 11301424) and Fundamental Research Funds for the Central Universities (No. JBK141111, JBK141121, JBK120509 and 14TD0046) of China. Li was also partially supported by the U.S. National Institutes of Health (No. RO1CA050597). We thank our editorial assistant, Ms. Martina Fu from Stanford University, for proofreading the manuscript and for many useful suggestions. We are also very grateful to the helpful comments of the Co-Editor, Associate Editor and referees that substantially improved the presentation of the paper.

S1 Iterative algorithm

The detailed algorithms of iterative screening-SCAD procedure are present as follows.

S1.1 Conditional random permutation C-SS (PC-SS)

Following the methods used by Fan, Feng and Song (2011) and Zhao and Li (2010), randomly permute $\mathbf{Y}$ to get $\mathbf{Y}_{\boldsymbol{\pi}} = (Y_{\pi_1}, \cdots, Y_{\pi_n})^T$ and compute $\widehat{g}_k^{\boldsymbol{\pi}}$, where $\boldsymbol{\pi}$ is a permutation of $\{1, \cdots, n\}$, based on the randomly coupled data $\{(Y_{\pi_i}, \mathbf{X}_i)\}_{i=1}^n$ that present no relationships between covariates and response. These estimates serve as the baseline of the marginal utilities under the null model (no relationship). To control the false selection rate under the null model, choose the screening threshold as the qth-ranked magnitude of $\{\widehat{g}_k^{\boldsymbol{\pi}},\ k = 1, \cdots, p\}$. In practice, $q = 1$, the largest marginal utility under the null model, is frequently used.

When the correlations among covariates are large, it is difficult to differentiate between the marginal utilities of the true variables, and the false ones. Enlightened by Fan, Ma and Dai (2014), we propose a PC-SS method, which performs conditional random permutation in the screening steps to determine the threshold.

0. *Determining* $\mathcal{M}^0$. For $k = 1, \cdots, p$, compute $\widehat{g}_k(0_p)$ as in (2.3) using $\{(Y_i, \mathbf{X}_i), i = 1, \cdots, n\}$. Select the top K variables by ranking $\widehat{g}_k(0_p)$, resulting in the index subset $\mathcal{M}^0$ to condition upon. Without loss of generality, we assume $\mathcal{M}^0 = \{1, \cdots, K\}$. Next estimate $\boldsymbol{\beta}_{\mathcal{M}^0} = (\beta_1, \cdots, \beta_K)^T$ by maximizing the smoothed

C-statistic (7) using $\{(Y_i, \mathbf{X}_{i\mathcal{M}^0}), i = 1, \cdots, n\}$. Here, $\mathbf{X}_{i\mathcal{M}^0} = (X_{i1}, \cdots, X_{iK})^T$. We write this estimator as $\widehat{\boldsymbol{\beta}}_{\mathcal{M}^0}$.

1. *Large-scale feature screening.* For all $k \notin \mathcal{M}^0$, compute $\widehat{g}_k((\widehat{\boldsymbol{\beta}}^T_{\mathcal{M}^0}, 0^T_{(\mathcal{M}^0)^c})^T)$ using $\{(Y_i, \mathbf{X}_i), i = 1, \cdots, n\}$. To determine the threshold for screening, we apply random permutation on the remaining covariates, i.e., $\mathbf{X}^c_{i\mathcal{M}^0} = (X_{i,K+1}, \cdots, X_{ip})^T$. Randomly permute $\{\mathbf{X}^c_{1\mathcal{M}^0}, \cdots, \mathbf{X}^c_{n\mathcal{M}^0}\}$ to get $\{\mathbf{X}^c_{\boldsymbol{\pi}_1\mathcal{M}^0}, \cdots, \mathbf{X}^c_{\boldsymbol{\pi}_n\mathcal{M}^0}\}$, where $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_n\}$ is a permutation of $\{1, \cdots, n\}$. Next for $k \notin \mathcal{M}^0$, compute $\widehat{g}^{\boldsymbol{\pi}}_k((\widehat{\boldsymbol{\beta}}^T_{\mathcal{M}^0}, 0^T_{(\mathcal{M}^0)^c})^T)$ based on the randomly coupled data $\{(Y_i, \mathbf{X}_{i\mathcal{M}^0}, \mathbf{X}^c_{i\mathcal{M}^0}), i = 1, \cdots, n\}$. Let γ^*_q be the qth-ranked magnitude of $\{|\widehat{g}^{\boldsymbol{\pi}}_k((\widehat{\boldsymbol{\beta}}^T_{\mathcal{M}^0}, 0^T_{(\mathcal{M}^0)^c})^T)|, k \notin \mathcal{M}^0\}$. Then, the active variable set is chosen as $\mathcal{A}^1 = \{k : |\widehat{g}^{\boldsymbol{\pi}}_k((\widehat{\boldsymbol{\beta}}^T_{\mathcal{M}^0}, 0^T_{(\mathcal{M}^0)^c})^T)| \geq \gamma^*_q, k \notin \mathcal{M}^0\} \cup \mathcal{M}^0$.

2. *Moderate-scale feature selection.* Apply the SCAD-penalized C-statistic (**??**) on $\mathcal{A}^1$ to select a subset of variables $\mathcal{M}^1$. Details about the implementation of SCAD are described in Section 4.3.

3. *Repeating.* Repeat steps 1 and 2, where we replace $\mathcal{M}^0$ in step 1 by $\mathcal{M}_l$, $l = 1, 2, \cdots$, and obtain $\mathcal{A}^{l+1}$ and $\mathcal{M}^{l+1}$ in step 2. Iterate until $\mathcal{M}^{l+1} = \mathcal{M}^k$ for some $k \leq l$ or $|\mathcal{M}^{l+1}| \geq \zeta_n$, for some prescribed positive integer ζ_n (such as $[n/\log(n)]$).

In practice we may have a priori knowledge that certain relevant features should be included, and could start with $\mathcal{M}^0$ containing these features in step 0. On the other hand, we could also set $\mathcal{M}^0 = \emptyset$ by taking $K = 1$. The associated algorithm is termed Greedy C-SS, which is detailed below.

S1.2 Greedy C-SS (GC-SS)

To further expedite computation, we implement a greedy version of the iterative screening-SCAD procedure. We skip step 0 and begin with step 1 in the algorithm above (i.e., take $\mathcal{M}^0 = \emptyset$), and select the top p_0 variables that have the largest norms of $\widehat{g}_k(0_p)$. In our simulation studies, p_0 is set as 1.

S2 More simulation studies of variable selection

Non-linear regression model. $Y = \exp\{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5\} + e$, where the noise e was generated from normal distribution $N(0, \sigma^2)$ with $\sigma = 0.5$. This example explores what happens when the model structure is nonlinear. Four methods were compared, which included conditional permutation screening-SCAD methods based on smoothed C-statistic (PC-SS) as in Section 1.1 with $K = 0$ (i.e., take $\mathcal{M}^0 = \emptyset$), Greedy screening-SCAD methods based on smoothed C-statistic (GC-SS) as in Section 1.2 with $p_0 = 1$, Permutation-SIS-SCAD (PSIS) as in Fan and Lv (2008), and Vanilla-SIS-SCAD (VSIS) as in Fan, Samworth and Wu (2009). Results are presented in Table 1. The results show that for nonlinear regression models, among the first three methods, the proposed methods had acceptable probabilities of

Table 1: Simulation results for non-linear regression model

	med.$\|\widehat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\| = 0.064$							
	$t = 0$				$t = 1$			
	PC-SS	GC-SS	PSIS	VSIS	PC-SS	GC-SS	PSIS	VSIS
perc.incl.true	0.98	0.94	0.85	0.99	0.97	0.94	0.87	0.99
med.model size	5	5	6	37	5	5	6	37
aver. model size	5.16	4.84	5.90	30.77	4.99	4.82	6.02	31.64
med.$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	0.068	0.068	0.751	2.752	0.066	0.067	0.735	2.688

including the true model, a smaller model size, and much smaller prediction error. The VSIS method however failed as it enhanced probabilities of including the true model at the expense of high false-positive rates .

Poisson regression model. In this example, we generated the response Y from a Poisson distribution $P(\lambda)$ with $\lambda = \exp\{2 \times (\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)\}$. The ordinal response was generated from the Poisson distribution; the PSIS-G and the VSIS-G are carried out under the Poisson distribution.

Table 2 indicates that the proposed methods showed similar performance as Fan, Samworth and Wu (2009)'s methods on model selection. This is not surprising, as their methods were carried under the correctly specified model. However, our methods had much smaller estimation errors than the competing methods.

Ordinal regression model. We evaluated the robustness of the proposed meth-

Table 2: Simulation results for Poisson regression model

	med.$\|\widehat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\| = 0.063$							
	$t = 0$				$t = 1$			
	PC-SS	GC-SS	PSIS-G	VSIS-G	PC-SS	GC-SS	PSIS-G	VSIS-G
perc.incl.true	0.96	0.94	0.98	0.99	0.97	0.92	0.98	1.00
med. model size	5	5	5	5	5	5	5	37
aver. model size	4.99	4.82	5.07	5.04	5.02	4.71	5.07	5.11
med.$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	0.063	0.067	1.001	0.999	0.066	0.065	1.006	1.002

ods via a comparison with Fan, Samworth and Wu (2009)'s methods when the distribution was misspecified. We generated the ordinal response Y as follows. Let $Y^* = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$, where the noise e was generated from $0.2 \cdot t(1)$. Define $Y = I(Y^* > -1.2) + I(Y^* > -0.6) + I(Y^* > -0.15) + I(Y^* > 0.3) + I(Y^* > 0.8) + I(Y^* > 1.5)$. We compared our proposed methods PC-SS and GC-SS with Permutation-SIS-SCAD (PSIS-G) and Vanilla-SIS-SCAD (VSIS-G; Fan, Samworth and Wu, 2009) for Poisson regression models. Table 3 shows that the proposed methods had an acceptable probability of including the true model, a smaller model size, and a much smaller prediction error as compared to the PSIS-G and VSIS-G. The PSIS-G and VSIS-G methods had low true-positive rates and large prediction errors.

Table 3: Simulation results for ordinal regression model

	med.$\|\widehat{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\| = 0.069$							
	$t = 0$				$t = 1$			
	PC-SS	GC-SS	PSIS	VSIS	PC-SS	GC-SS	PSIS	VSIS
perc.incl.true	0.95	0.90	0.83	0.84	0.97	0.88	0.84	0.86
med.model size	5	5	5	5	5	5	5	5
aver. model size	5.05	4.64	5.52	5.29	5.10	4.50	5.61	5.49
med.$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	0.074	0.071	0.522	0.655	0.075	0.071	0.500	0.686

S3 Some technical lemmas

Some technical Lemmas needed for our main results are stated and proved below. Lemmas 1 and 2 characterize the exponential tails, which are useful for the main proof. Lemma 3 is a Bernstein type inequality. Lemmas 4 and 5 provide a large deviation theory and a Bernstein type inequality for U-statistic, respectively.

Lemma 1 *Let W be a random variable. Suppose W has a conditional exponential tail: $P(|W| > t) \leq \exp(1 - (t/K)^r)$ for all $t \geq 0$, where $K > 0$ and $r \geq 1$. Then for all $m \geq 2$,*

$$\mathrm{E}(|W|^m) \leq eK^m m!. \tag{S3.1}$$

Proof. Recall that for any non-negative random variable W, $\mathrm{E}[W] = \int_0^\infty \mathrm{P}\{W \geq t\}dt$. Then we have

$$\begin{aligned}\mathrm{E}(|W|^m) &= \int_0^\infty \mathrm{P}\{|W|^m \geq t\}dt \\ &\leq \int_0^\infty \exp(1-(t^{1/m}/K)^r)dt \\ &= \frac{emK^m}{r}\Gamma(\frac{m}{r}).\end{aligned}$$

The lemma follows from the fact $r \geq 1$.

Lemma 2 *Let W_1, W_2 be independent random variables, satisfying $P(|W_i| > t) \leq \exp(1-(t/K)^r), \quad i = 1, 2$ for all $t \geq 0$, where $K > 0$ and $r \geq 1$. for all $t \geq 0$. Then for all $m \geq 2$,*

$$E(|W_1 + W_2|^m) \leq 2e(2K)^m m!. \tag{S3.2}$$

Proof. For any $t > 0$, we have

$$\begin{aligned}\mathrm{P}(|W_1 + W_2| > t) &\leq \mathrm{P}(|W_1| > t/2) + \mathrm{P}(|W_2| > t/2) \\ &\leq 2\exp(1-(t/2K)^r).\end{aligned}$$

Hence, for all $m \geq 2$,

$$\begin{aligned}\mathrm{E}(|W_1 + W_2|^m) &= \int_0^\infty \mathrm{P}\{|W_1 + W_2|^m \geq t\}dt \\ &\leq 2\int_0^\infty \exp(1-(t^{1/m}/2K)^r)dt \\ &= \frac{2em(2K)^m}{r}\Gamma(\frac{m}{r}) \\ &\leq 2e(2K)^m m!.\end{aligned}$$

Lemma 2 holds.

Lemma 3 *(Bernstein inequality, Lemma 2.2.11, van der Vaart and Wellner (1996)). For independent random variables* $Y_1, \cdots, Y_n$ *with mean zero such that* $\mathrm{E}[|Y_i|^m] \leq m! M^{m-2}\nu_i/2$ *for every* $m \geq 2$ *(and all* i*) and some constants* M *and* ν_i*. Then*

$$P(|Y_1 + \cdots + Y_n| > x) \leq 2\exp\{-x^2/(2(\nu + Mx))\},$$

for $v \geq \nu_1 + \cdots + \nu_n$*.*

Suppose $h(\cdot,\cdot)$ is a binary kernel of the U-statistic $U_n = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} h(W_i, W_j)$, where $W_1, W_2, \cdots, W_n$ are i.i.d. random variables or random vectors. Let $d_n = [\frac{n}{2}]$, the greatest integer $\leq n/2$. For any permutation $(i_1, \cdots, i_n)$ of $(1, \cdots, n)$, define

$$\Upsilon(W_{i_1}, \cdots, W_{i_n}) = \frac{1}{d}[h(W_{i_1}, W_{i_2}) + h(W_{i_3}, W_{i_4}) + \cdots + h(W_{i_{2d_n-1}}, W_{i_{2d_n}})].$$

Then we can rewrite U_n as

$$U_n = \frac{1}{n!} \sum_{i_1 \neq i_2 \neq \cdots \neq i_n}^{n} \Upsilon(W_{i_1}, \cdots, W_{i_n}). \tag{S3.3}$$

Note that $\Upsilon(\cdot)$ is the average of d_n i.i.d random variables. This type of representation was introduced and utilized by (Hoeffding, 1963).

Lemma 4 *If* $E[h(W_1, W_2)] = \mu$*, and* $E[\exp\{th(W_1, W_2)\}] < \infty$ *for any* $0 < t \leq t_0$*, then*

$$P(U_n - \mu > \delta) \leq \exp\{-\sup_{0<t\leq t_0}[td_n\delta - d_n \ln Q(t)]\}.$$

Here $Q(t) = E[\exp\{t(h(W_1, W_2) - \mu)\}]$*.*

Proof. Note that for any random variable W satisfying $E[\exp\{tW\}] < \infty$, for $0 < t \leq t_0$, it follows from the Markov's inequality that

$$P(W - E[W] > \delta) \leq \exp\{-t\delta\} E[\exp\{t(W - E[W])\}].$$

Since the exponential function is convex, it follows by Jensen's inequality that for $0 < t \leq t_0$,

$$\begin{aligned}
E[\exp\{td_n U_n\}] &= E\left[\exp\{\frac{td_n}{n!}\sum_{i_1\neq i_2\neq\cdots\neq i_n}^{n}\Upsilon(W_{i_1},\cdots,W_{i_n})\}\right] \\
&\leq \frac{1}{n!}\sum_{i_1\neq i_2\neq\cdots\neq i_n}^{n} E\left[\exp\{td_n\Upsilon(W_{i_1},\cdots,W_{i_n})\}\right] \\
&= \frac{1}{n!}\sum_{i_1\neq i_2\neq\cdots\neq i_n}^{n} E\left[\prod_{k=1}^{d_n}\exp\{th(W_{i_{2k-1}},W_{i_{2k}})\}\right] \\
&= \frac{1}{n!}\sum_{i_1\neq i_2\neq\cdots\neq i_n}^{n} \left[\prod_{k=1}^{d_n} E\exp\{th(W_{i_{2k-1}},W_{i_{2k}})\}\right] \\
&= \frac{1}{n!}\sum_{i_1\neq i_2\neq\cdots\neq i_n}^{n} \exp\left[\sum_{k=1}^{d_n}\ln E\exp\{th(W_{i_{2k-1}},W_{i_{2k}})\}\right] \\
&= \exp\left[d_n \ln E\exp\{th(W_1,W_2)\}\right] < \infty.
\end{aligned}$$

We can then obtain that

$$\begin{aligned}
P(U_n - \mu > \delta) &\leq \exp\{-td_n\delta\}E[\exp\{td_n(U_n-\mu)\}] && \text{(S3.4)} \\
&\leq \exp\{-td_n\delta\}\exp\left[d_n\ln E\exp\{th(W_1,W_2)\} - td_n\mu\right] \\
&= \exp\{-[td_n\delta - d_n\ln Q(t)]\}. && \text{(S3.5)}
\end{aligned}$$

Since (S3.5) is true for all $0 < t \leq t_0$,

$$\begin{aligned} P(U_n - \mu > \delta) &\leq \inf_{0<t\leq t_0} \exp\{-[td_n\delta - d_n \ln Q(t)]\} \\ &= \exp\{-\sup_{0<t\leq t_0} [td_n\delta - d_n \ln Q(t)]\}, \end{aligned}$$

and Lemma 4 holds.

Lemma 5 *If $E[h(W_1, W_2)] = \mu$, and for some $A > 0$ and any $m \geq 2$, $E|h(W_1, W_2) - \mu|^m \leq m!A^{m-2}\nu/2$, then for any $\delta > 0$,*

$$P(|U_n - \mu| > \delta) \leq 2\exp\{-\frac{d_n\delta^2}{2(\nu + A\delta)}\}.$$

Proof. With Taylor's expansion of $\exp\{t(h(W_1, W_2) - \mu)/d_n\}$ at 0, we have

$$\begin{aligned} E[\exp\{t(h(W_1, W_2) - \mu)\}] &= 1 + \frac{t^2}{2}E[h(W_1, W_2) - \mu]^2 + \sum_{m=3}^{\infty} \frac{t^m}{m!} E[h(W_1, W_2) - \mu]^m \\ &\leq 1 + \frac{t^2\nu}{2} + \frac{t^3 A\nu}{2} \sum_{m=0}^{\infty} (tA)^m. \end{aligned}$$

Furthermore, if $0 < t < \frac{1}{A}$,

$$E[\exp\{t(h(W_1, W_2) - \mu)\}] \leq 1 + \frac{t^2\nu}{2} + \frac{t^3 A\nu}{2}\frac{1}{1 - tA} = 1 + \frac{t^2\nu}{2(1 - tA)} < \infty.$$

Hence, by Lemma 4, we have for any $\delta > 0$ and any small $\varepsilon > 0$,

$$P(U_n - \mu > \delta) \leq \exp\{-\sup_{0<t\leq(\frac{1}{A}-\varepsilon)} [td_n\delta - d_n \ln Q(t)]\}. \tag{S3.6}$$

Note that $\ln x \leq x - 1$ for any $x \geq 0$, then for $0 < t < \frac{d_n}{A}$

$$\ln Q(t) \leq Q(t) - 1 = E[\exp\{t(h(W_1, W_2) - \mu)\}] - 1$$

$$\begin{aligned}
&\leq \frac{t^2\nu}{2} + \frac{t^3 A\nu}{2}\sum_{m=0}^{\infty}(tA)^m \\
&\leq \frac{t^2\nu}{2} + \frac{t^3 A\nu}{2}\frac{1}{1-tA} \\
&= \frac{t^2\nu}{2(1-tA)}, \qquad \text{(S3.7)}
\end{aligned}$$

whence

$$\sup_{0<t\leq\frac{1}{A}-\varepsilon}[td_n\delta - d_n\ln Q(t)] \geq \sup_{0<t\leq\frac{1}{A}-\varepsilon} d_n\left[t\delta - \frac{t^2\nu}{2(1-tA)}\right]. \qquad \text{(S3.8)}$$

By elementary calculus, we obtain the value of t that maximizes the expression in brackets (out of the two roots of the second degree polynomial equation, we choose the one which is $< \frac{1}{A}$) as, $t_{opt} = \frac{1}{A}(1 - \frac{1}{\sqrt{1+2\delta A/\nu}})$. Observing that $\sqrt{1+x} \leq 1 + x/2$, one gets

$$t_{opt} \leq \frac{1}{A}(1 - \frac{1}{1+\delta A/\nu}) = \frac{\delta}{\nu+\delta A} \equiv t' < \frac{1}{A}.$$

It then follows from (S3.8) that

$$\sup_{0<t\leq\frac{1}{A}-\varepsilon}[td_n\delta - d_n\ln Q(t)] \geq t'd_n\delta - \frac{t'^2 d_n\nu}{2(1-t'A)} = \frac{d_n\delta^2}{2(\nu+\delta A)}.$$

Combing with (S3.6) yields that

$$P(U_n - \mu > \delta) \;\; \leq \;\; \exp\{-\frac{d_n\delta^2}{2(\nu+\delta A)}\}. \qquad \text{(S3.9)}$$

By letting $U_n^* = -U_n = \frac{1}{n(n-1)}\sum_{i\neq j}^{n}[-h(W_i, W_j)]$ and $\mu^* = -\mu$, equivalently we have

$$P(U_n - \mu < -\delta) = P(U_n^* - \mu^* > \delta) \;\; \leq \;\; \exp\{-\frac{d_n\delta^2}{2(\nu+\delta A)}\}. \qquad \text{(S3.10)}$$

Thus

$$P(|U_n - \mu| > \delta) = P(U_n - \mu > \delta) + P(U_n - \mu < -\delta) \leq 2\exp\{-\frac{d_n\delta^2}{2(\nu + \delta A)}\}.$$

Lemma 5 holds.

S4 Proof of sure screening properties

Proof of Theorem 1

We first prove part (1). Recall that for $k = 1, \cdots, p_n$, $g_k(0) = E\widehat{g}_k(0) = E[I\{Y_1 > Y_2\}(X_{1k} - X_{2k})]$. Thus for any $k = 1, \cdots, p_n$

$$\begin{aligned}
& |\widehat{g}_k(0) - E\widehat{g}_k(0)| \\
= & \frac{1}{n(n-1)}\left|\sum_{i\neq j}^{n} I(Y_i > Y_j)(X_{ik} - X_{jk}) - E[I\{Y_1 > Y_2\}(X_{1k} - X_{2k})]\right| \\
= & \frac{1}{n(n-1)}\left|\sum_{i\neq j}^{n} I(Y_i > Y_j)(X_{ik} - X_{jk}) - E[G(Z_1, Z_2)(X_{1k} - X_{2k})]\right| \\
= & \frac{1}{n(n-1)}\Big|\sum_{i\neq j}^{n}\{I(Y_i > Y_j) - G(Z_i, Z_j)\}(X_{ik} - X_{jk})] \\
& + \sum_{i\neq j}^{n}\{G(Z_i, Z_j)(X_{ik} - X_{jk}) - E[G(Z_1, Z_2)(X_{1k} - X_{2k})]\}\Big| \\
\leq & T_{n1} + T_{n2}, \qquad \text{(S4.1)}
\end{aligned}$$

where

$$T_{n1} = \frac{1}{n(n-1)}\left|\sum_{i\neq j}^{n}\{I(Y_i > Y_j) - G(Z_1, Z_2)\}(X_{ik} - X_{jk})]\right|,$$

and

$$T_{n2} = \frac{1}{n(n-1)} \left| G(Z_1, Z_2)(X_{ik} - X_{jk}) - E[G(Z_1, Z_2)(X_{1k} - X_{2k})] \right|.$$

We first focus on T_{n1}. For all $m \geq 2$,

$$\begin{aligned} & E\left\{ \left| \left[I(Y_i > Y_j) - G(Z_i, Z_j) \right] (X_{ik} - X_{jk}) \right|^m \right\} \\ \leq \; & 2^m E[|X_{ik} - X_{jk}|^m] \\ \leq \; & 2e(4K_1)^m m! = m!(4K_1)^{m-2}(64eK_1^2/2), \end{aligned} \tag{S4.2}$$

where the last inequality is obtained based on Condition (C.1) and Lemma 2. Since $E\{[I(Y_i > Y_j) - G(Z_i, Z_j)] (X_{ik} - X_{jk})\} = 0$, it follows from Lemma 5 and equation (S4.2) that for any $\delta > 0$,

$$P\left(T_{n1} > \frac{\delta}{n} \right) \leq 2 \exp\{ -\frac{d_n \delta^2}{2n(64neK_1^2 + 4K_1\delta)} \}.$$

We now focus on T_{n2}. According to the Minkowski inequality, for any $m \geq 2$,

$$\begin{aligned} & E\left\{ \left| G(Z_i, Z_j)(X_{ik} - X_{jk}) - E[G(Z_1, Z_2)(X_{1k} - X_{2k})] \right|^m \right\} \\ \leq \; & 2^m E\left\{ \left| G(Z_i, Z_j)(X_{ik} - X_{jk}) \right|^m \right\} \\ \leq \; & 2^m E[|X_{ik} - X_{jk}|^m] \\ \leq \; & 2e(4K_1)^m m! = m!(4K_1)^{m-2}(64eK_1^2/2). \end{aligned}$$

Hence, T_{n2} has the same results as T_{n1}, that is for any $\delta > 0$,

$$P\left(T_{n2} > \frac{\delta}{n} \right) \leq 2 \exp\{ -\frac{d_n \delta^2}{2n(64neK_1^2 + 4K_1\delta)} \}.$$

Consequently, the union bound of probability yields that

$$P(|\widehat{g}_k(0) - E\widehat{g}_k(0)| > \frac{2\delta}{n}) \leq 4\exp\{-\frac{d_n\delta^2}{2n(64neK_1^2 + 4K_1\delta)}\}. \tag{S4.3}$$

Note that $d_n = [n/2] \geq (n-1)/2$, then by letting $c_2 = 256eK_1^2$ and $c_3 = 16K_1$, we obtain that for any $k = 1, \cdots, p_n$,

$$P(|\widehat{g}_k(0) - E\widehat{g}_k(0)| > \frac{2\delta}{n}) \leq 4\exp\{-\frac{\delta^2}{nc_2 + c_3\delta}\}. \tag{S4.4}$$

Thus, we have for any constant c_1,

$$P(|\widehat{g}_k(0) - E\widehat{g}_k(0)| > c_1 n^{-\kappa}) \leq 4\exp\{-\frac{c_1^2 n^{1-2\kappa}}{2(2c_2 + c_1c_3n^{-\kappa})}\}. \tag{S4.5}$$

Hence part (1) follows from the fact that

$$P(\max_{1\leq k\leq p_n} |\widehat{g}_k(0) - E\widehat{g}_k(0)| > c_1n^{-\kappa}) \leq \sum_{k=1}^{p_n} P(|\widehat{g}_k(0) - E\widehat{g}_k(0)| > c_1n^{-\kappa}).$$

We now prove part (2). Note that $E\widehat{g}_k(0) = E[I\{Y_1 > Y_2\}(X_{1k} - X_{2k})]$, and

$$|\widehat{g}_k(0) - E\widehat{g}_k(0)| \geq |E\widehat{g}_k(0)| - \widehat{g}_k(0),$$

then for $k \in \mathcal{M}_0$ on the event

$$\mathcal{A}_{nk} = \{|\widehat{g}_k(0) - E\widehat{g}_k(0)| < \delta n^{-\kappa}/2\},$$

we have

$$|\widehat{g}_k(0)| > |E\widehat{g}_k(0)| - \delta n^{-\kappa}/2 \geq \delta n^{-\kappa}/2.$$

Thus

$$P(|\widehat{g}_k(0)| \leq \delta n^{-\kappa}/2) \leq P(\mathcal{A}_{nk}^c) \leq 4\exp\{-\frac{\delta^2 n^{1-2\kappa}}{4(4c_2 + \delta c_3 n^{-\kappa})}\},$$

and

$$P\left(\mathcal{M}_0 \subset \widehat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - \sum_{k \in \mathcal{M}_0} P(\mathcal{A}_{nk}^c) \geq 1 - 4s_0 \exp\{-\frac{\delta^2 n^{1-2\kappa}}{4(4c_2 + \delta c_3 n^{-\kappa})}\}.$$

Proof of Theorem 2:

Note that

$$\sum_{i=1}^{p} |g_k(0_p)| = 2\sum_{i=1}^{p} |E[I\{Y_1 > Y_2\}(X_{1k} - X_{2k})|,$$

which implies that for any $\delta > 0$, the number of $\{k : |g_k(0_p)| > \delta n^{-\kappa}\}$ cannot exceed $O(n^\kappa \sum_{i=1}^{p} |E[I\{Y_1 > Y_2\}(X_{1k} - X_{2k})|)$. Then on the set $\mathcal{B}_n = \{\max_{1\leq k\leq p_n} |\widehat{g}_k(0) - E\widehat{g}_k(0)| \leq \delta n^{-\kappa}\}$, the number of $\{k : |\widehat{g}_k(0_p)| > 2\delta n^{-\kappa}\}$ can not exceed the number of $\{k : |g_k(0_p)| > \delta n^{-\kappa}\}$, which is bounded by $O(n^\kappa \sum_{k=1}^{p} |E[I\{Y_1 > Y_2\}(X_{1k} - X_{2k})|)$. Hence, by taking $\delta = c_4/2$, we have

$$\Pr\left\{|\widehat{\mathcal{M}}_{\gamma_n}| \leq O(n^\kappa \sum_{i=1}^{p} |E[I\{Y_1 > Y_2\}(X_{1k} - X_{2k})|)\right\} \geq \Pr(\mathcal{B}_n).$$

Then the desired result follows from Theorem 1(1).

S5 Conditions and Proof for the Oracle Property

S5.1 Regularity Conditions

Let $a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$ and $b_n = \max\{p''_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$. We first place the following conditions on the penalty functions:

(P.1) $a_n = O(n^{-1/2})$,

(P.2) $b_n \to 0$ as $n \to \infty$,

(P.3) $\liminf_{n\to\infty} \inf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$,

(P.4) there are constants D_1 and D_2 such that, when $\theta_1, \theta_2 > D_1\lambda_n$, $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \leq D_2|\theta_1 - \theta_2|$.

Condition (P.1) ensures both the unbiasedness property for large parameters and the existence. Condition (P.2) guarantees that the penalty function does not have much more influence than the smoothed AUC function on the penalized smoothed estimators. Condition (P.3) make the penalty function singular at the origin so that the penalized smoothed estimators possess the sparsity property. Condition (P.4) is a smoothness condition that is imposed on the penalty function.

The following conditions are necessary for obtaining the oracle property.

($C.1^*$) Write $Z = \tilde{\mathbf{X}}^T\tilde{\boldsymbol{\beta}}_0$. For $k = 1, \cdots, m_n$, let $\mu_k(Z) = E(X_k|Z)$ and $v_k(Z) = Var(X_k|Z)$. We assume that $\mu_k(\cdot)$ and $v_k(\cdot)$, $k = 1, \cdots, m_n$, have bounded continuous second order derivatives.

($C.2^*$) For $1 \leq k, l \leq m_n$, let $c_{kl}(Z) = Cov(X_k, X_l|Z)$. We assume that $c_{kl}(\cdot)$, $k, l = 1, \cdots, m_n$, have bounded continuous second order derivatives. Furthermore, we assume that $G(\cdot,\cdot)$ has bounded second order partial derivatives, where $G(z_1, z_2) = E[I\{Y_1 > Y_2\}|\tilde{\mathbf{X}}_1^T\tilde{\boldsymbol{\beta}}_0 = z_1, \tilde{\mathbf{X}}_2^T\tilde{\boldsymbol{\beta}}_0 = z_2]$.

$(C.3^*)$ Define $I(\tilde{\boldsymbol{\beta}}_0) = E[2Cov(\tilde{\mathbf{X}}|Z)G^{(1,0)}(Z,Z) + G(Z,Z)\frac{\partial Cov(\tilde{\mathbf{X}}|Z)}{\partial Z}]$, where $G^{(1,0)}(\cdot,\cdot)$ is the partial derivative of $G(\cdot,\cdot)$ with respect to the first variable. Assume that $I(\tilde{\boldsymbol{\beta}}_0)$ is a positive definite matrix with finite maximum eigenvalue, and the minimum eigenvalue is bounded away from 0.

$(C.4^*)$ For all $1 \le i \ne j \le n$, write $\widetilde{C}_s^{(i,j)}(\tilde{\boldsymbol{\beta}}) = \frac{1}{n(n-1)}\left[I(Y_i > Y_j)\Phi\left\{(\tilde{\mathbf{X}}_i^T\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{X}}_j^T\tilde{\boldsymbol{\beta}})/h\right\}\right]$. There is a large enough open subset ω_n of $\Omega_n \in \mathbf{R}^{m_n}$ that contains the true parameter point $\tilde{\boldsymbol{\beta}}_0$, such that for almost all $(Y_i, \tilde{\mathbf{X}}_i)$ and all $\tilde{\boldsymbol{\beta}} \in \omega_n$, $\left|\frac{\partial^3 \widetilde{C}_s^{(i,j)}(\tilde{\boldsymbol{\beta}})}{\partial\beta_k\partial\beta_l\partial\beta_t}\right| \le M_{nklt}((Y_i, \tilde{\mathbf{X}}_i))$, where $M_{nklt}((Y_i, \tilde{\mathbf{X}}_i))$, satisfying that there exists a constant M such that $E[M_{nklt}^2((Y_i, \tilde{\mathbf{X}}_i))] \le M < \infty$, for all $1 \le k,l,t \le m_n$.

$(C.5^*)$ Suppose $\beta_{10}, \cdots, \beta_{s_0 0}$ satisfy $\min\limits_{1\le k\le s_0} |\beta_{k0}|/\lambda_n \to \infty,$ as $n \to \infty$.

$(C.6^*)$, $m_n^4 = o(n)$ as $n \to \infty$.

Conditions $(C.1^*)$-$(C.4^*)$ are imposed on the second and the third derivatives of $\widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)$, $\mu_k(\cdot)$, $v_k(\cdot)$ and $c_{kl}(\cdot)$. These conditions are stronger than those for finite parameter situations, but they facilitate the technical derivations. Condition $(C.3^*)$ assumes that the information matrix of the smoothed C-statistic $\widetilde{C}_s(\tilde{\boldsymbol{\beta}})$ is positive definite, and has uniformly bounded eigenvalues. Under $(C.4^*)$, the variation of the tail for $\widetilde{C}_s(\tilde{\boldsymbol{\beta}})$ is assumed to be bounded. Similar conditions are imposed by Fan and Peng (2004) for generalized linear models. Condition $(C.5^*)$ explicitly shows the rate at which the penalized smoothed C-statistic can distinguish nonvanishing parameters from zero, is necessary for obtaining the oracle property.

S5.2 Proof of Theorem 3 and 4:

Proof of Theorem 3:

Our goal is to show that for any given $\varepsilon > 0$, there exists a constant B, large enough to make

$$\Pr\{\sup_{\|\mathbf{u}\|=1,\mathbf{u}^T\tilde{\boldsymbol{\beta}}_0=1} \mathrm{PC}_s((1-B^2\alpha_n^2)^{1/2}\tilde{\boldsymbol{\beta}}_0 + B\alpha_n\mathbf{u}) < \mathrm{PC}_s(\tilde{\boldsymbol{\beta}}_0)\} \geq 1-\varepsilon, \qquad \text{(S5.1)}$$

where $\alpha_n = \sqrt{m_n}(n^{-1/2} + a_n)$.

This implies that with a probability tending to 1, there is a local maximum $\widehat{\boldsymbol{\beta}}_n$ in the ball $\{(1-\delta^2\alpha_n^2)^{1/2}\tilde{\boldsymbol{\beta}}_0 + \delta\alpha_n\mathbf{u} : \|\mathbf{u}\| = 1, \mathbf{u}^T\tilde{\boldsymbol{\beta}}_0 = 1, \text{and } \delta < B\}$, hence satisfying $\|\widehat{\boldsymbol{\beta}}_n\| = 1$ and such that $\|\widehat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_0\| = O_p(\alpha_n)$.

Define $\boldsymbol{\beta}_n^* = (1-B^2\alpha_n^2)^{1/2}\tilde{\boldsymbol{\beta}}_0 + B\alpha_n\mathbf{u}$, using $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned} D_A(\boldsymbol{\beta}_n^*) &= PC_s(\boldsymbol{\beta}_n^*) - PC_s(\tilde{\boldsymbol{\beta}}_0) \leq \widetilde{C}_S(\boldsymbol{\beta}_n^*) - \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0) - \sum_{j=1}^{s_0}[p_{\lambda_n}(|\beta_j^*|) - p_{\lambda_n}(|\beta_{j0}|)] \\ &\equiv \mathrm{I}_{n1} + \mathrm{I}_{n2}. \end{aligned}$$

By Taylor's expansion we obtain

$$\begin{aligned} \mathrm{I}_{n1} &= \nabla^T\widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0) + \frac{1}{2}(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)^T\nabla^2\widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0) \\ &\quad + \frac{1}{6}\nabla^T\left((\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)^T\nabla^2\widetilde{C}_s(\boldsymbol{\beta}_n^{**})(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)\right) \\ &\equiv \mathrm{II}_{n1} + \mathrm{II}_{n2} + \mathrm{II}_{n3}, \end{aligned}$$

where $\boldsymbol{\beta}_n^{**}$ lies between $\boldsymbol{\beta}_n^*$ and $\tilde{\boldsymbol{\beta}}_0$.

Note that

$$\begin{aligned}
& |\mathrm{II}_{n1}| \\
= & |\nabla^T \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)| \\
= & \frac{1}{hn(n-1)} \left| \sum_{i \neq j} I(Y_i > Y_j)\phi(\frac{Z_i - Z_j}{h})(\tilde{\mathbf{X}}_i - \tilde{\mathbf{X}}_j)^T(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0) \right| \\
\leq & \left\| \frac{1}{hn(n-1)} \sum_{i \neq j} I(Y_i > Y_j)\phi(\frac{Z_i - Z_j}{h})(\tilde{\mathbf{X}}_i - \tilde{\mathbf{X}}_j) \right\| \|(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)\| \\
= & \left\{ \sum_{k=1}^{m_n} \left[[\frac{1}{hn(n-1)} \sum_{i \neq j} I(Y_i > Y_j)\phi(\frac{Z_i - Z_j}{h})(X_{ik} - X_{jk}) \right]^2 \right\}^{1/2} \|(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)\|.
\end{aligned}$$

Let $q_{nk}(\boldsymbol{\beta}_n^*) = \frac{1}{hn(n-1)} \sum\limits_{i \neq j} I(Y_i > Y_j)\phi(\frac{Z_i - Z_j}{h})(X_{ik} - X_{jk})$, for $k = 1, \cdots, n$. Then

$$q_{nk}^2(\boldsymbol{\beta}_n^*) = 2S_{n1}^k(\boldsymbol{\beta}_n^*) + 4S_{n2}^k(\boldsymbol{\beta}_n^*) + 6S_{n3}^k(\boldsymbol{\beta}_n^*),$$

where

$$\begin{aligned}
S_{n1}^k(\boldsymbol{\beta}_n^*) &= \frac{1}{h^2 n^2 (n-1)^2} \sum_{i \neq j} I(Y_i > Y_j)\phi^2(\frac{Z_i - Z_j}{h})(X_{ik} - X_{jk})^2, \\
S_{n2}^k(\boldsymbol{\beta}_n^*) &= \frac{1}{h^2 n^2 (n-1)^2} \sum_{i \neq j \neq l} I(Y_i > Y_j)\phi(\frac{Z_i - Z_j}{h})(X_{ik} - X_{jk}) \\
& \quad I\{Y_l > Y_j\}\phi(\frac{Z_l - Z_j}{h})(X_{lk} - X_{jk}), \\
S_{n3}^k(\boldsymbol{\beta}_n^*) &= \frac{1}{h^2 n^2 (n-1)^2} \sum_{i \neq j \neq l \neq t} I(Y_i > Y_j)\phi(\frac{Z_i - Z_j}{h})(X_{ik} - X_{jk}) \\
& \quad I(Y_l > Y_t)\phi(\frac{Z_l - Z_t}{h})(X_{lk} - X_{tk}).
\end{aligned}$$

By Condition $(C.1^*)$,

$$
\begin{aligned}
& E[S_{n1}^k(\boldsymbol{\beta}_n^*)] \\
= \ & \frac{1}{h^2 n(n-1)} E\Big\{ G(Z_1, Z_2)\phi^2(\frac{Z_1 - Z_2}{h})\{[\mu_k(Z_1) - \mu_k(Z_2)]^2 + v_k(Z_1) + v_k(Z_2)\} \Big\} \\
= \ & \frac{1}{hn(n-1)} E\left\{ \pi^{-1/2} G(Z_2, Z_2) v_k(Z_2) + o_P(1) \right\} \\
= \ & O(\frac{1}{hn(n-1)}),
\end{aligned}
$$

$$
\begin{aligned}
& E[S_{n2}^k(\boldsymbol{\beta}_n^*)] \\
= \ & \frac{n-2}{h^2 n(n-1)} E\Big\{ G(Z_1, Z_3) G(Z_2, Z_3) \\
& \quad \phi(\frac{Z_1 - Z_3}{h})\phi(\frac{Z_2 - Z_3}{h})\{[\mu_k(Z_1) - \mu_k(Z_3)][\mu_k(Z_2) - \mu_k(Z_3)] + v_k(Z_3)\} \Big\} \\
= \ & \frac{n-2}{n(n-1)} E[G^2(Z_3, Z_3) v_k(Z_3) + o_P(1) \Big] \\
= \ & O(\frac{n-2}{n(n-1)}),
\end{aligned}
$$

$$
\begin{aligned}
& E[S_{n3}^k(\boldsymbol{\beta}_n^*)] \\
= \ & \frac{(n-2)(n-3)}{h^2 n(n-1)} E^2\Big\{ G(Z_1, Z_2)\phi(\frac{Z_1 - Z_2}{h})[\mu_k(Z_1) - \mu_k(Z_2)] \Big\} \\
= \ & \frac{h^4 (n-2)(n-3)}{n(n-1)} E^2[G(Z_2, Z_2)\mu_k''(Z_2)/2 + G^{(1,0)}(Z_2, Z_2)\mu_k'(Z_2) + o_P(1)] \\
= \ & O(\frac{h^4 (n-2)(n-3)}{n(n-1)}).
\end{aligned}
$$

Since $nh \to \infty$ and $nh^4 \to 0$, we have $E[q_{nk}^2(\boldsymbol{\beta}_n^*)] = O_p(n^{-1})$. By similar argument and algorithm, we can also obtain that $Var[q_{nk}^2(\boldsymbol{\beta}_n^*)] = O_p(n^{-2})$. Hence $q_{nk}^2(\boldsymbol{\beta}_n^*) =$

$O_p(E[q_{nk}^2(\boldsymbol{\beta}_n^*)] + \sqrt{Var[q_{nk}^2(\boldsymbol{\beta}_n^*)]}) = O_p(n^{-1})$, which yields that

$$|\mathrm{II}_{n1}| = O_p(\sqrt{\frac{m_n}{n}})\|(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)\|. \tag{S5.2}$$

We next consider II_{n2},

$$\begin{aligned}
\mathrm{II}_{n2} &= \frac{1}{2}(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)^T \nabla^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0) \\
&= \frac{1}{2}(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)^T \left\{\nabla^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0) - E[\nabla^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)]\right\}(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0) \\
&\quad + \frac{1}{2}(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)^T E[\nabla^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)](\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0).
\end{aligned} \tag{S5.3}$$

Note that for any $\varepsilon > 0$, by Chebyshev inequality,

$$\begin{aligned}
&\Pr(\|\nabla^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0) - E[\nabla^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)]\| \geq \frac{\varepsilon}{m_n}) \\
\leq\ & \frac{m_n^2}{\varepsilon^2} E \sum_{k,l=1}^{m_n} \left\{ \left[\frac{\partial^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)}{\partial\beta_k \partial\beta_l}\right]^2 - E^2\left[\frac{\partial^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)}{\partial\beta_k \partial\beta_l}\right] \right\} \\
=\ & \frac{m_n^4}{\varepsilon^2 n}(1 + o(1)) = o(1).
\end{aligned} \tag{S5.4}$$

According to conditions $(C.1^*)$-$(C^*.2)$,

$$\begin{aligned}
& E\left[\frac{\partial^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)}{\partial\beta_k \partial\beta_l}\right] \\
=\ & \frac{1}{h^2 n(n-1)} \sum_{i \neq j} I(Y_i > Y_j)\phi'(\frac{Z_i - Z_j}{h})(X_{ik} - X_{jk})(X_{il} - X_{jl}) \\
=\ & \frac{1}{h^2} E\left[I(Y_1 > Y_2)\phi'(\frac{Z_1 - Z_2}{h})(X_{1k} - X_{2k})(X_{1l} - X_{2l})\right] \\
=\ & \frac{1}{h^2} E\Big[G(Z_1, Z_2)\phi'(\frac{Z_1 - Z_2}{h}) \\
& \left\{c_{kl}(Z_1) + c_{kl}(Z_2) + [\mu_k(Z_1) - \mu_k(Z_2)]\,[\mu_l(Z_1) - \mu_l(Z_2)]\right\}\Big]
\end{aligned}$$

$$\begin{aligned} &= -E[2c_{kl}(Z)G^{(1,0)}(Z,Z) + c'_{kl}(Z)G(Z,Z) + O(h)] \\ &= -E[2c_{kl}(Z)G^{(1,0)}(Z,Z) + c'_{kl}(Z)G(Z,Z)] + O(h). \end{aligned}$$

Let $\|\cdot\|_F$ be the Frobenius norm, that is for any $m \times n$-dimensional matrix A, $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|} = \sqrt{trac(A^T A)}$, then

$$\begin{aligned} &\left\| E[\nabla^2 \widetilde{C}_s(\tilde{\boldsymbol{\beta}}_0)] + E\left[2Cov(\tilde{\mathbf{X}}|Z)G^{(1,0)}(Z,Z) + G(Z,Z)\frac{\partial Cov(\tilde{\mathbf{X}}|Z)}{\partial Z}\right]\right\|_F \\ = \;& O(hm_n) = o(1). \end{aligned} \tag{S5.5}$$

According to Condition $(C^*.3)$, combining (S5.3)-(S5.5) leads to

$$\mathrm{II}_{n2} = -(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0)^T I(\tilde{\boldsymbol{\beta}}_0)(\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0) + o_p(1)\|\boldsymbol{\beta}_n^* - \tilde{\boldsymbol{\beta}}_0\|^2. \tag{S5.6}$$

By the Cauchy-Schwarz inequality and Condition $(C.4^*)$,

$$\begin{aligned} |\mathrm{II}_{n3}| &= \left|\frac{1}{6}\sum_{k,l,t}^{m_n}\frac{\partial^3 \widetilde{C}_s(\boldsymbol{\beta}^{**})}{\partial\beta_k\partial\beta_l\partial\beta_t}(\beta_k^{**}-\beta_{k0})(\beta_l^{**}-\beta_{l0})(\beta_t^{**}-\beta_{t0})\right| \\ &\le \frac{B^3\alpha_n^3}{n(n-1)}\left|\frac{1}{6}\sum_{i\neq j}^{n}\left(\sum_{k,l,t}^{m_n}\left[\frac{\partial^3 \widetilde{C}_s^{(i,j)}(\boldsymbol{\beta}^{**})}{\partial\beta_k\partial\beta_l\partial\beta_t}\right]^2\right)^{1/2}\right| \\ &\le B^3 M\alpha_n^3 O_p(m_n^{\frac{3}{2}}). \end{aligned} \tag{S5.7}$$

We now consider I_{n2},

$$\begin{aligned} \mathrm{I}_{n2} &= -\sum_{j=1}^{s_0}\left[p'_{\lambda_n}(|\beta_{j0}|)sgn(\beta_{j0})(\beta_j^*-\beta_{j0}) + p''_{\lambda_n}(|\beta_{j0}|)(\beta_j^*-\beta_{j0})^2(1+o(1))\right] \\ &\equiv \mathrm{II}_{n4} + \mathrm{II}_{n5}. \end{aligned}$$

The terms II_{n4} and II_{n5} can be dealt with as follows,

$$|\mathrm{II}_4| \leq \sum_{j=1}^{s_0} |p'_{\lambda_n}(|\beta_{j0}|)sgn(\beta_{j0})(\beta_j^* - \beta_{j0})| \leq \sqrt{s_0}\alpha_n a_n B, \tag{S5.8}$$

and

$$|\mathrm{II}_{n5}| \leq \sum_{j=1}^{s_0} |p''_{\lambda_n}(|\beta_{j0}|)(\beta_j^* - \beta_{j0})^2(1+o(1))| \leq \alpha_n^2 \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} s_n B^2 \tag{S5.9}$$

Since $m_n^4/n \to 0$ as $n \to \infty$, $\alpha_n = \sqrt{m_n}(n^{-1/2} + a_n)$, according to conditions (P.1), (P.2) and (S5.2) and (S5.6)-(S5.9), and allowing B to be large enough, all terms II_{n1}, II_{n3}, II_{n4} and II_{n5} are dominated by II_2, which is negative. This proves (S5.1), and then Theorem 3 follows.

Proof of Theorem 4

To prove Theorem 4, we first show that with probability tending to 1, for any given $\boldsymbol{\beta}^{(1)}$ satisfying $\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0^{(1)}\| = O_p(\sqrt{m_n/n})$ and $\|\boldsymbol{\beta}^{(1)}\| = 1$, and any constant B,

$$PC_s\{\begin{pmatrix} \boldsymbol{\beta}^{(1)} \\ 0_{p_n - s_0} \end{pmatrix}\} = \max_{\|\boldsymbol{\beta}^{(2)}\| \leq B(p_n/n)^{1/2}} PC_s\{\begin{pmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{pmatrix}\}. \tag{S5.10}$$

In fact, let $\varepsilon_n = B\sqrt{p_n/n}$, it is sufficient to prove that with probability tending to 1, as $n \to \infty$, for any $\boldsymbol{\beta}^{(1)}$ satisfying $\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0^{(1)}\| = O_p(\sqrt{m_n/n})$ and for $k = s_0 + 1, \cdots, m_n$, we have

$$\frac{\partial pc_s(\tilde{\boldsymbol{\beta}})}{\partial \beta_k} \quad < 0 \quad \text{for } 0 < \beta_k < \varepsilon, \tag{S5.11}$$

$$> 0 \quad \text{for } -\varepsilon < \beta_k < 0. \tag{S5.12}$$

By Taylor expansion,

$$\begin{aligned}
\frac{\partial PC_s(\tilde{\boldsymbol{\beta}})}{\partial \beta_k} &= \frac{\partial PC_s(\tilde{\boldsymbol{\beta}}_0)}{\partial \beta_k} + \sum_{l=1}^{m_n} \frac{\partial^2 PC_s(\tilde{\boldsymbol{\beta}}_0)}{\partial \beta_k \partial \beta_l}(\beta_l - \beta_{l0}) \\
&\quad + \frac{1}{2}\sum_{l,t=1}^{m_n} \frac{\partial^3 PC_s(\bar{\boldsymbol{\beta}})}{\partial \beta_k \partial \beta_l \partial \beta_t}(\beta_l - \beta_{l0})(\beta_t - \beta_{t0}) - p'_{\lambda_n}(|\beta_k|)sgn(\beta_k) \\
&\equiv M_{n1} + M_{n2} + M_{n3} + M_{n4},
\end{aligned}$$

where $\bar{\boldsymbol{\beta}}$ lies between $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}_0$.

By a standard arguments, we have $M_{n1} = O(1/\sqrt{n})$. Now, we consider M_{n2} and M_{n3}.

$$\begin{aligned}
M_{n2} &= \sum_{l=1}^{m_n} \left\{ \frac{\partial^2 PC_s(\tilde{\boldsymbol{\beta}}_0)}{\partial \beta_k \partial \beta_l} - \left[\frac{\partial^2 PC_s(\tilde{\boldsymbol{\beta}}_0)}{\partial \beta_k \partial \beta_l} \right] \right\} (\beta_l - \beta_{l0}) \\
&\quad - \sum_{l=1}^{m_n} I_{k,l}(\tilde{\boldsymbol{\beta}}_0)(\beta_l - \beta_{l0}) + O(h) \sum_{l=1}^{m_n} (\beta_l - \beta_{l0}). \qquad \text{(S5.13)}
\end{aligned}$$

Since that $nh^4 \to 0$ and $m_n^4/n \to 0$ as $n \to \infty$, it follows form $\|\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_0\| = O_p(\sqrt{m_n/n})$ that the third item on the right-hand side of (S5.13) is $o_p(\sqrt{m_n/n})$. According to condition $(C^*.3)$, the eigenvalues of $I_{k,l}(\tilde{\boldsymbol{\beta}}_0)$ are bounded, then

$$\sum_{l=1}^{m_n} I_{k,l}(\tilde{\boldsymbol{\beta}}_0) = O(1),$$

which yields that the second item on the right-hand side of (S5.13) is $O_p(\sqrt{m_n/n})$. As for the first term on the right-hand side of (S5.13), by the Cauchy-Schwarz inequality and using the calculation of (S5.4), we have

$$\left| \sum_{l=1}^{m_n} \left\{ \frac{\partial^2 PC_s(\tilde{\boldsymbol{\beta}}_0)}{\partial \beta_k \partial \beta_l} - \left[\frac{\partial^2 PC_s(\tilde{\boldsymbol{\beta}}_0)}{\partial \beta_k \partial \beta_l} \right] \right\} (\beta_l - \beta_{l0}) \right|$$

$$\begin{aligned}
&\leq \quad \|\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_0\| \left[\sum_{l=1}^{m_n} \left\{ \frac{\partial^2 PC_s(\tilde{\boldsymbol{\beta}}_0)}{\partial\beta_k \partial\beta_l} - \left[\frac{\partial^2 PC_s(\tilde{\boldsymbol{\beta}}_0)}{\partial\beta_k \partial\beta_l} \right] \right\}^2 \right]^{1/2} \\
&= \quad O_p(m_n/n).
\end{aligned}$$

This entails that $M_{n2} = O_p(\sqrt{m_n/n})$. By the same argument of (S5.7) in the proof of Theorem 3, we can obtain that $M_{n3} = o_p(\sqrt{m_n/n})$, and consequently

$$M_{n1} + M_{n2} + M_{n3} = O_p(\sqrt{m_n/n}). \tag{S5.14}$$

Since $\sqrt{m_n/n}/\lambda_n \to 0$ and $\liminf_{n\to\infty} \inf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$, we obtain that

$$\frac{\partial PC_s(\tilde{\boldsymbol{\beta}})}{\partial\beta_k} \quad = \quad \lambda_n \left[O_p(\sqrt{m_n/n}/\lambda_n) - p'_{\lambda_n}(|\beta_k|)/\lambda_n sgn(\beta_k) \right].$$

That is the sight of $\partial PC_s(\tilde{\boldsymbol{\beta}})/\partial\beta_k$ is completely determined by the sign of β_k. Hence, (S5.11) and (S5.12) follow. This prove part (i), that is with probability tending to 1, $\widehat{\boldsymbol{\beta}}$ has the form $(\widehat{\boldsymbol{\beta}}^{(1)T}, 0^T_{m_n-s_0})^T$.

We now prove part (ii).Note that for $\beta_k \neq 0$, $p'_{\lambda_n}(|\beta_k|)sgn(\beta_k) \approx \{p'_{\lambda_n}(|\beta_{k0}|)/|\beta_{k0}|\}\beta_k$, then we can obtain

$$\begin{aligned}
-\nabla PC_s(\boldsymbol{\beta}_0^{(1)}) \quad &= \quad [\nabla^2 PC_s(\boldsymbol{\beta}_0^{(1)}) - \Sigma_{\lambda_n}(\boldsymbol{\beta}_0^{(1)})](\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) - \Sigma_{\lambda_n}(\dot{\boldsymbol{\beta}}^{(1)})\boldsymbol{\beta}_0^{(1)} \\
&\quad + \frac{1}{2}(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)})^T \nabla^2\{\nabla PC_s(\check{\boldsymbol{\beta}}^{(1)})\}(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}), \qquad \text{(S5.15)}
\end{aligned}$$

where $\check{\boldsymbol{\beta}}^{(1)}$ and $\check{\boldsymbol{\beta}}^{(1)}$ lie between $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\boldsymbol{\beta}_0^{(1)}$ By regular arguments we can easily prove that

$$\frac{1}{2}(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)})^T \nabla^2\{\nabla PC_s(\check{\boldsymbol{\beta}}^{(1)})\}(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) = O_p(\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\|^2) = o_p(n^{-1/2}).$$

According to the calculation of II_{n2} in the proof of Theorem 3, by Conditions $(C.1^*)$-$(C.5^*)$ and Condition (P.4),

$$\|\nabla^2 PC_s(\boldsymbol{\beta}_0^{(1)}) - \Sigma_{\lambda_n}(\dot{\boldsymbol{\beta}}^{(1)}) + I(\boldsymbol{\beta}_0^{(1)}) + \Sigma_{\lambda_n}(\boldsymbol{\beta}_0^{(1)})\|_F = O_p(\sqrt{m_n/n}) + O_p(n^{-1/2}) + O(h).$$

Note that since $nh^4 \to 0$ and $m_n^4/n \to 0$, $hm_n \to 0$ as $n \to \infty$ as $n \to \infty$, which implies that

$$\nabla PC_s(\boldsymbol{\beta}_0^{(1)}) = [I(\boldsymbol{\beta}_0^{(1)}) + \Sigma_{\lambda_n}(\boldsymbol{\beta}_0^{(1)})](\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) + \mathbf{b} + o_p(n^{-1/2}). \tag{S5.16}$$

Let $\tilde{\mu}(Z) = E(\tilde{\mathbf{X}}^{(1)}|Z)$, then we can obtain that

$$\begin{aligned}
& E[\nabla PC_s(\boldsymbol{\beta}_0^{(1)})] \\
= \; & E\left\{G(Z_1, Z_2)\phi(\frac{Z_1 - Z_2}{h})[\tilde{\mathbf{X}}^{(1)}(Z_1) - \tilde{\mathbf{X}}^{(1)}(Z_2)]/h\right\} \\
= \; & E\left\{G(Z_1, Z_2)\phi(\frac{Z_1 - Z_2}{h})[\tilde{\mu}(Z_1) - \tilde{\mu}(Z_2)]/h\right\} \\
= \; & h^2 E\left[G(Z, Z)\tilde{\mu}_k''(Z)/2 + G^{(1,0)}(Z, Z)\tilde{\mu}_k'(Z)\right] + o_P(h^2).
\end{aligned}$$

According to the same argument of II_{n1} in the proof of Theorem 3, since $nh \to \infty$ and $nh^4 \to 0$ as $n \to \infty$,

$$\begin{aligned}
& Cov[\nabla PC_s(\boldsymbol{\beta}_0^{(1)})] \\
= \; & \frac{1}{n}E\left[G^2(Z, Z)Cov(\tilde{\mathbf{X}}^{(1)}|Z) + o_P(1)\right] \\
= \; & \frac{1}{n}I^*(\boldsymbol{\beta}_0^{(1)}) + o(n^{-1}).
\end{aligned}$$

Thus, by the central limit theory of U-statistic (Lee, 1990), we have

$$\sqrt{n}\nabla PC_s(\boldsymbol{\beta}_0^{(1)}) \xrightarrow{\mathcal{L}} N(0_{s_0}, I^*(\boldsymbol{\beta}_0^{(1)})).$$

Based on the Slutsky's theorem, it follows from (S5.16) that

$$\sqrt{n}[I(\boldsymbol{\beta}_0^{(1)}) + \Sigma_{\lambda_n}(\boldsymbol{\beta}_0^{(1)})]\left(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)} + [I(\boldsymbol{\beta}_0^{(1)}) + \Sigma_{\lambda_n}(\boldsymbol{\beta}_0^{(1)})]^{-1}\mathbf{b}\right) \xrightarrow{\mathcal{L}} N(0_{s_0}, I^*(\boldsymbol{\beta}_0^{(1)})).$$

References

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models, *J. Am. Statist. Assoc.*, **106**, 544-557.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion), *J.Roy. Statisti. Soc. B.*, **70**, 849-911.

Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Varying Coefficient Models. *J. Am. Statist. Assoc.*, **109**, 1270-1284.

Fan J. and Peng H. (2004). Nonconcave penalized likelihood with a diverging number of parameters, *Ann. Statist.*, **32**, 928-961.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model, *The Journal of Machine Learning Research*, **10**, 2013-2038.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *J. Am. Statist. Assoc.*, **58**, 13-30.

Lee, A. J. (1990). *U-statistics. Theory and Practice*, Marcel Dekker, New York.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*, Springer, New York.

Zhao, D. S., and Li, Y. (2012). Principled Sure Independence Screening for Cox Models With Ultra-High-

Dimensional Covariates. *Journal of Multivariate Analysis,* **105,** 397-411.