

Supplementary Material for “Subgroup Identification and Variable Selection in the High-Dimensional Heterogeneous Cox Model”

Yunshu HUANG

Department of Mathematics, Franklin and Marshall College, Lancaster, PA, 17604, United States

E-mail: huangys1001@gmail.com

Jiehui WANG*

School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei, 430072, China

E-mail: jiehui.wang@whu.edu.cn

S1 Some Lemmas

Lemma S1.1 *Assuming that Condition 5 hold,*

$$\|\sqrt{n}U_{\mathcal{D}}(\boldsymbol{\theta}_0)\| = O_p(\sqrt{s}).$$

Proof. Since $\|\sqrt{n}U_{\mathcal{D}}(\boldsymbol{\theta}_0)\|^2 = ntr\{(U_{\mathcal{D}}(\boldsymbol{\theta}_0))^{\otimes 2}\}$,

$$\begin{aligned} E\{\|\sqrt{n}U_{\mathcal{D}}(\boldsymbol{\theta}_0)\|^2\} &= ntr\{E(U_{\mathcal{D}}(\boldsymbol{\theta}_0))^{\otimes 2}\} \\ &= tr\left\{E\int_0^\tau \mathbf{V}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)S^{(0)}(\boldsymbol{\theta}_0, t)d\Lambda_0(t)\right\}. \end{aligned} \quad (\text{S1.1})$$

For any random variable \mathbf{G} , $E\{(\mathbf{G}_i - \bar{\mathbf{G}})^{\otimes 2}\} \leq E\{\mathbf{G}_i^{\otimes 2}\}$ for any i , where $\bar{\mathbf{G}} = \frac{1}{n} \sum_{i=1}^n \mathbf{G}_i$, and $\mathbf{A} \leq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is nonnegative definite. Hence

$$\begin{aligned} V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t) &= \frac{\sum_{i=1}^n \{\mathbf{B}_{i\mathcal{D}} - E_{\mathcal{D}}(\boldsymbol{\theta}_0, t)\}^{\otimes 2} Y_i(t) \exp(\boldsymbol{\theta}_{0\mathcal{D}}^\top \mathbf{B}_{i\mathcal{D}}(t))}{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\theta}_{0\mathcal{D}}^\top \mathbf{B}_{i\mathcal{D}}(t))} \\ &\leq \frac{\sum_{i=1}^n \mathbf{B}_{i\mathcal{D}}(t)^{\otimes 2} Y_i(t) \exp(\boldsymbol{\theta}_{0\mathcal{D}}^\top \mathbf{B}_{i\mathcal{D}}(t))}{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\theta}_{0\mathcal{D}}^\top \mathbf{B}_{i\mathcal{D}}(t))}. \end{aligned}$$

Then

$$V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)S^{(0)}(\boldsymbol{\theta}_0, t) \leq n^{-1} \sum_{i=1}^n \mathbf{B}_{i\mathcal{D}}(t)^{\otimes 2} Y_i(t) \exp(\boldsymbol{\theta}_{0\mathcal{D}}^\top \mathbf{B}_{i\mathcal{D}}(t)),$$

which gives

$$\begin{aligned} E\left\{\sup_{t \in [0, \tau]} tr[V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)S^{(0)}(\boldsymbol{\theta}_0, t)]\right\} &\leq E\left\{\sup_{t \in [0, \tau]} tr(\mathbf{B}_{i\mathcal{D}}(t)^{\otimes 2} Y_i(t) \exp(\boldsymbol{\theta}_{0\mathcal{D}}^\top \mathbf{B}_{i\mathcal{D}}(t)))\right\} \\ &= E\left\{\sup_{t \in [0, \tau]} \|\mathbf{B}_{i\mathcal{D}}(t)\|^2 Y_i(t) \exp(\boldsymbol{\theta}_{0\mathcal{D}}^\top \mathbf{B}_{i\mathcal{D}}(t))\right\}. \end{aligned}$$

By Condition 5, we have

$$E\left\{\sup_{t \in [0, \tau]} tr[V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)S^{(0)}(\boldsymbol{\theta}_0, t)]\right\} = O(s).$$

Combining with (S1.1), it follows

$$E\{\|\sqrt{n}U_{\mathcal{D}}(\boldsymbol{\theta}_0)\|^2\} = O(s).$$

Applying Lemma 2.1 in [2], the lemma is concluded. \square

Lemma S1.2 Assume that Conditions 4 and 6 hold. Then $\sup_{\boldsymbol{\theta} \in \mathcal{K}} \|\mathcal{I}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}})\| = O_p(1)$, $\|\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1}\| = O_p(1)$ and $\sup_{\boldsymbol{\theta} \in \mathcal{K}} \|\mathcal{I}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) - \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}})\| = o_p(1)$.

Proof. First we have

$$\begin{aligned} \mathcal{I}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) - \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) &= \int_0^\tau \{V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t) - v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t)\} s^{(0)}(\boldsymbol{\theta}_0, t) \lambda_0(t) dt \\ &\quad + \int_0^\tau V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t) \{S^{(0)}(\boldsymbol{\theta}_0, t) - s^{(0)}(\boldsymbol{\theta}_0, t)\} \lambda_0(t) dt \\ &:= H_1(\boldsymbol{\theta}_{\mathcal{D}}) + H_2(\boldsymbol{\theta}_{\mathcal{D}}). \end{aligned}$$

Using Lemma 4.1 (i) and (ii) in [2], we obtain

$$\|H_1(\boldsymbol{\theta}_{\mathcal{D}})\|^2 \leq \Lambda_0(\tau) \int_0^\tau \|V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t) - v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t)\|^2 (s^{(0)}(\boldsymbol{\theta}_0, t))^2 \lambda_0(t) dt,$$

and

$$\|H_2(\boldsymbol{\theta}_{\mathcal{D}})\|^2 \leq \Lambda_0(\tau) \int_0^\tau \|V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t)\|^2 (S^{(0)}(\boldsymbol{\theta}_0, t) - s^{(0)}(\boldsymbol{\theta}_0, t))^2 \lambda_0(t) dt.$$

By Condition 4 (i) and (ii), we have

$$\begin{aligned} \|V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t)\| &\leq \|V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t) - v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t)\| \\ &\quad + \|v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t) - v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{0\mathcal{D}}, t)\| + \|v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{0\mathcal{D}}, t)\| \\ &= O_p(1). \end{aligned}$$

Then $\sup_{\boldsymbol{\theta} \in \mathcal{K}} \|H_1(\boldsymbol{\theta}_{\mathcal{D}})\|^2 = o_p(1)$ and $\sup_{\boldsymbol{\theta} \in \mathcal{K}} \|H_2(\boldsymbol{\theta}_{\mathcal{D}})\|^2 = o_p(1)$. Therefore,

$$\sup_{\boldsymbol{\theta} \in \mathcal{K}} \|\mathcal{I}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) - \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}})\| \leq \sup_{\boldsymbol{\theta} \in \mathcal{K}} \|H_1(\boldsymbol{\theta}_{\mathcal{D}})\| + \sup_{\boldsymbol{\theta} \in \mathcal{K}} \|H_2(\boldsymbol{\theta}_{\mathcal{D}})\| = o_p(1). \quad (\text{S1.2})$$

By Condition 4 (ii), we have

$$\|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}})\|^2 \leq \Lambda_0(\tau) \int_0^\tau \|v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}, t)\|^2 (s^{(0)}(\boldsymbol{\theta}_0, t))^2 \lambda_0(t) dt = O_p(1).$$

This gives

$$\sup_{\boldsymbol{\theta} \in \mathcal{K}} \|\mathcal{I}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}})\| \leq \sup_{\boldsymbol{\theta} \in \mathcal{K}} \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}})\| + \sup_{\boldsymbol{\theta} \in \mathcal{K}} \|\mathcal{I}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) - \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}})\| = O_p(1).$$

Besides, we have

$$\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1} = \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2} \{I + \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2} (\mathcal{I}_{\mathcal{D}\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}) \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}\}^{-1} \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}.$$

Let $\boldsymbol{\mathcal{H}} = I + \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2} (\mathcal{I}_{\mathcal{D}\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}) \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}$, then $\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1} = \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2} \boldsymbol{\mathcal{H}}^{-1} \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}$. Using Bauer-Fike inequality in [1], we obtain

$$|\lambda(\boldsymbol{\mathcal{H}}) - 1| \leq \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2} (\mathcal{I}_{\mathcal{D}\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}) \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}\| \leq \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}\| \|\mathcal{I}_{\mathcal{D}\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}\| \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}\|.$$

From (S1.2) and Condition 6, it follows $|\lambda(\boldsymbol{\mathcal{H}}) - 1| = o_p(1)$, and then $\lambda(\boldsymbol{\mathcal{H}}^{-1}) = 1 + o_p(1)$. Since $\boldsymbol{\mathcal{H}}$ is symmetrical, $\|\boldsymbol{\mathcal{H}}^{-1}\| = O_p(1)$. Therefore, we get $\|\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1}\| \leq \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}\| \|\boldsymbol{\mathcal{H}}^{-1}\| \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2}\| = O_p(1)$. \square

Lemma S1.3 Assume that Condition 4 hold.

$$(i) \quad \|\mathcal{W}_{\mathcal{D}\mathcal{D}}\| = O_p(s/\sqrt{n}).$$

$$(ii) \quad \text{For any consistent estimator } \hat{\boldsymbol{\theta}} \text{ of } \boldsymbol{\theta}_0, \text{ if } s = o(n^{1/2}), \text{ then } \|\mathcal{W}_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}})\| = o_p(1).$$

Proof. (i) Since $\mathcal{W}_{\mathcal{D}\mathcal{D}} = n^{-1} \int_0^\tau V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t) d\bar{M}(t)$ is symmetrical, $\|\mathcal{W}_{\mathcal{D}\mathcal{D}}\|^2 = r_\sigma(\mathcal{W}_{\mathcal{D}\mathcal{D}}^2) \leq \text{tr}(\mathcal{W}_{\mathcal{D}\mathcal{D}}^2)$ and

$$\text{tr}(\mathcal{W}_{\mathcal{D}\mathcal{D}}^2) = n^{-1} \sum_{i,j=1}^{pK_0+s} \left\{ n^{-1/2} \int_0^\tau V_{\mathcal{D}\mathcal{D}}^{(i,j)}(\boldsymbol{\theta}_0, t) d\bar{M}(t) \right\}^2,$$

where $V_{\mathcal{D}\mathcal{D}}^{(i,j)}(\boldsymbol{\theta}_0, t)$ denotes the (i, j) th entry of $V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)$. Let

$$\begin{aligned} \pi_{ij}(t) &= n^{-1/2} \int_0^t V_{\mathcal{D}\mathcal{D}}^{(i,j)}(\boldsymbol{\theta}_0, u) d\bar{M}(u), \\ \mathcal{X}_{ij}(t) &= \int_0^t \{v_{\mathcal{D}\mathcal{D}}^{(i,j)}(\boldsymbol{\theta}_0, u)\}^2 s^{(0)}(\boldsymbol{\theta}_0, u) d\Lambda_0(u). \end{aligned}$$

By Condition 4 (ii) and (iii), $\mathcal{X}_{ij}(t)$ is bounded. Note that $\pi_{ij}(t)$ is a locally square integrable martingale with mean zero and predictable quadratic variation process with

$$\langle \pi_{ij}(t) \rangle = \int_0^t \{V_{\mathcal{D}\mathcal{D}}^{(i,j)}(\boldsymbol{\theta}_0, u)\}^2 S^{(0)}(\boldsymbol{\theta}_0, u) d\Lambda_0(u).$$

By Condition 4 (i), $\langle \pi_{ij}(t) \rangle \rightarrow \mathcal{X}_{ij}(t)$ in probability. By Chebyshev inequality, $\pi_{ij}^2(\tau)$ is bounded in probability. Hence $\text{tr}(\mathcal{W}_{\mathcal{D}\mathcal{D}}^2)$ is of order $O_p(s^2/n)$.

(ii) By Condition 4 (i) and (ii), we have

$$\begin{aligned} & \sup_{t \in [0, \tau]} \|V_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}, t) - V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)\| \\ & \leq \sup_{t \in [0, \tau]} \|V_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}, t) - v_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}, t)\| + \sup_{t \in [0, \tau]} \|V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t) - v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)\| \\ & + \sup_{t \in [0, \tau]} \|v_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}, t) - v_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)\| \\ & = o_p(1). \end{aligned}$$

Note that

$$\mathcal{W}_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}) - \mathcal{W}_{\mathcal{D}\mathcal{D}} = n^{-1} \int_0^\tau \{V_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}, t) - V_{\mathcal{D}\mathcal{D}}(\boldsymbol{\theta}_0, t)\} d\bar{M}(t).$$

Therefore, $\|\mathcal{W}_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}) - \mathcal{W}_{\mathcal{D}\mathcal{D}}\| = o_p(1)$. Then, by (i), we have

$$\|\mathcal{W}_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}})\| \leq \|\mathcal{W}_{\mathcal{D}\mathcal{D}}\| + \|\mathcal{W}_{\mathcal{D}\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}) - \mathcal{W}_{\mathcal{D}\mathcal{D}}\| = o_p(1).$$

□

Lemma S1.4 For a $(pK_0 + s) \times 1$ -dimensional unit vector \mathbf{c}_n , define

$$\phi_n = -\sqrt{n} \mathbf{c}_n^\top \Sigma_{\mathcal{D}\mathcal{D}}^{1/2} \left(\frac{\partial U(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0^\top} \right)_{\mathcal{D}\mathcal{D}}^{-1} U_{\mathcal{D}}(\boldsymbol{\theta}_0).$$

Assume that $\|\mathcal{I}_{\mathcal{D}\mathcal{D}} - \Sigma_{\mathcal{D}\mathcal{D}}\| = O_p(s/\sqrt{n})$. Under Condition 6, $\phi_n = \phi_{n1} + o_p(1)$, where $\phi_{n1} = -\sqrt{n} \mathbf{c}_n^\top \Sigma_{\mathcal{D}\mathcal{D}}^{-1/2} U_{\mathcal{D}}(\boldsymbol{\theta}_0)$.

Proof. Let $\mathcal{L} = \mathbf{I} + \mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2} \mathcal{W}_{\mathcal{D}\mathcal{D}} \mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2}$. Using the Bauer-Fiker inequality in [1],

$$|\lambda(\mathcal{L}) - 1| \leq \|\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2} \mathcal{W}_{\mathcal{D}\mathcal{D}} \mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2}\| \leq \|\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2}\|^2 \|\mathcal{W}_{\mathcal{D}\mathcal{D}}\|.$$

Applying Lemmas S1.2 and S1.3, we have

$$\lambda(\mathcal{L}) = 1 + O_p(s/\sqrt{n}). \quad (\text{S1.3})$$

Noting that

$$-\left(\frac{\partial U(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0^\top}\right)_{\mathcal{D}\mathcal{D}}^{-1} = -\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1} + \mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2} \{\mathbf{I} - \mathcal{L}^{-1}\} \mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2},$$

it follows that

$$\begin{aligned} \phi_n &= -\sqrt{n} \mathbf{c}_n^\top \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1/2} U_{\mathcal{D}}(\boldsymbol{\theta}_0) \\ &\quad + \sqrt{n} \mathbf{c}_n^\top (\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1} - \mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1}) U_{\mathcal{D}}(\boldsymbol{\theta}_0) \\ &\quad + \mathbf{c}_n^\top \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{1/2} \mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2} \{\mathbf{I} - \mathcal{L}^{-1}\} \mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2} \sqrt{n} U_{\mathcal{D}}(\boldsymbol{\theta}_0) \\ &:= \phi_{n1} + \phi_{n2} + \phi_{n3}. \end{aligned}$$

We first consider term ϕ_{n2} . Since $\|\mathbf{c}_n\| = 1$, we have

$$|\phi_{n2}| \leq \sqrt{n} \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1}\| \cdot \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}} - \mathcal{I}_{\mathcal{D}\mathcal{D}}\| \cdot \|\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1}\| \cdot \|U_{\mathcal{D}}(\boldsymbol{\theta}_0)\| = O_p(\sqrt{s^3/n}) = o_p(1), \quad (\text{S1.4})$$

where we use Lemmas S1.1 and S1.2, Condition 6, and the fact that $s = o(n^{1/3})$.

Since $\mathbf{I} - \mathcal{L}^{-1}$ is symmetrical, $r_\sigma(\mathbf{I} - \mathcal{L}^{-1}) = \|\mathbf{I} - \mathcal{L}^{-1}\|$. Noting that $\|\mathbf{c}_n\| = 1$, we have

$$|\phi_{n3}| \leq r_\sigma(\mathbf{I} - \mathcal{L}^{-1}) \|\boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{1/2}\| \|\mathcal{I}_{\mathcal{D}\mathcal{D}}^{-1/2}\|^2 \|\sqrt{n} U_{\mathcal{D}}(\boldsymbol{\theta}_0)\| \leq r_\sigma(\mathbf{I} - \mathcal{L}^{-1}) O_p(\sqrt{s}), \quad (\text{S1.5})$$

by using of Condition 6, and Lemmas S1.1-S1.2. (S1.3) and (S1.5) yield that $\phi_{n3} = O_p(\sqrt{s^3/n})$, and we have $r_\sigma(\mathbf{I} - \mathcal{L}^{-1}) = O_p(s/\sqrt{n})$. Thus, the assumption that $s = o(n^{1/3})$ implies $\phi_n = \phi_{n1} + o_p(1)$. \square

Lemma S1.5 *Define*

$$\tilde{\Theta} = \{\boldsymbol{\beta} \in \mathbb{R}^{np}, \boldsymbol{\eta} \in \mathbb{R}^q : \max_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{0i}\| \leq \epsilon_n, \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\| \leq \epsilon_n\}.$$

For any $(\boldsymbol{\beta}, \boldsymbol{\eta}) \in \tilde{\Theta}$ and large enough n , we have $Q(\boldsymbol{\beta}, \boldsymbol{\eta}) \geq Q(\boldsymbol{\beta}^*, \boldsymbol{\eta})$.

Proof. By mean value theorem, we have

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{\eta}) - Q(\boldsymbol{\beta}^*, \boldsymbol{\eta}) &= \frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \boldsymbol{\beta}^\top} \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \frac{\partial P_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ &=: w_1 + w_2, \end{aligned}$$

where $\tilde{\beta} = m\beta + (1-m)\beta^*$ for some $m \in (0, 1)$. For w_2 ,

$$\begin{aligned} w_2 &= \left. \frac{\partial P_n(\beta)}{\partial \beta^\top} \right|_{\beta=\tilde{\beta}} (\beta - \beta^*) \\ &= \lambda_1 \sum_{1 \leq i \leq j \leq n} \rho_{\lambda_1}^{(1)'}(\|\tilde{\beta}_i - \tilde{\beta}_j\|) \frac{(\tilde{\beta}_i - \tilde{\beta}_j)^\top}{\|\tilde{\beta}_i - \tilde{\beta}_j\|} (\beta_i - \beta_i^*) \\ &\quad + \lambda_1 \sum_{1 \leq j \leq i \leq n} \rho_{\lambda_1}^{(1)'}(\|\tilde{\beta}_j - \tilde{\beta}_i\|) \frac{-(\tilde{\beta}_j - \tilde{\beta}_i)^\top}{\|\tilde{\beta}_j - \tilde{\beta}_i\|} (\beta_i - \beta_i^*) \\ &= \lambda_1 \sum_{1 \leq i \leq j \leq n} \rho_{\lambda_1}^{(1)'}(\|\tilde{\beta}_i - \tilde{\beta}_j\|) \frac{(\tilde{\beta}_i - \tilde{\beta}_j)^\top}{\|\tilde{\beta}_i - \tilde{\beta}_j\|} \{(\beta_i - \beta_i^*) - (\beta_j - \beta_j^*)\}. \end{aligned}$$

On one hand, when subjects i and j are from different groups, that is $i \in \mathcal{G}_{0,k}$ and $j \in \mathcal{G}_{0,k'}$, we have

$$\begin{aligned} \|\beta_{0i} - \beta_{0j}\| &= \|\beta_{0i} - \tilde{\beta}_i + \tilde{\beta}_i - \tilde{\beta}_j + \tilde{\beta}_j - \beta_{0j}\| \\ &\leq \|\beta_{0i} - \tilde{\beta}_i\| + \|\tilde{\beta}_i - \tilde{\beta}_j\| + \|\tilde{\beta}_j - \beta_{0j}\| \\ \|\tilde{\beta}_i - \tilde{\beta}_j\| &\geq \|\beta_{0i} - \beta_{0j}\| - \|\tilde{\beta}_i - \beta_{0i}\| - \|\tilde{\beta}_j - \beta_{0j}\| \\ &\geq \|\beta_{0i} - \beta_{0j}\| - 2 \max_i \|\tilde{\beta}_i - \beta_{0i}\| \\ &= \|\alpha_{0k} - \alpha_{0k'}\| - 2 \max_i \|\tilde{\beta}_i - \beta_{0i}\|. \end{aligned}$$

For any $(\beta, \eta) \in \Theta$, $\max_i \|\beta_i - \beta_{0i}\| \leq \epsilon_n$. By (??), we have $\max_k \|\alpha_k - \alpha_{0k}\| \leq \epsilon_n$ for $\alpha = T^*(\beta)$. Then β^* satisfies that $\max_i \|\beta_i^* - \beta_{0i}\| \leq \epsilon_n$. By the definition of $\tilde{\beta}$, we have

$$\max_i \|\tilde{\beta}_i - \beta_{0i}\| \leq m \max_i \|\beta_i - \beta_{0i}\| + (1-m) \max_i \|\beta_i^* - \beta_{0i}\| \leq m\epsilon_n + (1-m)\epsilon_n = \epsilon_n.$$

Then, we have

$$\|\tilde{\beta}_i - \tilde{\beta}_j\| \geq b - 2\epsilon_n > a_1 \lambda_1.$$

By Condition 2, $\rho_{\lambda_1}^{(1)}(t)$ is a constant when $t > a_1 \lambda_1$. Thus, when subjects i and j are from different groups, $\rho_{\lambda_1}^{(1)'}(\|\tilde{\beta}_i - \tilde{\beta}_j\|) \equiv 0$.

On the other hand, $\beta_i^* = \beta_j^*$ when i and j are from the same group. For $\tilde{\beta} = m\beta + (1-m)\beta^*$, we have $\tilde{\beta}_i - \tilde{\beta}_j = m(\beta_i - \beta_j)$, then $\frac{(\tilde{\beta}_i - \tilde{\beta}_j)^\top}{\|\tilde{\beta}_i - \tilde{\beta}_j\|} = \frac{(\beta_i - \beta_j)^\top}{\|\beta_i - \beta_j\|}$ and

$$\rho_{\lambda_1}^{(1)'}(\|\tilde{\beta}_i - \tilde{\beta}_j\|) \frac{(\tilde{\beta}_i - \tilde{\beta}_j)^\top}{\|\tilde{\beta}_i - \tilde{\beta}_j\|} \{(\beta_i - \beta_i^*) - (\beta_j - \beta_j^*)\} = \rho_{\lambda_1}^{(1)'}(\|\tilde{\beta}_i - \tilde{\beta}_j\|) \|\beta_i - \beta_j\|.$$

Note that

$$\begin{aligned} \max_k \max_{i,j \in \mathcal{G}_{0,k}} \|\tilde{\beta}_i - \tilde{\beta}_j\| &= \max_k \max_{i,j \in \mathcal{G}_{0,k}} \|\tilde{\beta}_i - \beta_i^* + \beta_i^* - \beta_j^* + \beta_j^* - \tilde{\beta}_j\| \\ &\leq 2 \max_i \|\tilde{\beta}_i - \beta_i^*\| \leq 2 \max_i (\|\tilde{\beta}_i - \beta_{0i}\| + \|\beta_i^* - \beta_{0i}\|) \leq 4\epsilon_n. \end{aligned}$$

By Condition 2, we have

$$w_2 = \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}, i < j} \lambda_1 \rho_{\lambda_1}^{(1)'}(\|\tilde{\beta}_i - \tilde{\beta}_j\|) \|\beta_i - \beta_j\| \geq \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}, i < j} \lambda_1 \rho_{\lambda_1}^{(1)'}(4\epsilon_n) \|\beta_i - \beta_j\|.$$

For w_1 , define

$$\begin{aligned} \mathbf{U}_i &= \left. \frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\eta}_A)}{\partial \boldsymbol{\beta}_i} \right|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \\ &= -\frac{1}{n} \int_0^\tau \mathbf{X}_i dN_i(t) + \frac{1}{n} \int_0^\tau \frac{Y_i(t) \mathbf{X}_i \exp(\tilde{\boldsymbol{\beta}}_i^\top \mathbf{X}_i + \boldsymbol{\eta}^\top \mathbf{Z})}{\frac{1}{n} \sum_{j=1}^n Y_j(t) \exp(\tilde{\boldsymbol{\beta}}_j^\top \mathbf{X}_j + \boldsymbol{\eta}^\top \mathbf{Z})} d\tilde{N}(t), \end{aligned}$$

where $\tilde{N}(t) = \frac{1}{n} \sum_{i=1}^n N_i(t)$. Then after some calculation, we have

$$\begin{aligned} w_1 &= \sum_{i=1}^n \mathbf{U}_i^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) = \sum_{k=1}^{K_0} \sum_{i \in \mathcal{G}_{0,k}} \mathbf{U}_i^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) = \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}} \frac{\mathbf{U}_i^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{|\mathcal{G}_{0,k}|} \\ &= \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}} \frac{\mathbf{U}_i^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{2|\mathcal{G}_{0,k}|} + \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}} \frac{\mathbf{U}_j^\top (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)}{2|\mathcal{G}_{0,k}|} \\ &= \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}} \frac{(\mathbf{U}_i - \mathbf{U}_j)^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{2|\mathcal{G}_{0,k}|} \\ &= \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}, i < j} \frac{(\mathbf{U}_i - \mathbf{U}_j)^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{|\mathcal{G}_{0,k}|} \\ &\geq - \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}, i < j} \frac{2 \max_i \|\mathbf{U}_i\| \cdot \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{|\mathcal{G}_{\min}|}, \end{aligned}$$

where $|\mathcal{G}_{\min}| = \min_{k=1, \dots, K_0} |\mathcal{G}_{0,k}|$. Let Ω_{L_1} denote the event that $\max_i |X_i| \leq L_1$ and $\max_j \sup_{t \in [0, \tau]} |Z_j(t)| \leq L_1$ for $L_1 > 0$. By Condition 1 (iii), we have

$$P(\Omega_{L_1}^c) \leq \sum_{i=1}^{pK_0+q} P(\sup_{t \in [0, \tau]} |B_i(t)| > L_1) \leq (pK_0 + q)M_1 \exp\{-M_2 L_1^a\}$$

with probability at least $1 - (pK_0 + q)M_1 \exp\{-M_2 L_1^a\}$ for some constants $M_1, M_2 > 0$. This gives that there exist constant C_1 such that $\|B_i(t)\| \leq C_1$. If $\log q = O(n^\alpha)$, then there a constant C_2 such that $\max_i \|n\mathbf{U}_i\| \leq C_2$ with probability tending to one. By Condition 2, we have $\lim_{n \rightarrow \infty} \rho_{\lambda_1}^{(1)'}(4\epsilon_n) > 0$. Therefore, for large enough n ,

$$Q(\boldsymbol{\beta}, \boldsymbol{\eta}) - Q(\boldsymbol{\beta}^*, \boldsymbol{\eta}) = w_1 + w_2 \geq \sum_{k=1}^{K_0} \sum_{i,j \in \mathcal{G}_{0,k}, i < j} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| [\lambda_1 \rho_{\lambda_1}'(4\epsilon_n) - 2C_2/|n\mathcal{G}_{\min}|] \geq 0.$$

This completes the proof of Lemma S1.5. \square

Lemma S1.6 *Define*

$$\boldsymbol{\xi} = -\frac{1}{n} \sum_{i=1}^n \int_0^\tau (\mathbf{B}_i(t) - E(\boldsymbol{\theta}_0, t)) dM_i(t).$$

Under Conditions 4 and 7, if $v_n = \max_j \sigma_j^2/u_n$ is bounded, then for any sequence $\{u_n\}$ bounded away from zero, there exist positive constants c_0 and c_1 such that

$$P(|\xi_j| > n^{-1/2} u_n) \leq c_0 \exp(-c_1 u_n),$$

where ξ_j is the j th components of $\boldsymbol{\xi}$.

Proof. We write ξ_j as

$$\begin{aligned}\xi_j &= -\frac{1}{n} \sum_{i=1}^n \int_0^\tau (\mathbf{B}_{ij}(t) - e_j(\boldsymbol{\theta}_0, t)) dM_i(t) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \int_0^\tau (E_j(\boldsymbol{\theta}_0, t) - e_j(\boldsymbol{\theta}_0, t)) dM_i(t) \\ &:= \xi_{j1}(\tau) + \xi_{j2}(\tau).\end{aligned}$$

Since $\xi_{j1}(\tau) = \frac{1}{n} \sum_{i=1}^n \psi_{ij}$, where $\{\psi_{ij}\}_{i=1}^n$ is a sequence of i.i.d. random variables with mean zero satisfying Condition 7. It follows from the Bernstein exponential inequality, for any $e > 0$,

$$P(|n\xi_{j1}(\tau)| > e) \leq 2 \exp\{-e^2/2(n\sigma_j^2 + Me)\}. \quad (\text{S1.6})$$

Let $\bar{N}(t) = \sum_{i=1}^n N_i(t)$ and $\Delta N_i(t) = \sum_{i=1}^n \Delta N_i(t)$, where $\Delta N_i(t) = N_i(t) - N_i(t^-)$ denotes the jump of $N_i(\cdot)$ at time t . Since no two counting processes N_i jumps at same time, $|\Delta \bar{N}(t)| \leq 1$. Let $\bar{\Lambda}(t) = \sum_{i=1}^n \Lambda_i(t)$. By continuity of the compensator $\Lambda_i(t)$, $|\Delta \bar{\Lambda}(t)| = 0$. Then $|\Delta \bar{M}(t)| = |\Delta \bar{N}(t)| \leq 1$. It can be shown that

$$\begin{aligned}|\Delta(\sqrt{n}\xi_{j2}(t))| &\leq n^{-1/2}|E_j(\boldsymbol{\theta}_0, t) - e_j(\boldsymbol{\theta}_0, t)| \\ &\leq n^{-1/2} \sup_{t \in [0, \tau]} \|E(\boldsymbol{\theta}_0, t) - e(\boldsymbol{\theta}_0, t)\|_\infty \\ &= n^{-1/2} c_n,\end{aligned}$$

which is bounded almost surely by Condition 4 (iv). Note that

$$\begin{aligned}\langle \sqrt{n}\xi_{j2}(t) \rangle &= n^{-1} \int_0^t (E_j(\boldsymbol{\theta}_0, t) - e_j(\boldsymbol{\theta}_0, t))^2 d\langle \bar{M}(u) \rangle \\ &= \int_0^t (E_j(\boldsymbol{\theta}_0, t) - e_j(\boldsymbol{\theta}_0, t))^2 S^{(0)}(\boldsymbol{\theta}_0, t) d\Lambda_0(t) \\ &\leq \int_0^t \|E(\boldsymbol{\theta}_0, t) - e(\boldsymbol{\theta}_0, t)\|_\infty^2 S^{(0)}(\boldsymbol{\theta}_0, t) d\Lambda_0(t) := b_n^2(t).\end{aligned}$$

Obviously,

$$b_n^2(t) \leq b_n^2(\tau) \leq c_n^2 \int_0^\tau S^{(0)}(\boldsymbol{\theta}_0, t) d\Lambda_0(t).$$

Since

$$\int_0^\tau S^{(0)}(\boldsymbol{\theta}_0, t) d\Lambda_0(t) \leq \int_0^\tau s^{(0)}(\boldsymbol{\theta}_0, t) d\Lambda_0(t) + f_n \Lambda_0(\tau),$$

by Condition 4, there exist constants $0 \leq h < \infty$ and $0 < \varpi < \infty$, independent of j , such that $|\Delta(\sqrt{n}\xi_{j2}(t))| \leq h$ and $\langle \sqrt{n}\xi_{j2}(t) \rangle \leq \varpi^2$. It follows from Lemma 2.1 of [6], for any $u_n > 0$,

$$P(|\xi_{j2}(\tau)| > n^{-1/2}u_n) = P(|\sqrt{n}\xi_{j2}(\tau)| > u_n) \leq 2 \exp\left\{-\frac{u_n^2}{2(hu_n + \varpi^2)}\right\}.$$

There exists a constant $c > 0$ such that

$$P(|\xi_{j2}(\tau)| > n^{-1/2}u_n) \leq 2 \exp\{-cu_n\} \quad (\text{S1.7})$$

uniformly over j . Note that

$$P(|\xi_j(\tau)| > n^{-1/2}u_n) \leq P(|\xi_{j1}(\tau)| > 0.5n^{-1/2}u_n) + P(|\xi_{j2}(\tau)| > 0.5n^{-1/2}u_n).$$

It follows from (S1.6) and (S1.7), $P(|\xi_j(\tau)| > n^{-1/2}u_n)$ is bounded by

$$2 \exp \left\{ -\frac{u_n}{4(2\sigma_j^2 u_n^{-1} + Mn^{-1/2})} \right\} + 2 \exp\{-0.5cu_n\}.$$

Then there exist positive constants c_0 and c_1 satisfy $P(|\xi_j(\tau)| > n^{-1/2}u_n) < c_0 \exp\{-c_1u_n\}$ uniformly over j , if $\max_j \sigma_j^2 = O(u_n)$. \square

Lemma S1.7 *Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}) = (T^*(\hat{\boldsymbol{\beta}}^*), \hat{\boldsymbol{\eta}}) \in \mathbb{R}^{pK_0+q}$. If Condition 2 holds, then $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\eta}})$ is a strict local minimizer of $Q^{\mathcal{M}_2}(\boldsymbol{\beta}^*, \boldsymbol{\eta})$ when the following conditions hold*

$$U_{\hat{\mathcal{A}}}(\hat{\boldsymbol{\theta}}) + \lambda_2 \rho_{\lambda_2}^{(2)'}(|\hat{\boldsymbol{\eta}}_{\hat{\mathcal{A}}}|) \circ \text{sgn}(\hat{\boldsymbol{\eta}}_{\hat{\mathcal{A}}}) = \mathbf{0}, \quad (\text{S1.8})$$

$$U_{\hat{\mathcal{C}}}(\hat{\boldsymbol{\theta}}) + \frac{P_n^{(1)}(\hat{\boldsymbol{\beta}}^*)}{\partial \boldsymbol{\beta}^\top} = \mathbf{0}, \quad (\text{S1.9})$$

$$\|U_{\hat{\mathcal{A}}^c}(\hat{\boldsymbol{\theta}})\|_\infty < \lambda_2 \rho_{\lambda_2}^{(2)'}(0+), \quad (\text{S1.10})$$

$$\lambda_{\min} \left\{ \frac{1}{n} \int_0^\tau V_{\hat{\mathcal{A}}\hat{\mathcal{A}}}(\hat{\boldsymbol{\theta}}, t) d\bar{N}(t) \right\} > \lambda_2 \kappa(\rho_{\lambda_2}^{(2)}, \hat{\boldsymbol{\eta}}_{\hat{\mathcal{A}}}), \quad (\text{S1.11})$$

$$\lambda_{\min} \left\{ \frac{1}{n} \int_0^\tau V_{\hat{\mathcal{C}}\hat{\mathcal{C}}}(\hat{\boldsymbol{\theta}}, t) d\bar{N}(t) \right\} > \lambda_1 \kappa(\rho_{\lambda_1}^{(1)}, \hat{\boldsymbol{\beta}}^*), \quad (\text{S1.12})$$

where $\mathcal{C} = \{1, \dots, pK_0\}$ and \circ represents Hadamard product.

Proof. Let $\mathcal{J} = \{(\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathbb{R}^{np} \times \mathbb{R}^q : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j = \boldsymbol{\alpha}_k, i, j \in \mathcal{G}_{0,k}, 1 \leq k \leq K_0, \boldsymbol{\eta}_{\hat{\mathcal{A}}^c} = \mathbf{0}\}$. (S1.8) and (S1.9) imply that $\hat{\boldsymbol{\theta}}$ is a stationary point. (S1.11) and (S1.12) ensure that the objective function $Q^{\mathcal{M}_2}(\boldsymbol{\beta}^*, \boldsymbol{\eta})$ is strictly convex in a neighborhood of $\hat{\boldsymbol{\theta}}$ in \mathcal{J} . Hence, $\hat{\boldsymbol{\theta}}$ is a strict local minimizer of $Q^{\mathcal{M}_2}(\boldsymbol{\beta}^*, \boldsymbol{\eta})$ in the subspace \mathcal{J} .

It remains to show that for any $(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(1)}) \in \mathcal{M}_2 \setminus \mathcal{J}$ that lies in a sufficiently small neighborhood of $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\eta}})$ such that $Q(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(1)}) > Q(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\eta}})$. To this end, let $\boldsymbol{\eta}^{(2)}$ be the projection of $\boldsymbol{\eta}^{(1)}$ onto the subspace \mathcal{J} . Since $Q(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(2)}) > Q(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\eta}})$, we only need to prove that $Q(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(1)}) > Q(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(2)})$. Note that

$$\begin{aligned} Q(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(1)}) - Q(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(2)}) &= \sum_{j \in \hat{\mathcal{A}}^c, \eta_j^{(1)} \neq 0} \frac{\partial Q(\hat{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\eta}})}{\partial \eta_j} \eta_j^{(1)} \\ &= \sum_{j \in \hat{\mathcal{A}}^c, \eta_j^{(1)} \neq 0} \{U_j(\hat{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\eta}}) + \lambda_2 \rho_{\lambda_2}^{(2)'}(|\tilde{\boldsymbol{\eta}}_j|) \text{sgn}(\tilde{\boldsymbol{\eta}}_j)\} \eta_j^{(1)}, \end{aligned}$$

where $\tilde{\boldsymbol{\eta}}$ is a point on the line segment between $\boldsymbol{\eta}^{(1)}$ and $\boldsymbol{\eta}^{(2)}$. It follows from (S1.10) and continuity that $|U_j(\hat{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\eta}})| < \lambda_2 \rho_{\lambda_2}^{(2)'}(|\tilde{\boldsymbol{\eta}}_j|) \text{sgn}(\tilde{\boldsymbol{\eta}}_j)$ for all $j \in \hat{\mathcal{A}}^c$, where $U_j(\hat{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\eta}}) = U_{\hat{\mathcal{A}}_j^c}(\hat{\boldsymbol{\theta}})$. Using the fact that $\text{sgn}(\tilde{\boldsymbol{\eta}}_j) = \text{sgn}(\eta_j^{(1)})$, it follows $Q(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(1)}) > Q(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\eta}^{(2)})$. This completes the proof of Lemma S1.7. \square

S2 Some details of the algorithm

S2.1 Coordinate Descent Algorithm

For the fixed $\beta = \hat{\beta}$, we consider the following objective function:

$$Q(\hat{\beta}, \eta, \Xi) = \tilde{g}(\Xi; \Xi') + \sum_{i=1}^q P_{\lambda_2}^{(2)}(|\eta_i|).$$

According to [4], we consider the following weighted objective function to balance the regularization strengths on different components of η :

$$\tilde{Q}(\hat{\beta}, \eta, \Xi) = \tilde{g}(\Xi; \Xi') + \sum_{i=1}^q \mathbf{G}_{ii} P_{\lambda_2}^{(2)}(|\eta_i|),$$

where $\mathbf{G} = \mathbf{Z}'\mathbf{H}\mathbf{Z}$ with $\mathbf{Z} = (\mathbf{Z}'(T_{ij}), j \in \mathcal{R}_i, i = 1, \dots, n)$. It suffices to minimize the following objective function in order to update η :

$$\bar{Q}(\hat{\beta}, \eta, \Xi) = \frac{1}{2} \eta^\top \mathbf{H} \eta - \eta^\top \mathbf{b} + \sum_{i=1}^q \mathbf{G}_{ii} P_{\lambda_2}^{(2)}(|\eta_i|),$$

where $\mathbf{b} = \mathbf{Z}'(\mathbf{H}\Xi' - \mathbf{H}\mathbf{X}\beta - \nabla g(\Xi'))$. The first order derivative at η_j can be estimated by solving

$$\frac{\partial \bar{Q}(\hat{\beta}, \eta, \Xi)}{\partial \eta_j} = \mathbf{G}_{jj} \eta_j + \eta_{-j}^\top \mathbf{G}_{j,-j} + \mathbf{G}_{j,j} P_{\lambda_2}^{(2)'}(|\eta_j|) = 0,$$

where η_{-i} represents the vector composing of the remaining $(i-1)$ elements of η after the i th element is removed, $\mathbf{G}_{i,-i}$ denote the vector formed by the i th column of \mathbf{G} , with the i th element removed. Define $f_j = (b_j - \eta_{-j}^\top \mathbf{G}_{j,-j}) \mathbf{G}_{j,-j}$, the coordinate descent algorithm for the group-SCAD penalty can be solved by using the univariate soft thresholding operator:

$$\eta_j = \begin{cases} \text{ST}(f_j, \lambda_2), & |f_j| \leq 2\lambda_2; \\ \frac{\text{sign}(f_j)(|f_j| - \frac{\gamma_2 \lambda_2}{\gamma_2 - 1})}{1 - 1/(\gamma_2 - 1)}, & 2\lambda_2 < |f_j| \leq \gamma_2 \lambda_2; \\ f_j, & |f_j| > \gamma_2 \lambda_2, \end{cases} \quad (\text{S2.1})$$

For the group-MCP penalty with parameter γ_2 ,

$$\eta_j = \begin{cases} \frac{\text{ST}(f_j, \lambda_2)}{1 - 1/\gamma_2}, & |f_j| \leq \gamma_2 \lambda_2; \\ f_j, & |f_j| > \gamma_2 \lambda_2, \end{cases} \quad (\text{S2.2})$$

where $\text{ST}(t, \lambda) = \text{sign}(t)(|t| - \lambda)_+$ is the soft thresholding rule, and $(x)_+ = x$ if $x > 0$ and 0 otherwise. The Coordinate Descent Algorithm is summarized as follows.

Algorithm S.1. Coordinate Descent Algorithm

1. Set the initial value of $\tilde{\boldsymbol{\eta}}$;
 - repeat**
 2. Compute $\tilde{\boldsymbol{\Xi}}'$, $\nabla g(\tilde{\boldsymbol{\Xi}}')$, and $\tilde{\mathbf{b}}$;
 3. For $j = 1, \dots, q$, cyclically update the j th component $\hat{\eta}_j$ of $\hat{\boldsymbol{\eta}}$ using equations (S2.1) and (S2.2) for SCAD and MCP, respectively;
 4. Set $\tilde{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}$;
 - until** convergence of $\hat{\boldsymbol{\eta}}$.
-

S2.2 Majorized ADMM Algorithm

For the fixed $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$, let $\hat{\mathcal{A}}$ is the active set of $\hat{\boldsymbol{\eta}}$. we only need to consider the objective function as follows:

$$Q(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \boldsymbol{\Xi}) = \tilde{g}(\boldsymbol{\Xi}; \boldsymbol{\Xi}') + \sum_{1 \leq i < j \leq n} P_{\lambda_1}^{(1)}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|) \quad (\text{S2.3})$$

Let $\boldsymbol{\delta}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j$, we reformulate the optimal problem (??) as minimizing the following objective function:

$$Q(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \boldsymbol{\Xi}, \boldsymbol{\delta}) = \tilde{g}(\boldsymbol{\Xi}; \boldsymbol{\Xi}') + \sum_{1 \leq i < j \leq n} P_{\lambda_1}^{(1)}(\|\boldsymbol{\delta}_{ij}\|) \quad (\text{S2.4})$$

subject to $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} = \mathbf{0}$,

where $\boldsymbol{\delta} = \{\boldsymbol{\delta}_{i,j}^\top, i < j\}^\top$. Following [3], the augmented Lagrangian for (S2.4) is

$$\begin{aligned} Q'(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \boldsymbol{\Xi}, \boldsymbol{\delta}; \boldsymbol{\nu}, \boldsymbol{\varrho}, \boldsymbol{\Xi}') &= \tilde{g}(\boldsymbol{\Xi}; \boldsymbol{\Xi}') + \sum_{1 \leq i < j \leq n} P_{\lambda_1}^{(1)}(\|\boldsymbol{\delta}_{ij}\|) \\ &+ \sum_{i=1}^n \langle \nu_i, \boldsymbol{\Xi}_i(T_i) - \boldsymbol{\beta}_i^\top \mathbf{X}_i - \hat{\boldsymbol{\eta}}^\top \mathbf{Z}_i(T_i) \rangle + \sum_{1 \leq i < j \leq n} \langle \boldsymbol{\varrho}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} \rangle \\ &+ \frac{\rho}{2} \sum_{i=1}^n (\boldsymbol{\Xi}_i(T_i) - \boldsymbol{\beta}_i^\top \mathbf{X}_i - \hat{\boldsymbol{\eta}}^\top \mathbf{Z}_i(T_i))^2 + \frac{\rho}{2} \sum_{1 \leq i < j \leq n} \|\boldsymbol{\varrho}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\|^2, \end{aligned} \quad (\text{S2.5})$$

where the Lagrangian dual variables $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)^\top$ and $\boldsymbol{\varrho} = (\boldsymbol{\varrho}_{ij}^\top, i < j)^\top$ is the penalty parameter. This optimal problem is equivalent to (S2.4).

At the m th iteration, for a given $(\boldsymbol{\beta}^{(m-1)}, \boldsymbol{\Xi}^{(m-1)(t)}, \boldsymbol{\delta}^{(m-1)}; \boldsymbol{\nu}^{(m-1)}, \boldsymbol{\varrho}^{(m-1)}, \boldsymbol{\Xi}'^{(m-1)})$, cluster size $K^{(m-1)}$ and subgroup set $\mathcal{G}^{(m-1)}$, we update the parameters by the following steps:

Step 1. Updata $\boldsymbol{\beta}^{(m)}$ by minimizing

$$Q'(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \boldsymbol{\Xi}^{(m-1)}, \boldsymbol{\delta}^{(m-1)}; \boldsymbol{\nu}^{(m-1)}, \boldsymbol{\varrho}^{(m-1)}, \boldsymbol{\Xi}'^{(m-1)}).$$

For a given $(\boldsymbol{\Xi}, \boldsymbol{\delta}; \boldsymbol{\nu}, \boldsymbol{\varrho}, \boldsymbol{\Xi})$, it is equivalent to the minimizing the following function:

$$\begin{aligned} &\sum_{i=1}^n \langle \nu_i, Y_i(t) - \boldsymbol{\beta}_i^\top \mathbf{X}_i - \hat{\boldsymbol{\eta}}^\top \mathbf{Z}_i(t) \rangle + \sum_{1 \leq i < j \leq n} \langle \boldsymbol{\varrho}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} \rangle \\ &+ \frac{\rho}{2} \sum_{i=1}^n (Y_i(t) - \boldsymbol{\beta}_i^\top \mathbf{X}_i - \hat{\boldsymbol{\eta}}^\top \mathbf{Z}_i(t))^2 + \frac{\rho}{2} \sum_{1 \leq i < j \leq n} \|\boldsymbol{\varrho}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\|^2. \end{aligned}$$

This is also equivalent to minimizing

$$\begin{aligned} & \langle \boldsymbol{\nu}, \boldsymbol{\Xi} - \mathbf{Z}\hat{\boldsymbol{\eta}} - \widetilde{\mathbf{X}}_{\mathcal{G}}\boldsymbol{\alpha} \rangle + \langle \boldsymbol{\varrho}, \widetilde{\mathbf{A}}\boldsymbol{\alpha} - \boldsymbol{\delta} \rangle \\ & + \frac{\rho}{2} \|\boldsymbol{\Xi} - \mathbf{Z}\hat{\boldsymbol{\eta}} - \widetilde{\mathbf{X}}_{\mathcal{G}}\boldsymbol{\alpha}\|^2 + \frac{\rho}{2} \|\widetilde{\mathbf{A}}\boldsymbol{\alpha} - \boldsymbol{\delta}\|^2, \end{aligned} \quad (\text{S2.6})$$

where $\widetilde{\mathbf{A}} = \mathbf{A}\widetilde{\mathbf{W}}_{\mathcal{G}}$. By setting the first derivative of (S2.6) to equal to zero, we get

$$\begin{aligned} \boldsymbol{\alpha}^{(m)} &= (\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} + \widetilde{\mathbf{X}}_{\mathcal{G}(m-1)}^\top \widetilde{\mathbf{X}}_{\mathcal{G}(m-1)})^{-1} (\widetilde{\mathbf{X}}_{\mathcal{G}(m-1)}^\top (\boldsymbol{\Xi} + \boldsymbol{\nu}/\rho - \mathbf{Z}\hat{\boldsymbol{\eta}}) + \widetilde{\mathbf{A}}^\top (\boldsymbol{\delta} - \boldsymbol{\varrho}/\rho)), \\ \boldsymbol{\beta}^{(m)} &= \widetilde{\mathbf{W}}_{\mathcal{G}(m-1)} \boldsymbol{\alpha}^{(m)}. \end{aligned} \quad (\text{S2.7})$$

Step 2. Update $\boldsymbol{\Xi}^{(m)}$ by minimizing

$$Q'(\boldsymbol{\beta}^{(m)}, \hat{\boldsymbol{\eta}}, \boldsymbol{\Xi}, \boldsymbol{\delta}^{(m-1)}; \boldsymbol{\nu}^{(m-1)}, \boldsymbol{\varrho}^{(m-1)}, \boldsymbol{\Xi}'^{(m-1)}).$$

It is equivalent to the minimizing the following function:

$$\begin{aligned} & \langle \boldsymbol{\Xi}, \nabla g(\boldsymbol{\Xi}') \rangle + \frac{1}{2} \|\boldsymbol{\Xi} - \boldsymbol{\Xi}'\|_H^2 \\ & + \sum_{i=1}^n \langle \nu_i, \Xi_i(T_j) - \mathbf{X}_i^\top \boldsymbol{\beta}_i - \mathbf{Z}_i'(T_j) \hat{\boldsymbol{\eta}} \rangle + \frac{\rho}{2} \sum_{i=1}^n (\Xi_i(T_j) - \boldsymbol{\beta}_i^\top \mathbf{X}_i - \hat{\boldsymbol{\eta}}^\top \mathbf{Z}_i(T_j))^2. \end{aligned} \quad (\text{S2.8})$$

That is,

$$\Xi_i^{(m)}(T_j) = (\tilde{h}_i + \rho)^{-1} [-\nabla_i g(\boldsymbol{\Xi}'^{(m-1)}(T_j)) - \nu_i^{(m-1)} + \rho(\mathbf{X}_i^\top \boldsymbol{\beta}_i^{(m)} + \hat{\boldsymbol{\eta}}^\top \mathbf{Z}_i(T_j))]. \quad (\text{S2.9})$$

Thus, $\boldsymbol{\Xi}'^{(m)}(T_j)$ is updated by

$$\boldsymbol{\Xi}'^{(m)}(T_j) = \mathbf{X}_i^\top \boldsymbol{\beta}_i + \hat{\boldsymbol{\eta}}^\top \mathbf{Z}_i(T_j). \quad (\text{S2.10})$$

Step 3. Update $\boldsymbol{\delta}_{ij}^{(m)}$ by minimizing

$$Q'(\boldsymbol{\beta}^{(m)}, \hat{\boldsymbol{\eta}}, \boldsymbol{\Xi}^{(m)}, \boldsymbol{\delta}; \boldsymbol{\nu}^{(m-1)}, \boldsymbol{\varrho}^{(m-1)}, \boldsymbol{\Xi}'^{(m)}).$$

For a given $(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \boldsymbol{\Xi}, \boldsymbol{\nu}, \boldsymbol{\varrho}, \boldsymbol{\Xi}')$,

$$\boldsymbol{\delta}_{ij} = \arg \min_{\boldsymbol{\delta}_{ij}} \frac{1}{2} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j + \frac{\boldsymbol{\varrho}_{ij}}{\rho} - \boldsymbol{\delta}_{ij}\|^2 + \frac{1}{\rho} P_{\lambda_1}^{(1)}(\|\boldsymbol{\delta}_{ij}\|). \quad (\text{S2.11})$$

We can get the closed form of $\boldsymbol{\delta}_{ij}^{(m)}$. For the SCAD penalty with $\gamma_1 > 1/\rho + 1$, we have

$$\boldsymbol{\delta}_{ij}^{(m)} = \begin{cases} \text{ST}(\boldsymbol{l}_{ij}^{(m-1)}, \lambda_1/\rho), & \|\boldsymbol{l}_{ij}^{(m-1)}\| \leq \lambda_1 + \lambda_1/\rho; \\ \frac{(\rho(\gamma_1-1)-\gamma_1\lambda_1/\|\boldsymbol{l}_{ij}^{(m-1)}\|)\boldsymbol{l}_{ij}^{(m-1)}}{\rho\gamma_1-\rho-1}, & 2\lambda_1 + \lambda_1/\rho < \|\boldsymbol{l}_{ij}^{(m-1)}\| \leq \gamma_1\lambda_1; \\ \boldsymbol{l}_{ij}^{(m-1)}, & \|\boldsymbol{l}_{ij}^{(m-1)}\| > \gamma_1\lambda_1, \end{cases} \quad (\text{S2.12})$$

where $\boldsymbol{l}_{ij}^{(m-1)} = \boldsymbol{\beta}_i^{(m)} - \boldsymbol{\beta}_j^{(m)} + \frac{\boldsymbol{\delta}_{ij}^{(m-1)}}{\rho}$. For the MCP penalty with $\gamma_1 > 1/\rho$, we have

$$\boldsymbol{\delta}_{ij}^{(m)} = \begin{cases} \text{ST}(\frac{\rho\boldsymbol{l}_{ij}^{(m-1)}}{\rho-1/\gamma_1}, \frac{\lambda_1}{\rho-1/\gamma_1}), & \|\boldsymbol{l}_{ij}^{(m-1)}\| \leq \gamma_1\lambda_1; \\ \boldsymbol{l}_{ij}^{(m-1)}, & \|\boldsymbol{l}_{ij}^{(m-1)}\| > \gamma_1\lambda_1. \end{cases} \quad (\text{S2.13})$$

Step 4. Update $\boldsymbol{\nu}^{(m)}$ and $\boldsymbol{\varrho}^{(m)}$ by

$$\begin{aligned}\nu_i^{(m)} &= \nu_i^{(m-1)} + \mathfrak{e} \varrho(Y_i^{(m)} - \mathbf{X}_i^\top \boldsymbol{\beta}_i^{(m)} - \hat{\boldsymbol{\eta}}^\top \mathbf{Z}_i(t)); \\ \varrho_{ij}^{(m)} &= \varrho_{ij}^{(m-1)} + \mathfrak{e} \varrho(\boldsymbol{\beta}_i^{(m)} - \boldsymbol{\beta}_j^{(m)} - \boldsymbol{\delta}_{ij}^{(m)}),\end{aligned}\tag{S2.14}$$

where the constant $\mathfrak{e} \in (0, (1 + \sqrt{5})/2)$.

Step 5. Update $K^{(m)}$ and $\mathcal{G}^{(m)}$ by $\tilde{\boldsymbol{\delta}}_{ij}^{(m)}$, where

$$\tilde{\boldsymbol{\delta}}_{ij}^{(m)} = \arg \min_{\tilde{\boldsymbol{\delta}}_{ij}} \frac{1}{2} \|\boldsymbol{\beta}_i^{(m)} - \boldsymbol{\beta}_j^{(m)} - \tilde{\boldsymbol{\delta}}_{ij}\|^2 + P_{\lambda_1}^{(1)}(\|\tilde{\boldsymbol{\delta}}_{ij}\|).\tag{S2.15}$$

Here $\tilde{\boldsymbol{\delta}}_{ij} = \mathbf{0}$ means that individuals i and j are in the same subgroup and can be used to update \mathcal{G} and K .

As suggested by [5], we set $K^{(0)} = \lfloor \sqrt{n} \rfloor$ to ensure that it is sufficiently large, where $\lfloor a \rfloor$ denotes the largest integer no greater than a . A cluster analysis method can then be applied to determine $\mathcal{G}^{(0)} = (\mathcal{G}_1^{(0)}, \dots, \mathcal{G}_{K^{(0)}}^{(0)})$. Let $\boldsymbol{\Xi}^{(0)} = \boldsymbol{\Xi}'^{(0)} = \boldsymbol{\beta}^{(0)\top} \mathbf{X} + \hat{\boldsymbol{\eta}}^\top \mathbf{Z}$, $\boldsymbol{\delta}^{(0)} = \mathbf{A} \boldsymbol{\beta}^{(0)}$, $\boldsymbol{\nu}^{(0)} = \mathbf{0}$, and $\boldsymbol{\varrho}^{(0)} = \mathbf{0}$. Define $r^{(m)} = \|\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)}\| + |K^{(m)} - K^{(m-1)}|$. The Majorized ADMM Algorithm is summarized as follows.

Algorithm S.2. Majorized ADMM Algorithm

1. Set $m \leftarrow 0$, initialize $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\Xi}^{(0)}, \boldsymbol{\delta}^{(0)}; \boldsymbol{\nu}^{(0)}, \boldsymbol{\varrho}^{(0)}, \boldsymbol{\Xi}'^{(0)})$, $K^{(0)}$ and $\mathcal{G}^{(0)}$;
 - repeat**
 2. $m \leftarrow m + 1$
 3. Update $\boldsymbol{\beta}^{(m)}$ by (S2.2), $(\boldsymbol{\Xi}^{(m)}, \boldsymbol{\Xi}'^{(m)})$ by (S2.9) and (S2.10), and $\boldsymbol{\delta}^{(m)}$ by (S2.12) and (S2.13) for SCAD and MCP respectively;
 4. Update $\tilde{\boldsymbol{\delta}}_{ij}^{(m)}$ by (S2.15), and then update $K^{(m)}$ and \mathcal{G}^m according to $\tilde{\boldsymbol{\delta}}_{ij}^{(m)}$;
 - until** convergence of $r^{(m)}$.
-

References

- [1] Bhatia, R.: Matrix Analysis, Springer-Verlag, New York, 1997.
- [2] Bradic, J., Fan, J., and Jiang, J.: Regularization for Cox's proportional hazards model with np-dimensionality. *The Annals of Statistics*, **39**, 3092–3120 (2011).
- [3] Hu, X., Huang, J., Liu, L., Sun, D., and Zhao, X.: Subgroup analysis in the heterogeneous Cox model. *Statistics in Medicine*, **40**(3), 739–757 (2021).
- [4] Lin, W., and Lv, J.: High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, **108**, 247–264 (2013).
- [5] Ma, S., Huang, J., Zhang, Z., and Liu, M.: Exploration of heterogeneous treatment effects via concave fusion. *The International Journal of Biostatistics*, **16**, 2018–2026 (2020).
- [6] van de Geer S.: Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, **23**, 1779–1801 (1995).