PENALIZED PAIRWISE PSEUDO LIKELIHOOD FOR VARIABLE SELECTION WITH NONIGNORABLE MISSING DATA

Jiwei Zhao[1], Yang Yang[1] and Yang Ning[2]

[1]*State University of New York at Buffalo,* [2]*Cornell University*

Supplementary Material

S1 Theoretical Derivations

Verification of the Assumption 2. For the lower bound of $\rho_{-}(s)$, denote $F_{ij} = \{|y_i| \leq \tau\} \cap \{|y_j| \leq \tau\}$, where τ is a positive constant, we have

$$\begin{aligned} \nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*) &\geq \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \{\psi''(y_{i\backslash j} \boldsymbol{x}_{i\backslash j}^T \boldsymbol{\gamma}^*) y_{i\backslash j}^2 \boldsymbol{x}_{i\backslash j}^{\otimes 2} I(F_{ij})\} \\ &\geq c_3 \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \{y_{i\backslash j}^2 \boldsymbol{x}_{i\backslash j}^{\otimes 2} I(F_{ij})\} \triangleq \mathbf{W}, \end{aligned}$$

where $c_3 = \exp(-4B\tau)\{1 + \exp(4B\tau)\}^{-2}$.

According to the arguments in the proof of Theorem 3.10 in Ning et al. (2017), for any $\boldsymbol{v} \in \mathcal{F}$, where

$$\mathcal{F} = \{\boldsymbol{\Delta} \in R^p : \|\boldsymbol{\Delta}\|_0 = s, \|\boldsymbol{\Delta}\|_2 = 1\},$$

we have

$$\begin{aligned} |\boldsymbol{v}^T\mathbf{W}\boldsymbol{v} - \boldsymbol{v}^T E(\mathbf{W})\boldsymbol{v}| &\leq \|\boldsymbol{v}\|_1^2 \|\mathbf{W} - E(\mathbf{W})\|_\infty \\ &\leq s\|\mathbf{W} - E(\mathbf{W})\|_\infty. \end{aligned}$$

Hence, $\rho_-(\mathbf{W}, s) \geq \rho_-(E(\mathbf{W}), s) - s\|\mathbf{W} - E(\mathbf{W})\|_\infty$. Note that the kernel function of $\mathbf{W}$ is bounded, i.e., $\|c_3 y_{i\backslash j}^2 \boldsymbol{x}_{i\backslash j}^{\otimes 2} I(F_{ij})\|_\infty \leq 16 c_3 M^2 \tau^2$. Then the Hoeffding's inequality can be applied to the centered U-statistics $W_{jk} - E(W_{jk})$. For some constant $t > 0$ to be chosen, there exist some universal constants c_4, $c_5 > 0$, such that

$$\begin{aligned} \Pr\left(s\|\mathbf{W} - E(\mathbf{W})\|_\infty > t\right) &\leq \sum_{j,k} \Pr\left(|W_{jk} - E(W_{jk})| > \frac{t}{s}\right) \\ &\leq c_4 p^2 \exp\left(-\frac{c_5 t^2 n}{s^2}\right). \end{aligned}$$

If Y follows the normal linear model, without loss of generality, we assume $Y|\boldsymbol{X} \sim N(\alpha + \beta^T \boldsymbol{X}, \phi)$, then

$$\begin{aligned} & E\left(y_{i\backslash j}^2 I(F_{ij}) | \boldsymbol{x}_i, \boldsymbol{x}_j\right) \\ = & \frac{1}{\sqrt{2\pi}} \int_{-\tau}^{\tau}\int_{-\tau}^{\tau} y_{i\backslash j}^2 \exp\left\{-\frac{(y_i - \alpha - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + (y_j - \alpha - \boldsymbol{x}_j^T\boldsymbol{\beta})^2}{2\phi}\right\} dy_i dy_j \\ \geq & \frac{1}{\sqrt{2\pi}} \int_{-\tau}^{\tau}\int_{-\tau}^{\tau} y_{i\backslash j}^2 \exp\left\{-\frac{y_i^2 + y_j^2 + 2B^2 + 2B|y_i| + 2B|y_j|}{2\phi}\right\} dy_i dy_j \triangleq c_6. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \boldsymbol{v}^T E(\mathbf{W})\boldsymbol{v} &= \boldsymbol{v}^T E(E(\mathbf{W}|\boldsymbol{x}))\boldsymbol{v} \geq c_6 \boldsymbol{v}^T E\boldsymbol{x}_{i\backslash j}^{\otimes 2}\boldsymbol{v} \\ &= 2c_6 \boldsymbol{v}^T E(\boldsymbol{x}_i \boldsymbol{x}_i^T)\boldsymbol{v} \geq 2c_6 \lambda_{\min}(\Sigma_x) \end{aligned}$$

and hence, $\rho_-(E(\mathbf{W}), s) \geq 2c_6\lambda_{\min}(\Sigma_x)$, where $\Sigma_x = \mathrm{Cov}(\boldsymbol{X})$. By the Hoeffding equality, taking $t = c_6\lambda_{\min}(\Sigma_x)$ we have

$$\rho_-(s) \geq \rho_-(\mathbf{W}, s) \geq c_6\lambda_{\min}(\Sigma_x),$$

with probability at least $1 - c_4p^2\exp(-c_5c_6^2\lambda_{\min}^2(\Sigma_x)n/s^2)$.

For the upper bound of $\rho_+(s)$, notice that

$$\nabla^2\mathcal{L}(\boldsymbol{\gamma}^*) \leq \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} y_{i\backslash j}^2\boldsymbol{x}_{i\backslash j}^{\otimes 2} \triangleq \mathbf{W}'$$

Similar as before, we have

$$\rho_+(s) \leq \rho_+(\mathbf{W}', s) \leq \rho_+(E(\mathbf{W}'), s) + s\|\mathbf{W}' - E(\mathbf{W}')\|_\infty$$

If $Y|\boldsymbol{X} \sim N(\alpha + \beta^T\boldsymbol{X}, \phi)$, we have

$$E(y_{i\backslash j}^2|\boldsymbol{x}_i, \boldsymbol{x}_j) = 2\phi + (\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\beta})^2$$

and hence

$$\begin{aligned}
\rho_+(E(\mathbf{W}'), s) &\leq E(2\phi(\boldsymbol{x}_{i\backslash j}^T\boldsymbol{x}_{i\backslash j})^2) + E\{(\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\beta})^2(\boldsymbol{x}_{i\backslash j}^T\boldsymbol{v})^2\} \\
&\leq 4\phi\lambda_{\max}(\Sigma_x) + \frac{1}{2}E(\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\beta})^4 + \frac{1}{2}E(\boldsymbol{x}_{i\backslash j}^T\boldsymbol{v})^4 \\
&\leq 4\phi\lambda_{\max}(\Sigma_x) + 16B^4 + 16M^4.
\end{aligned}$$

Following the similar argument as above, we have

$$\rho_+(s) \leq \rho_+(\mathbf{W}', s) \leq t + 4\phi\lambda_{\max}(\Sigma_x) + 16B^4 + 16M^4$$

with probability at least $1 - c_1 p^2 \exp(-c_2 t^2 n/s^2)$, for any constant $t > 0$. For simplicity, after taking $t = c_6 \lambda_{\min}(\Sigma_x)$, we have

$$\rho_+(s) \leq \rho_+(\mathbf{W}', s) \leq c_6 \lambda_{\min}(\Sigma_x) + 4\phi \lambda_{\max}(\Sigma_x) + 16B^4 + 16M^4$$

with probability at least $1 - c_4 p^2 \exp(-c_5 c_6^2 \lambda_{\min}^2(\Sigma_x) n/s^2)$. The choices of ρ_* and ρ^* can be decided accordingly. When $Y|\boldsymbol{X}$ follows a logistic regression, based on the arguments in the proof of Theorem 3.10 in Ning et al. (2017) and the above steps, the same conclusion follows. This completes the verification by taking $C_1 = 2c_4$ and $C_2 = c_5 c_6^2 \lambda_{\min}^2(\Sigma_x)$.

□

Proof of Lemma 3. First, by Lemma 1,

$$\nabla \mathcal{L}(\boldsymbol{\gamma}^*) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \{\psi'(y_{i\backslash j} \boldsymbol{x}_{i\backslash j}^T \boldsymbol{\gamma}^*) y_{i\backslash j} \boldsymbol{x}_{i\backslash j} - y_{i\backslash j} \boldsymbol{x}_{i\backslash j}\}$$

is a mean-zero U-statistic of order 2. Given the Assumption 1, we have

$$\|\{\psi'(y_{i\backslash j} \boldsymbol{x}_{i\backslash j}^T \boldsymbol{\gamma}^*) y_{i\backslash j} \boldsymbol{x}_{i\backslash j} - y_{i\backslash j} \boldsymbol{x}_{i\backslash j}\}\|_\infty \leq 2M |y_{i\backslash j}|.$$

By the sub-exponential tail condition on y_i, for any $x > 0$ and $u = 1, \ldots, p$,

$$\begin{aligned}
& \Pr(|\{\psi'(y_{i\backslash j} \boldsymbol{x}_{i\backslash j}^T \boldsymbol{\gamma}^*) y_{i\backslash j} \boldsymbol{x}_{i\backslash j} - y_{i\backslash j} \boldsymbol{x}_{i\backslash j}\}_u| > x) \\
\leq \; & \Pr(|y_{i\backslash j}| > x/(2M)) \\
\leq \; & \Pr(|y_i| > x/(4M)) + \Pr(|y_i| > x/(4M)) \\
\leq \; & 2c_1 \exp\{-c_2 x/(4M)\}.
\end{aligned}$$

By Lemma 2 with $k = \lfloor n/2 \rfloor$, we have

$$\begin{aligned} & \Pr(\|\nabla \mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty > C_3\sqrt{\log p/n}) \leq \sum_{u=1}^{p} \Pr(|\nabla_u \mathcal{L}(\boldsymbol{\gamma}^*)| > C_3\sqrt{\log p/n}) \\ \leq \; & 2p \exp\left[-\min\left\{ \frac{c_2^2 C_3^2 k \log p}{2^9 c_1^2 M^2 n}, \frac{c_2 C_3 k (\log p)^{1/2}}{2^5 c_1 M n^{1/2}} \right\} \right], \end{aligned}$$

which completes the proof by defining $C_4 = \frac{c_2^2 C_3^2}{3 \cdot 2^9 c_1^2 M^2}$ and $C_5 = \frac{c_2 C_3}{3 \cdot 2^5 c_1 M}$, where we use the fact that $k/n > 1/3$. □

Proof of Lemma 4. We restrict all vectors on S in this proof. For the sake of easy presentation, the subscript S is omitted throughout. From the Taylor's expansion, we have

$$\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \boldsymbol{\gamma}^* = -\{\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}}_1)\}^{-1} \nabla \mathcal{L}(\boldsymbol{\gamma}^*),$$

where $\widetilde{\boldsymbol{\gamma}}_1 = \boldsymbol{\gamma}^* + t_1(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \boldsymbol{\gamma}^*)$, $0 \leq t_1 \leq 1$. Therefore

$$\|\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \boldsymbol{\gamma}^*\|_\infty \leq \|\{\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}}_1)\}^{-1}\|_{L_\infty} \|\nabla \mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty.$$

For $\|\nabla \mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty$, based on the proof of Lemma 3, we have $\|\nabla \mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty \leq C_3\sqrt{\frac{\log s^*}{n}}$ with probability at least $1 - 2s^* \exp\left[-\min\left\{ C_4 \log s^*, C_5 n^{1/2} (\log s^*)^{1/2} \right\} \right]$.

For $\|\{\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}}_1)\}^{-1}\|_{L_\infty}$, following the similar argument in Ning et al. (2017), we have $\|\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \boldsymbol{\gamma}^*\|_1 \leq c_{11} s^* \sqrt{\frac{\log s^*}{n}}$ with probability at least $1 - c_{12} p^{-1}$ and

$$\|\{\nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)\}^{-1} \{\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}}_1) - \nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)\}\|_{L_\infty} \leq s^* \min\{e^b - 1, 1 - e^{-b}\},$$

where

$$\begin{aligned} b &= \max_{i,j} |y_{i\setminus j} \boldsymbol{x}_{i\setminus j}^T (\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}^*)| \\ &\leq \max_{i,j} |Y_i - Y_j| \|\boldsymbol{X}_i - \boldsymbol{X}_j\|_\infty \|\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}^*\|_1 \\ &\leq 12 c_2^{-1} \log(n) M c_{11} s^* \sqrt{\frac{\log s^*}{n}} \end{aligned}$$

with probability at least $1 - c_1 n^{-1} - c_{12} p^{-1}$ by taking $\delta = 3c_2^{-1}\log(n)$ defined in Assumption 1. Therefore, $\|\{\nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)\}^{-1}\{\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}}_1) - \nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)\}\|_{L_\infty}$ is bounded by a term with the order of $\log(n)(s^*)^2\sqrt{\frac{\log s^*}{n}} = o_p(1)$ with a high probability. Then we can choose a sufficiently large n, such that $\|\{\nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)\}^{-1}\{\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}}_1) - \nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)\}\|_{L_\infty} \leq 1/2$. Then based on the Theorem 2.3.4 in Golub and Van Loan (1996), we have

$$\|\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}}_1)^{-1}\|_{L_\infty} \leq \frac{\|\nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)^{-1}\|_{L_\infty}}{1 - \|\{\nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)\}^{-1}\{\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}}_1) - \nabla^2 \mathcal{L}(\boldsymbol{\gamma}^*)\}\|_{L_\infty}} < 2C,$$

and this completes the proof. □

Proof of Lemma 5. According to the Assumption 3, since $q_\lambda(t)$ satisfies the Lipschitz continuity condition, we have

$$-\zeta_- \|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|^2 \leq (q'_\lambda(\boldsymbol{\gamma}_2) - q'_\lambda(\boldsymbol{\gamma}_1))^T (\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) \leq -\zeta_+ \|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|^2,$$

which implies that the convex function $-\mathcal{Q}(\boldsymbol{\gamma})$ satisfies

$$(\nabla(-\mathcal{Q}_\lambda(\boldsymbol{\gamma}_2)) - \nabla(-\mathcal{Q}_\lambda(\boldsymbol{\gamma}_1)))^T (\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) \leq \zeta_- \|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2,$$

and

$$(\nabla(-\mathcal{Q}_\lambda(\boldsymbol{\gamma}_2)) - \nabla(-\mathcal{Q}_\lambda(\boldsymbol{\gamma}_1)))^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) \geq \zeta_+\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2.$$

According to Theorem 2.1.5 and Theorem 2.1.9 in Nesterov (2013), the above two expressions are equivalent definitions of strong smoothness and strong convexity respectively. In other words, $-\mathcal{Q}_\lambda(\boldsymbol{\gamma})$ satisfies

$$-\mathcal{Q}_\lambda(\boldsymbol{\gamma}_2) \leq -\mathcal{Q}_\lambda(\boldsymbol{\gamma}_1) - \nabla\mathcal{Q}(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) + \frac{\zeta_-}{2}\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2,$$

and

$$-\mathcal{Q}_\lambda(\boldsymbol{\gamma}_2) \geq -\mathcal{Q}_\lambda(\boldsymbol{\gamma}_1) - \nabla\mathcal{Q}(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) + \frac{\zeta_+}{2}\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2.$$

For our loss function $\mathcal{L}(\boldsymbol{\gamma})$, by Taylor's expansion and the mean value theorem, we have

$$\mathcal{L}(\boldsymbol{\gamma}_2) = \mathcal{L}(\boldsymbol{\gamma}_1)+\nabla\mathcal{L}(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2-\boldsymbol{\gamma}_1)+\frac{1}{2}(\boldsymbol{\gamma}_2-\boldsymbol{\gamma}_1)^T\nabla^2\mathcal{L}(t\boldsymbol{\gamma}_1+(1-t)\boldsymbol{\gamma}_2)(\boldsymbol{\gamma}_2-\boldsymbol{\gamma}_1),$$

where $0 \leq t \leq 1$. Since we assume $\|(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1)_{\bar{S}}\|_0 \leq s^*$, which implies $\|\boldsymbol{\gamma}_2-\boldsymbol{\gamma}_1\|_0 \leq 2s^*$. Therefore, by the definition of sparse eigenvalue, we have

$$\frac{(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1)^T}{\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2}\nabla^2\mathcal{L}(t\boldsymbol{\gamma}_1 + (1-t)\boldsymbol{\gamma}_2)\frac{(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1)}{\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2} \in [\rho_-(\nabla^2\mathcal{L}, 2s^*), \rho_+(\nabla^2\mathcal{L}, 2s^*)].$$

Plugging this into the RHS of the Taylor expansion, we have

$$\mathcal{L}(\boldsymbol{\gamma}_2) \geq \mathcal{L}(\boldsymbol{\gamma}_1) + \nabla\mathcal{L}(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) + \frac{\rho_-(\nabla^2\mathcal{L}, 2s^*)}{2}\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2,$$

and

$$\mathcal{L}(\boldsymbol{\gamma}_2) \leq \mathcal{L}(\boldsymbol{\gamma}_1) + \nabla\mathcal{L}(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) + \frac{\rho_+(\nabla^2\mathcal{L}, 2s^*)}{2}\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2.$$

Putting all of the above four inequalities together, we have

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_2) \geq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_1) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) + \frac{\rho_-(\nabla^2\mathcal{L}, 2s^*) - \zeta_-}{2}\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2,$$

and

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_2) \leq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_1) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) + \frac{\rho_+(\nabla^2\mathcal{L}, 2s^*) - \zeta_+}{2}\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2.$$

□

Proof of Theorem 1. From the Karush-Kuhn-Tucker condition, we have

$$\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}) + \lambda\widehat{\boldsymbol{\xi}} = 0,$$

where $\widehat{\boldsymbol{\xi}} \in \partial\|\widehat{\boldsymbol{\gamma}}\|_1$ represents the subgradient, i.e., $\hat{\xi}_j = \text{sign}(\hat{\gamma}_j)$, if $\hat{\gamma}_j \neq 0$; $\hat{\xi}_j \in [-1, 1]$ if $\hat{\gamma}_j = 0$. Next, we show that, there exists some $\boldsymbol{\xi}_{\text{O}} \in \partial\|\widehat{\boldsymbol{\gamma}}_{\text{O}}\|_1$, such that $\widehat{\boldsymbol{\gamma}}_{\text{O}}$ satisfies the exactly same condition as above

$$\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}_{\text{O}}) + \lambda\boldsymbol{\xi}_{\text{O}} = 0.$$

For $j \in S$, by the condition of the weakest signal strength and the result of Lemma 4, with probability at least $1 - \delta_2$, when n is sufficiently large,

$$|(\widehat{\boldsymbol{\gamma}}_{\text{O}})_j| \geq |\gamma_j^*| - \|\widehat{\boldsymbol{\gamma}}_{\text{O}} - \boldsymbol{\gamma}^*\|_\infty \geq 2\nu - 2CC_3\sqrt{\log s^*/n} > \nu, \qquad \text{(S1.1)}$$

then by the condition of the penalty function, we have

$$(\nabla\mathcal{Q}_\lambda(\widehat{\boldsymbol{\gamma}}_{\text{O}}) + \lambda\boldsymbol{\xi}_{\text{O}})_j = (\nabla\mathcal{P}_\lambda(\widehat{\boldsymbol{\gamma}}_{\text{O}}))_j = p'_\lambda((\widehat{\boldsymbol{\gamma}}_{\text{O}})_j) = 0.$$

For $j \in \bar{S}$, $(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}})_j = 0$, so $(\nabla \mathcal{Q}_\lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}))_j = 0$, therefore

$$(\nabla \widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) + \lambda \boldsymbol{\xi}_{\mathrm{O}})_j = (\nabla \mathcal{L}(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) + \lambda \boldsymbol{\xi}_{\mathrm{O}})_j,$$

so we can define $(\boldsymbol{\xi}_{\mathrm{O}})_j = (-\frac{\nabla \mathcal{L}(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}})}{\lambda})_j$. Note that we choose $\lambda \asymp \sqrt{\log p/n}$, and from the proof of Lemma 3, with probability at least $1 - \delta_1$,

$$\|\nabla \mathcal{L}(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}})\|_\infty \leq C_3 \sqrt{\log p/n}. \tag{S1.2}$$

So we have $\boldsymbol{\xi}_{\mathrm{O}} \in [-1, 1]$, and therefore we've found $\boldsymbol{\xi}_{\mathrm{O}}$, such that $\boldsymbol{\xi}_{\mathrm{O}} \in \partial \|\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}\|_1$, and $\nabla \widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) + \lambda \boldsymbol{\xi}_{\mathrm{O}} = 0$, with probability at least $1 - \delta_1 - \delta_2$, by (S1.1), (S1.2) and the fact that $P(A \bigcap B) \geq P(A) + P(B) - 1$, where A and B are two arbitrary events.

Next, we show that $\|(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}_{\mathrm{O}})_{\bar{S}}\|_0 \leq s^*$. Due to the analysis of the convergence properties based on the MM algorithm, presented in Zou and Li (2008), we only need to prove this result in the l-th iteration, i.e., for $\widehat{\boldsymbol{\gamma}}^{(l)}$. In the l-th iteration, we define $G^{(l)} = \{k : \gamma_k^* = 0, \hat{\omega}_k^{(l-1)} \geq p'_\lambda(c_8\lambda), k = 1, \ldots, p\}$, representing the covariates who are unimportant but heavily penalized. Its complement $\overline{G^{(l)}} = \{k : \gamma_k^* \neq 0, \text{ or } \hat{\omega}_k^{(l-1)} < p'_\lambda(c_8\lambda), k = 1, \ldots, p\}$. It's clear that $S \subset \overline{G^{(l)}}$. If we define $H := \overline{G^{(l)}} - S = \{k : \gamma_k^* = 0, \hat{\omega}_k^{(l-1)} < p'_\lambda(c_8\lambda), k = 1, \ldots, p\}$, it's also clear that S and H are disjoint. We are going to first show that $|\overline{G^{(l)}}| \leq 2s^*$ by induction.

For $l = 1$, because we have $\hat{\omega}_k^{(0)} = \lambda$, $\overline{G^{(1)}} = S$, hence $|\overline{G^{(1)}}| \leq s^*$. Now

we assume that $|\overline{G^{(l)}}| \leq 2s^*$ for some integer l and our goal is to prove that $|\overline{G^{(l+1)}}| \leq 2s^*$.

Suppose $\widehat{\boldsymbol{\gamma}}^{(l)}$ is the solution in the l-th iteration, from the Karush-Kuhn-Tucker condition, we have

$$\nabla \mathcal{L}(\widehat{\boldsymbol{\gamma}}^{(l)}) + \widehat{\boldsymbol{\omega}}^{(l-1)} \circ \boldsymbol{\xi}^{(l)} = 0,$$

where $\boldsymbol{\xi}^{(l)} \in \partial \|\widehat{\boldsymbol{\gamma}}^{(l)}\|_1$. In the following, we denote $\boldsymbol{\delta} = \widehat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma}^*$. By the mean value theorem, we have

$$\nabla \mathcal{L}(\widehat{\boldsymbol{\gamma}}^{(l)}) - \nabla \mathcal{L}(\boldsymbol{\gamma}^*) = \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}})\boldsymbol{\delta},$$

where $\widetilde{\boldsymbol{\gamma}} = t\boldsymbol{\gamma}^* + (1-t)\widehat{\boldsymbol{\gamma}}^{(l)}$, which implies

$$0 \leq \boldsymbol{\delta}^T \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\gamma}})\boldsymbol{\delta} = -\boldsymbol{\delta}^T \widehat{\boldsymbol{\omega}}^{(l-1)} \circ \boldsymbol{\xi}^{(l)} - \nabla \mathcal{L}(\boldsymbol{\gamma}^*)^T \boldsymbol{\delta}.$$

For the second term, Holder's inequality implies

$$\nabla \mathcal{L}(\boldsymbol{\gamma}^*)^T \boldsymbol{\delta} \geq -\|\nabla \mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty \|\boldsymbol{\delta}\|_1.$$

For the first term, also use Holder's inequality, we have

$$\begin{aligned} \boldsymbol{\delta}^T(\widehat{\boldsymbol{\omega}}^{(l-1)} \circ \boldsymbol{\xi}^{(l)}) &= \boldsymbol{\delta}_S^T(\widehat{\boldsymbol{\omega}}^{(l-1)} \circ \boldsymbol{\xi}^{(l)})_S + |\boldsymbol{\delta}_H^T \widehat{\boldsymbol{\omega}}_H^{(l-1)}| + |\boldsymbol{\delta}_G^T \widehat{\boldsymbol{\omega}}_G^{(l-1)}| \\ &\geq -\|\boldsymbol{\delta}_S\|_1 \|\widehat{\boldsymbol{\omega}}_S^{(l-1)}\|_\infty + \|\boldsymbol{\delta}_H\|_1 \|\widehat{\boldsymbol{\omega}}_H^{(l-1)}\|_{\min} + \|\boldsymbol{\delta}_G\|_1 \|\widehat{\boldsymbol{\omega}}_G^{(l-1)}\|_{\min}. \end{aligned}$$

Combining these two inequalities, we have

$$-\|\boldsymbol{\delta}_S\|_1 \|\widehat{\boldsymbol{\omega}}_S^{(l-1)}\|_\infty + \|\boldsymbol{\delta}_H\|_1 \|\widehat{\boldsymbol{\omega}}_H^{(l-1)}\|_{\min} + \|\boldsymbol{\delta}_G\|_1 \|\widehat{\boldsymbol{\omega}}_G^{(l-1)}\|_{\min} - \|\nabla \mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty \|\boldsymbol{\delta}\|_1 \leq 0.$$

Hence

$$p'_\lambda(c_8\lambda)\|\boldsymbol{\delta}_G\|_1 \leq \|\boldsymbol{\delta}_G\|_1\|\widehat{\boldsymbol{\omega}}_G^{(l-1)}\|_{\min} \leq \|\nabla\mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty\|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}_S\|_1\|\widehat{\boldsymbol{\omega}}_S^{(l-1)}\|_\infty.$$

Therefore, we have

$$\left[p'_\lambda(c_8\lambda) - \|\nabla\mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty\right]\|\boldsymbol{\delta}_G\|_1 \leq \left[\|\widehat{\boldsymbol{\omega}}_S^{(l-1)}\|_\infty + \|\nabla\mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty\right]\|\boldsymbol{\delta}_{\bar{G}}\|_1,$$

which implies

$$\|\boldsymbol{\delta}_G\|_1 \leq \frac{\|\widehat{\boldsymbol{\omega}}_S^{(l-1)}\|_\infty + \|\nabla\mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty}{p'_\lambda(c_8\lambda) - \|\nabla\mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty}\|\boldsymbol{\delta}_{\bar{G}}\|_1 \leq c_{13}\|\boldsymbol{\delta}_{\bar{G}}\|_1,$$

which is equivalent to

$$\|\widehat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma}^*\|_1 \leq (1+c_{13})\|\widehat{\boldsymbol{\gamma}}^{(l)}_{G^{(l)}} - \boldsymbol{\gamma}^*_{G^{(l)}}\|_1.$$

Similarly, we can also show that

$$\|\widehat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma}^*\|_2 \leq (1+c_{13})\|\widehat{\boldsymbol{\gamma}}^{(l)}_{I^{(l)}} - \boldsymbol{\gamma}^*_{I^{(l)}}\|_2.$$

Next, following the proof of Lemma A.3 in Yang et al. (2014), based on the Assumption 2 and the condition that $s^*\sqrt{\frac{\log p}{n}} = o_p(1)$, with probability at least $1-\delta_3$, we can establish the following crude rates of convergence for $l \geq 1$:

$$\|\widehat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma}^*\|_2 \leq c_{14}\rho_*^{-1}\sqrt{s^*}\lambda. \tag{S1.3}$$

By the concavity of p_λ, for any $k \in A := \overline{G^{(l+1)}} - S$, we have $|\hat{\gamma}_k^{(l)}| \geq c_8\lambda$. Therefore we have

$$\sqrt{|A|} \leq \|\widehat{\boldsymbol{\gamma}}_A^{(l)}\|_2/(c_8\lambda) = \|\widehat{\boldsymbol{\gamma}}_A^{(l)} - \boldsymbol{\gamma}_A^*\|_2/(c_8\lambda) \leq c_{14}\rho_*^{-1}\sqrt{s^*}/c_8 \leq \sqrt{s^*},$$

where the first inequality follows from $|A| \leq \sum_{k\in A} |\hat{\gamma}_k^{(l)}|^2/(c_8\lambda)^2$, and the last inequality follows from the appropriate choice of c_{14} by the similar argument in Yang et al. (2014). Note that this implies that $|\overline{G^{(l+1)}}| \leq 2s^*$. Therefore, by induction, $|\overline{G^{(l)}}| \leq 2s^*$ for any $l \geq 1$. Then, from (S1.3) we can follow the similar arguments in Zhang (2013); Yang et al. (2014) to conclude that $\|(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}_{\mathrm{O}})_{\bar{S}}\|_0 \leq s^*$, with probability at least $1 - \delta_3$.

Next we are showing $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}_{\mathrm{O}}$ when n is sufficiently large. By Lemma 5, it yields

$$\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}) \geq \widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) + \nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}})^T(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) + \frac{\rho_-(\nabla^2\mathcal{L}, 2s^*) - \zeta_-}{2}\|\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}_{\mathrm{O}}\|_2^2,$$

and

$$\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) \geq \widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}) + \nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}})^T(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \widehat{\boldsymbol{\gamma}}) + \frac{\rho_-(\nabla^2\mathcal{L}, 2s^*) - \zeta_-}{2}\|\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \widehat{\boldsymbol{\gamma}}\|_2^2.$$

By the convexity of L_1 norm, we have

$$\lambda\|\widehat{\boldsymbol{\gamma}}\|_1 \geq \lambda\|\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}\|_1 + \lambda(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}_{\mathrm{O}})^T\boldsymbol{\xi}_{\mathrm{O}},$$

and

$$\lambda\|\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}\|_1 \geq \lambda\|\widehat{\boldsymbol{\gamma}}\|_1 + \lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \widehat{\boldsymbol{\gamma}})^T\widehat{\boldsymbol{\xi}}.$$

Adding the above four inequalities, we have

$$0 \geq (\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}) + \lambda\widehat{\boldsymbol{\xi}})^T(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \widehat{\boldsymbol{\gamma}}) + (\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) + \lambda\boldsymbol{\xi}_{\mathrm{O}})^T(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) + (\rho_-(\nabla^2\mathcal{L}, 2s^*) - \zeta_-)\|\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}_{\mathrm{O}}\|_2^2.$$

Since $\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}) + \lambda\widehat{\boldsymbol{\xi}} = 0$, $\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) + \lambda\boldsymbol{\xi}_{\mathrm{O}} = 0$, $\rho_-(\nabla^2\mathcal{L}, 2s^*) - \zeta_- > 0$, we must have $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}_{\mathrm{O}}$, i.e., we conclude that $\widehat{\boldsymbol{\gamma}}$ is the oracle estimator $\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}$. Also, since $\min_{j\in S}|(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}})_j| > 0$ and the fact that $\mathrm{supp}(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) \subset S$, we have

$$\mathrm{supp}(\widehat{\boldsymbol{\gamma}}) = \mathrm{supp}(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) = \mathrm{supp}(\boldsymbol{\gamma}^*),$$

with probability at least $1 - \delta_1 - \delta_2 - \delta_3$, where this high probability comes from (S1.1), (S1.2), (S1.3) in the process of this proof, and the fact that $P(A\bigcap B\bigcap C) \geq P(A) + P(B\bigcap C) - 1 \geq P(A) + P(B) + P(C) - 2$ where A, B and C are three arbitrary events, and this completes the proof. □

S2 More Simulation Studies

In general the assumption imposed on the missing data mechanism is unverifiable. Although the assumption (2.3) we discuss in this paper is already very flexible, it is still plausible to be violated in real applications. Therefore, in the next four simulations, we evaluate the robustness of our proposed method when the assumption (2.3) is slightly violated. The simulation settings (S5)–(S8) are as follows:

(S5): same as (S1) except that $\Pr(R = 1|Y, \boldsymbol{X}) = I_{\{Y+0.1X_3>\gamma_1\}}I_{\{X_1>\gamma_2\}}$.

(S6): same as (S2) except that $\Pr(R = 1|Y, \boldsymbol{X}) = I_{\{Y+0.1X_3>\gamma_1\}}I_{\{X_1>\gamma_2\}}$.

(S7): same as (S3) except that $\Pr(R = 1|Y, \boldsymbol{X}) = I_{\{X_1>\gamma\}}\cdot\left\{\frac{2Y+3}{5} - \frac{0.1(|X_3|\wedge 3)}{1+0.1(|X_3|\wedge 3)}\right\}$.

(S8): same as (S4) except that $\Pr(R=1|Y,\boldsymbol{X}) = I_{\{X_1>\gamma\}}\cdot\left\{\frac{2Y+3}{5} - \frac{0.1(|X_3|\wedge 3)}{1+0.1(|X_3|\wedge 3)}\right\}$.

Similar as before, we count the number of false positives (#FP) and the number of false negatives (#FN) and report them in a boxplot in each setting in Figures 1–4 respectively. We also list the mean and standard deviation (SD) of #FP and #FN for each setting in Tables 1–2. It can be seen that, although the assumption (2.3) is slightly violated, our proposed method still performs better than the one assuming MAR in many scenarios. This phenomenon shows that our proposed method possesses some robustness to the misspecification of the missing data mechanism assumption.

Finally, we provide some results on the computing time of our proposed method. We report the mean and standard deviation (SD) of the computing time for simulation settings (S1)–(S2) in Table 3. The simulations are conducted on an OS X system version 10.9.5 with 2.2 GHz Intel Core i7 CPU and 16GB memory. It's not surprising that our proposed method is more time-consuming than the others. This phenomenon is consistent with the theoretical implication. In theory, from the algorithms we developed in Section 3, the computing time of the proposed method is equivalent to solving a standard penalized logistic regression with sample size $n(n-1)/2$, while the computing time of the method assuming no missing data (or assuming

MAR) is the same as to solving a standard penalized logistic regression with sample size N (or n). Eventually we will make our algorithm publicly available by creating an R package with some core part implemented by C.

S3 Real Data Analyses

In this Section, we present two data analyses to demonstrate the usefulness of our proposed method in real applications. The first study concerns the melanoma cancer through the observation-controlled Eastern Cooperative Oncology Group (ECOG) phase III clinical trial E1684. The second study (GEO GDS3289) investigates the association between prostate cancer tumors and genomic biomarkers, sponsored by the US National Institutes of Health.

S3.1 Melanoma Study

Melanoma is the most dangerous type of skin cancer and its incidence is increasing at a rate that exceeds all solid tumors. Although education efforts have resulted in earlier detection of melanoma, high-risk melanoma patients continue to have high relapse and mortality rate of 50% or higher. Several post-operative (adjuvant) chemotherapies have been proposed for this class of melanoma patients, and the one which seems to provide the most

significant impact on relapse-free survival and survival is Interferon Alpha-2b (IFN). This immunotherapy was evaluated in E1684, an observation-controlled Eastern Cooperative Oncology Group (ECOG) phase III clinical trial Kirkwood et al. (1996).

In this trial, there are in total $N = 286$ patients and all the patients were randomized to one of two treatment trials: high dose interferon or observation. In this analysis, the outcome variable Y, was taken to be binary, and was assigned a 1 if the patient had an overall survival time greater than or equal to 0.55 years, and 0 otherwise. There are several prognostic factors that were identified as potentially important predictors: X_1, treatment (two levels); X_2, age (in years); X_3, nodes1 (four levels); X_4, sex (two levels); X_5, perform (two levels); and X_6, logarithm of Breslow thickness (in mm). Among all six covariates, X_3 and X_6 have missing values and the total number of completely observed samples is $n = 234$. The data set is available from Ibrahim et al. (2001).

To illustrate the proposed method, we assume that the original data set fits into a logistic regression and we minimize the penalized pairwise pseudo likelihood (2.8) to obtain the estimates. In contrast, under the MAR assumption, the corresponding estimates can be calculated by a penalized logistic regression with the completely observed subjects. We examine both

methods using three penalty functions: LASSO, SCAD and MCP. The variable selection and parameter estimation results are reported in Table 4.

The comparison of the results shown by both methods is as follows. Variables sex, perform, log(Breslow) are never selected by any method or any penalty, showing some agreement of the two methods. However, variable age is selected by the proposed method but not the method assuming MAR; variable nodes1 is selected by either method and either penalty, but the proposed method always show an elevation of the parameter estimate; the selection of the variable treatment depends on the method and the penalty.

A similar data set was previously analyzed in Ibrahim et al. (2001) and Garcia et al. (2010), and the latter showed that, both variable age and variable treatment can be selected by the adaptive LASSO method but not by the SCAD method. Variable age is negatively associated with a longer survival time, and its effect is not significant in the maximum likelihood estimate (MLE) method; while variable treatment is positively associated with a longer survival time, and its effect is significant according to the MLE. Both these agreements and disagreements of these methods reveal some more information that is contained in the data but cannot be disclosed if only one single method is explored. This could certainly provide

more insight of the data to investigators and clinicians.

S3.2 Prostate Cancer Study

We also analyze a data set from a study (GEO GDS3289) investigating the association between prostate cancer tumors and genomic biomarkers (Tomlins et al., 2007). The whole data set can be accessed from the website of the National Center for Biotechnology Information of the National Institutes of Health. Briefly, this data set contains $N = 104$ samples, out of which 34 are benign epithelium samples ($Y = 0$) and 70 non-benign samples ($Y = 1$).

There are missing values for various biomarkers in this data set. In our analysis, we include $p = 64$ biomarkers in total and six of them have missing values with the number of missing samples for each biomarker ranging from 1 to 53. The missing values result in a complete data set with the sample size $n = 49$, and there are 36 non-benign samples in this complete data set. We adopt the penalized logistic regression in this analysis, and we examine the results under two different assumptions: one assuming MAR, and the other assuming (2.3), with three different representative penalty functions: LASSO, SCAD and MCP. Similar to the previous data analysis, the variable selection and the parameter estimation results are reported in Table 5.

Our major findings and the comparison with previous literature can be summarized as follows. First, some biomarkers like RHOB, can be selected by either method or either penalty function. Second, some other biomarkers, for example, MME, ANXA1, CLDN4 and SOX4 can be selected by our proposed method but not the method assuming MAR. Interestingly, they were all investigated in the previous literature Kälin et al. (2011); Geary et al. (2014); Maeda et al. (2012); Wang et al. (2013) and clinically concluded to be associated with the prostate cancer. Although we cannot reach a uniform conclusion that our method outperforms the MAR method in this real data exploration, the analysis demonstrates that it can reveal some extra genetic information by using our proposed method. This illustrates the potential usefulness of our proposed method and it will be very interesting to medical investigators and clinical practitioners.

Bibliography

Garcia, R. I., J. G. Ibrahim, and H. Zhu (2010). Variable selection for regression models with missing data. *Statistica Sinica 20*(1), 149–165.

Geary, L. A., K. A. Nash, H. Adisetiyo, M. Liang, C.-P. Liao, J. H. Jeong, E. Zandi, and P. Roy-Burman (2014). Caf-secreted annexin a1 induces

prostate cancer cells to gain stem cell–like features. *Molecular Cancer Research 12*(4), 607–621.

Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2001). *Bayesian Survival Analysis*. Springer.

Kälin, M. et al. (2011). Novel prognostic markers in the serum of patients with castration-resistant prostate cancer derived from quantitative analysis of the pten conditional knockout mouse proteome. *European Urology 60*(6), 1235–1243.

Kirkwood, J. M., M. H. Strawderman, M. S. Ernstoff, T. J. Smith, E. C. Borden, and R. H. Blum (1996). Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the eastern cooperative oncology group trial est 1684. *Journal of Clinical Oncology 14*(1), 7–17.

Maeda, T., M. Murata, H. Chiba, A. Takasawa, S. Tanaka, T. Kojima, N. Masumori, T. Tsukamoto, and N. Sawada (2012). Claudin-4-targeted therapy using clostridium perfringens enterotoxin for prostate cancer. *The Prostate 72*(4), 351–360.

Nesterov, Y. (2013). *Introductory Lectures on Convex Optimization: A Basic Course*, Volume 87. Springer Science & Business Media.

Ning, Y., T. Zhao, and H. Liu (2017). A likelihood ratio framework for high-dimensional semiparametric regression. *The Annals of Statistics 45*(6), 2299–2327.

Tomlins, S. A. et al. (2007). Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics 39*(1), 41–51.

Wang, L., J. Zhang, X. Yang, Y. Chang, M. Qi, Z. Zhou, and B. Han (2013). Sox4 is associated with poor prognosis in prostate cancer and promotes epithelial–mesenchymal transition in vitro. *Prostate Cancer and Prostatic Diseases 16*(4), 301–307.

Yang, Z., Y. Ning, and H. Liu (2014). On semiparametric exponential family graphical models. *arXiv preprint arXiv:1412.8697*.

Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli 19*(5B), 2277–2293.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*(4), 1509–1566.

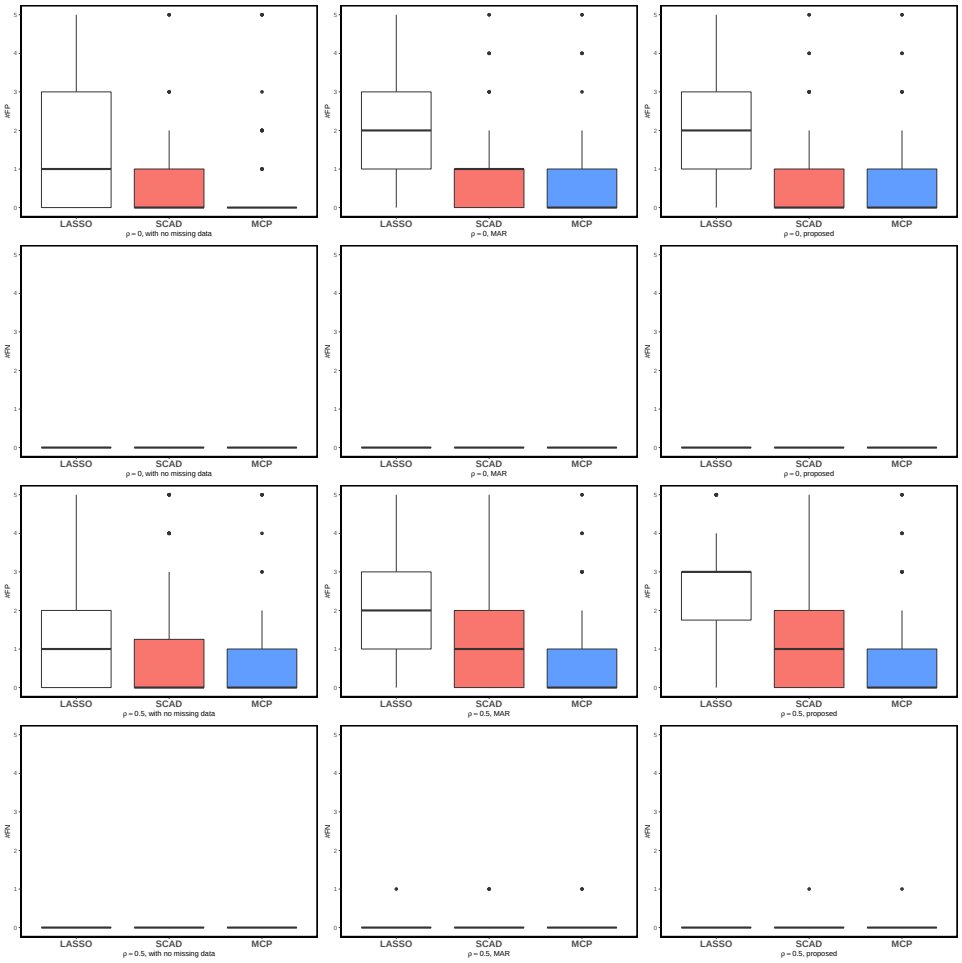

Figure 1: Boxplots of #FP and #FN in simulation setting (S5). The three columns represent the methods with no missing data, MAR and proposed, respectively. The first and third rows show #FP while the second and fourth rows show #FN. The first two rows are for the case with $\rho = 0$ and the last two rows are for the case with $\rho = 0.5$.

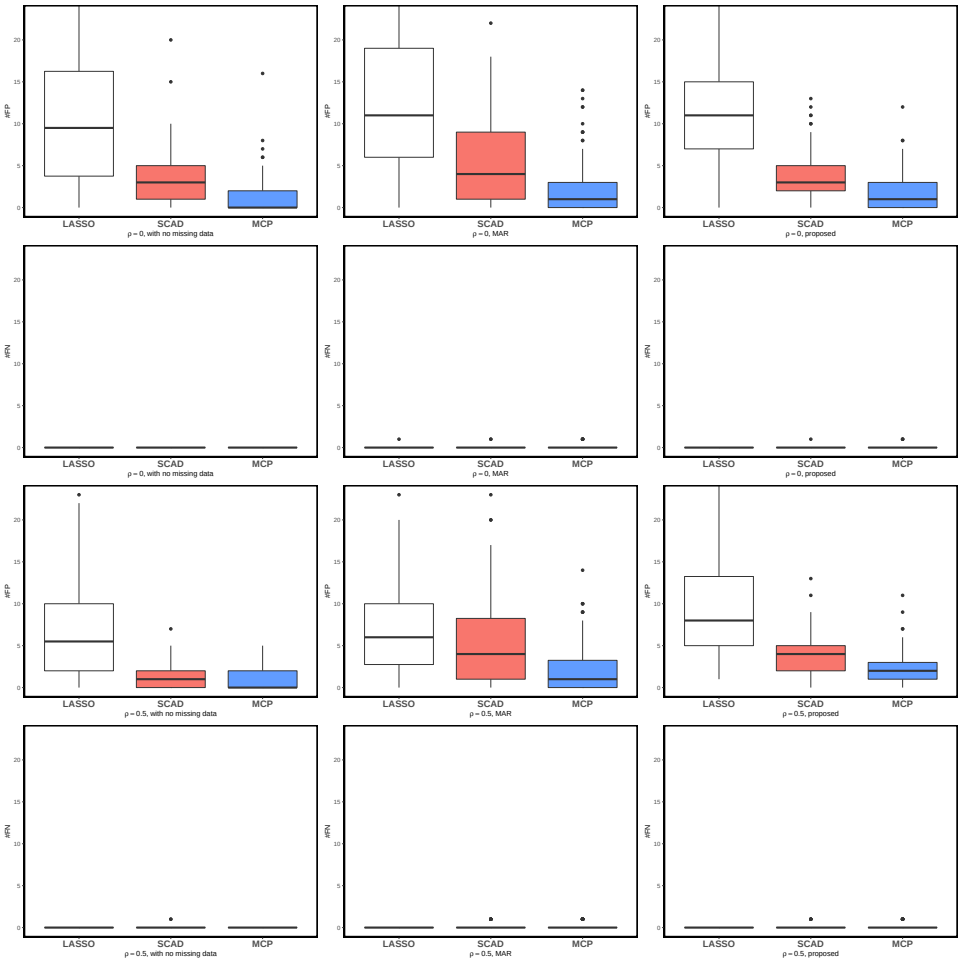

Figure 2: Boxplots of #FP and #FN in simulation setting (S6). The three columns represent the methods with no missing data, MAR and proposed, respectively. The first and third rows show #FP while the second and fourth rows show #FN. The first two rows are for the case with $\rho = 0$ and the last two rows are for the case with $\rho = 0.5$.

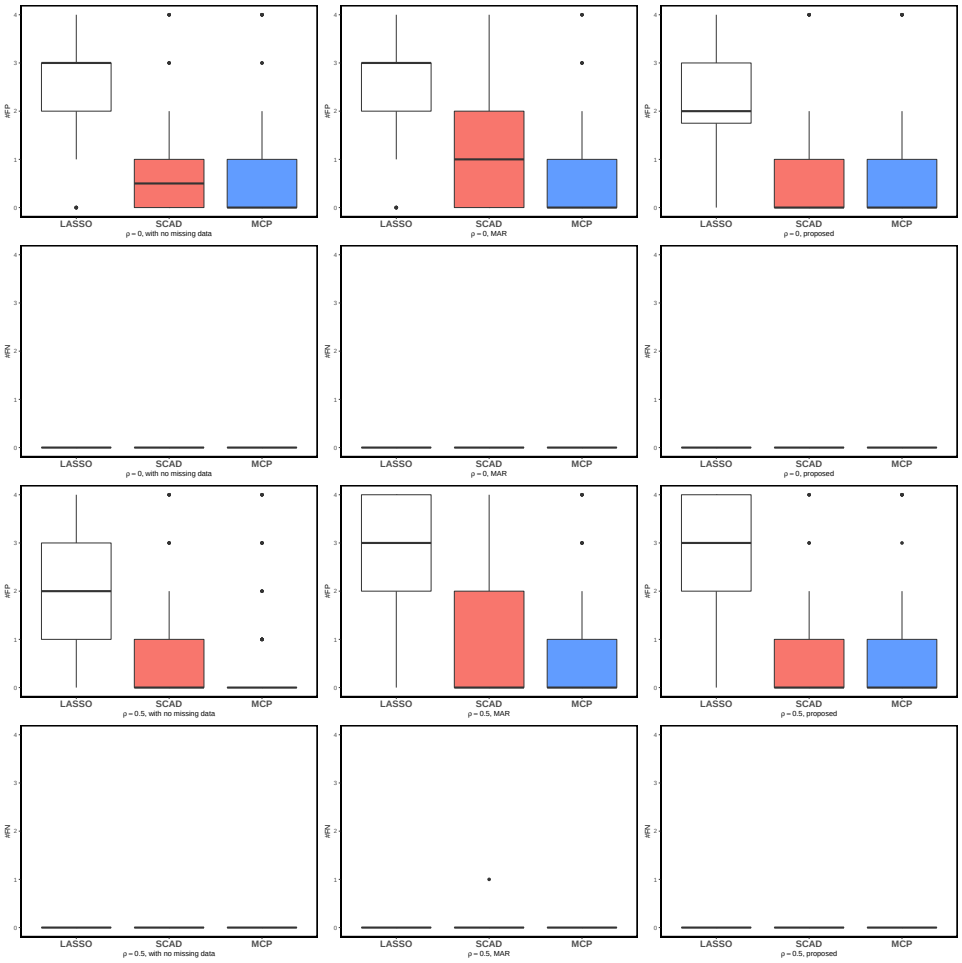

Figure 3: Boxplots of #FP and #FN in simulation setting (S7). The three columns represent the methods with no missing data, MAR and proposed, respectively. The first and third rows show #FP while the second and fourth rows show #FN. The first two rows are for the case with $\rho = 0$ and the last two rows are for the case with $\rho = 0.5$.

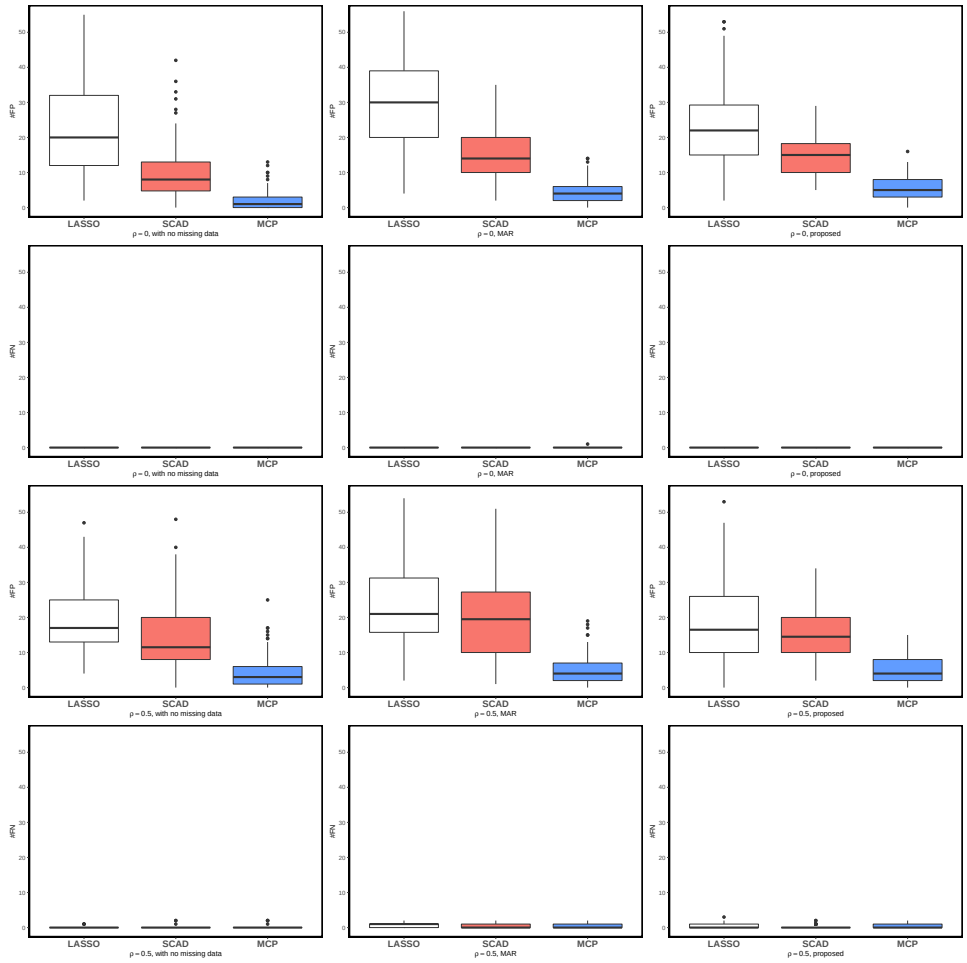

Figure 4: Boxplots of #FP and #FN in simulation setting (S8). The three columns represent the methods with no missing data, MAR and proposed, respectively. The first and third rows show #FP while the second and fourth rows show #FN. The first two rows are for the case with $\rho = 0$ and the last two rows are for the case with $\rho = 0.5$.

Table 1: Mean and standard deviation (SD; in parentheses) of #FP and #FN in simulation settings (S5)–(S6). The proposed method is compared to two other methods: the method with no missing data, which uses all simulated data; and the method assuming MAR, which uses completely observed samples only.

	Method	Penalty	$\rho = 0$		$\rho = 0.5$	
			#FP	#FN	#FP	#FN
p=8	with no missing data	LASSO	1.42 (1.44)	0 (0)	1.32 (1.48)	0 (0)
		SCAD	0.62 (1.10)	0 (0)	1.10 (1.62)	0 (0)
		MCP	0.46 (1.14)	0 (0)	0.65 (1.27)	0 (0)
	MAR	LASSO	2.19 (1.48)	0 (0)	1.90 (1.49)	0.01 (0.10)
		SCAD	0.87 (1.16)	0 (0)	1.15 (1.29)	0.03 (0.17)
		MCP	0.65 (1.13)	0 (0)	0.77 (1.18)	0.02 (0.14)
	proposed	LASSO	1.98 (1.21)	0 (0)	2.48 (1.38)	0 (0)
		SCAD	0.79 (1.17)	0 (0)	1.20 (1.40)	0.01 (0.10)
		MCP	0.58 (1.11)	0 (0)	0.81 (1.25)	0.01 (0.10)
p=200	with no missing data	LASSO	11.20 (9.45)	0 (0)	8.12 (8.68)	0 (0)
		SCAD	3.37 (3.41)	0 (0)	1.37 (1.59)	0.02 (0.14)
		MCP	1.37 (2.30)	0 (0)	1.01 (1.31)	0 (0)
	MAR	LASSO	14.00 (11.95)	0.01 (0.10)	7.91 (7.80)	0 (0)
		SCAD	5.35 (5.29)	0.02 (0.14)	5.72 (5.53)	0.10 (0.30)
		MCP	2.47 (3.41)	0.06 (0.24)	2.35 (2.92)	0.17 (0.38)
	proposed	LASSO	11.11 (5.96)	0 (0)	9.68 (6.59)	0 (0)
		SCAD	3.67 (2.85)	0.01 (0.10)	4.02 (2.27)	0.05 (0.22)
		MCP	2.10 (2.30)	0.03 (0.17)	2.46 (1.94)	0.10 (0.30)

Table 2: Mean and standard deviation (SD; in parentheses) of #FP and #FN in simulation settings (S7)–(S8). The proposed method is compared to two other methods: the method with no missing data, which uses all simulated data; and the method assuming MAR, which uses completely observed samples only.

	Method	Penalty	$\rho = 0$		$\rho = 0.5$	
			#FP	#FN	#FP	#FN
p=8	with no missing data	LASSO	2.51 (1.12)	0 (0)	2.35 (1.17)	0 (0)
		SCAD	0.79 (1.03)	0 (0)	0.51 (1.02)	0 (0)
		MCP	0.60 (1.10)	0 (0)	0.41 (1.02)	0 (0)
	MAR	LASSO	2.57 (1.09)	0 (0)	2.68 (1.03)	0 (0)
		SCAD	0.92 (1.04)	0 (0)	0.81 (1.18)	0.01 (0.10)
		MCP	0.64 (1.01)	0 (0)	0.61 (1.07)	0 (0))
	proposed	LASSO	2.21 (1.13)	0 (0)	2.54 (1.18)	0 (0)
		SCAD	0.59 (0.99)	0 (0)	0.59 (0.96)	0 (0)
		MCP	0.56 (1.04)	0 (0)	0.53 (1.03)	0 (0)
p=500	with no missing data	LASSO	22.97 (13.77)	0 (0)	19.54 (10.03)	0.04 (0.20)
		SCAD	9.64 (7.99)	0 (0)	14.72 (9.87)	0.05 (0.30)
		MCP	2.36 (2.77)	0 (0)	4.35 (4.56)	0.07 (0.36)
	MAR	LASSO	30.02 (12.19)	0 (0)	24.93 (14.94)	0.58 (0.59)
		SCAD	15.10 (7.15)	0 (0)	19.72 (11.47)	0.36 (0.56)
		MCP	4.46 (3.19)	0.01 (0.10)	4.83 (4.28)	0.55 (0.73)
	proposed	LASSO	23.37 (12.42)	0 (0)	19.27 (12.49)	0.52 (0.64)
		SCAD	14.66 (5.30)	0 (0)	15.40 (7.03)	0.19 (0.44)
		MCP	5.50 (3.32)	0 (0)	5.05 (3.77)	0.42 (0.67)

Table 3: Mean and standard deviation (SD; in parentheses) of computing time (in seconds) in simulation settings (S1)–(S2).

		Method	LASSO	SCAD	MCP
p=8	$\rho = 0$	with no missing data	0.04(0.00)	0.02(0.00)	0.02(0.00)
		MAR	0.04(0.01)	0.02(0.00)	0.02(0.00)
		proposed	0.45(0.06)	1.47(0.76)	1.05(0.15)
	$\rho = 0.5$	with no missing data	0.04(0.01)	0.02(0.00)	0.02(0.01)
		MAR	0.04(0.02)	0.02(0.00)	0.02(0.00)
		proposed	0.44(0.06)	1.53(0.62)	1.48(0.22)
p=200	$\rho = 0$	with no missing data	0.17(0.02)	0.05(0.00)	0.06(0.01)
		MAR	0.08(0.01)	0.05(0.01)	0.05(0.01)
		proposed	9.74(1.89)	47.71(9.24)	22.44(4.36)
	$\rho = 0.5$	with no missing data	0.17(0.02)	0.05(0.01)	0.05(0.01)
		MAR	0.07(0.01)	0.04(0.01)	0.05(0.01)
		proposed	6.91(1.41)	36.81(7.07)	19.58(3.82)

Table 4: The variable selection and parameter estimation results in the melanoma study contrasting the method assuming MAR and the proposed method.

	LASSO		SCAD		MCP	
	MAR	proposed	MAR	proposed	MAR	proposed
$\lvert\{i : \hat{\gamma}_i \neq 0\}\rvert$	2	3	2	2	1	2
treatment	-0.035	-0.024	-0.022	0.000	0.000	0.000
age	0.000	0.014	0.000	0.016	0.000	0.016
nodes1	0.422	0.564	0.539	0.691	0.528	0.691
sex	0.000	0.000	0.000	0.000	0.000	0.000
perform	0.000	0.000	0.000	0.000	0.000	0.000
log(Breslow)	0.000	0.000	0.000	0.000	0.000	0.000

Table 5: The variable selection and parameter estimation results in the prostate cancer study contrasting the method assuming MAR and the proposed method.

	LASSO		SCAD		MCP	
	MAR	proposed	MAR	proposed	MAR	proposed
$\lvert\{i : \hat{\gamma}_i \neq 0\}\rvert$	13	13	10	11	3	6
RHOB	-4.593	-3.224	-3.640	-0.868	-2.459	-0.885
MME	-0.111	-0.821	0.000	-0.276	0.000	-0.588
ANXA1	0.000	-0.280	0.000	-0.950	-1.140	-0.690
FAM89A	-2.917	-2.574	-0.937	-0.779	0.000	0.000
SETD5	1.279	2.727	0.555	0.434	0.000	0.000
CLDN4	0.000	0.577	0.000	0.483	0.000	0.586
SOX4	0.000	3.352	0.000	1.199	0.000	2.016
IMAGE:133130	0.000	2.455	0.000	1.418	0.000	3.487
ADAM22	-1.240	0.000	-0.535	0.000	-6.052	0.000
AMACR	0.098	1.066	0.000	0.345	0.000	0.000
ODF2	3.039	0.042	0.519	0.000	0.000	0.000
ST14	2.513	0.490	0.843	0.000	0.000	0.000
IMAGE:490971	0.638	0.000	1.615	0.000	0.000	0.000
RND3	-2.251	0.000	-0.182	0.000	0.000	0.000
KIAA0020	8.011	0.000	4.559	0.000	0.000	0.000
SLC25A6	1.660	0.000	0.485	0.000	0.000	0.000
MYO6	0.000	0.965	0.000	0.385	0.000	0.000
MYC	0.000	0.117	0.000	0.000	0.000	0.000
EFEMP2	-0.977	0.000	0.000	0.000	0.000	0.000
SERPING1	0.000	0.000	0.000	-0.163	0.000	0.000
others	0.000	0.000	0.000	0.000	0.000	0.000