High-Dimensional Variable Selection with Right Censored Length-biased Data

Di He[1,2], Yong Zhou[3,4] and Hui Zou[5]

[1] *School of Statistics and Management, Shanghai University of Finance and Economics, China*

[2] *School of Economics, Nanjing University, China*

[3] *Key Laboratory of Advanced Theory and Application in Statistics and Data Science, MOE*

[4] *Academy of Statistics and Interdisciplinary Sciences, East China Normal University, China*

[5] *School of Statistics, University of Minnesota, USA*

Supplementary Material

The supplementary file contains proofs of the theorems and full detailed tables of our simulation studies.

S1 Proof of Theorem 1.

By Foldes and Rejto (1981), the Kaplan-Meier estimator $\hat{S}_C(t)$ is uniformly consistent over $[0, t_0]$

$$\sup_{t\in[0,t_0]} |\hat{S}_C(t) - S_C(t)| = O((n/\log n)^{-\frac{1}{2}}), a.s.$$

thus,

$$\sup_{t\in[0,t_0]} |\hat{\pi}(t) - \pi(t)| \leq \sup_{t\in[0,t_0]} \int_0^t |\hat{S}_C(t) - S_C(t)| = O((n/\log n)^{-\frac{1}{2}}), a.s.$$

We can have $\frac{1}{M} < \hat{\pi}(Y) < \frac{1}{m}$ by condition (C).

Consider the folded concave penalized problem (3.2) in the paper with $P_\lambda(\cdot)$ satisfying (i)-(iv) in condition (E). Under condition (A), applying theorems 1-2 in Fan et al. (2014), we have the convergence of the LLA solution $\hat{\boldsymbol{\beta}}$ initialized by $\hat{\boldsymbol{\beta}}^{\text{initial}}$ to $\tilde{\boldsymbol{\beta}}^{\text{oracle}}$ after two iterations with probability at least $1 - \delta_0 - \delta_1 - \delta_2$, where

$$\delta_0 = \Pr(\|\hat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^*\|_{\max} > a_0\lambda),$$

$$\delta_1 = \Pr(\|\nabla_{\mathcal{A}^c}\ell_n(\tilde{\boldsymbol{\beta}}^{\text{oracle}})\|_{\max} > a_1\lambda),$$

$$\delta_2 = \Pr(\|\tilde{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} \leq a\lambda).$$

We can derive the explicit upper bounds for δ_1 and δ_2 , which only depends on the behavior of the oracle estimator.

Let $\widetilde{\mathbf{H}}_{\mathcal{A}} = W^{\frac{1}{2}}\widetilde{\mathbf{X}}_{\mathcal{A}}(\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\widetilde{\mathbf{X}}_{\mathcal{A}})^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}}^T W^{\frac{1}{2}}$. Since $\log\tilde{T} = \mathbf{X}_{\mathcal{A}}^T\boldsymbol{\beta}^*_{\mathcal{A}} + \epsilon$, by the estimating equation we have $\tilde{y} - \widetilde{\mathbf{X}}_{\mathcal{A}}\boldsymbol{\beta}^*_{\mathcal{A}} = \frac{1}{n}\mathbf{X}^T\left(\mathbf{D}\mathbf{y} - \mathbf{D}\mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}^*_{\mathcal{A}}\right) = \frac{1}{n}\mathbf{X}^T\epsilon$. Since $\widetilde{\mathbf{X}}_{\mathcal{A}^c} = \frac{1}{n}\mathbf{X}^T\mathbf{D}\mathbf{X}_{\mathcal{A}^c}$, so $\nabla_{\mathcal{A}^c}\ell_n(\tilde{\boldsymbol{\beta}}^{\text{oracle}}) = 2\widetilde{\mathbf{X}}_{\mathcal{A}^c}^T W^{\frac{1}{2}}(W^{\frac{1}{2}}\tilde{y} - \widetilde{\mathbf{H}}_{\mathcal{A}}W^{\frac{1}{2}}\tilde{y}) = \frac{2}{n^2}\mathbf{X}_{\mathcal{A}^c}^T\mathbf{D}\mathbf{X}W^{\frac{1}{2}}(\mathbf{I} - \widetilde{\mathbf{H}}_{\mathcal{A}})W^{\frac{1}{2}}\mathbf{X}^T\epsilon$.

To simplify notation, denote $\mathbf{u}_j^T = \mathbf{e}_j^T\mathbf{X}_{\mathcal{A}^c}^T\mathbf{D}\mathbf{X}W^{\frac{1}{2}}(\mathbf{I} - \widetilde{\mathbf{H}}_{\mathcal{A}})W^{\frac{1}{2}}\mathbf{X}^T$, where $\mathbf{e}_j$ is the unit vector with jth element being 1. Since $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$ being i.i.d. sub-Gaussian(σ) for some fixed constant $\sigma > 0$, by Hoeffding bound, we have

$$\delta_1 = \Pr(\|\nabla_{\mathcal{A}^c}\ell_n(\tilde{\boldsymbol{\beta}}^{\text{oracle}})\|_{\max} > a_1\lambda) \leq \sum_{j\in\mathcal{A}^c} \Pr(|\mathbf{u}_j^T\epsilon| > \frac{a_1 n^2\lambda}{2}) \\ \leq 2\sum_{j\in\mathcal{A}^c} \exp\left(-\frac{a_1^2 n^4\lambda^2}{8\sigma^2\cdot\|\mathbf{u}_j\|_2^2}\right),$$

$$\begin{aligned}\|\mathbf{u}_j\|_2^2 &= \mathbf{e}_j^T\mathbf{X}_{\mathcal{A}^c}^T\mathbf{D}\mathbf{X}W^{\frac{1}{2}}(\mathbf{I}-\widetilde{\mathbf{H}}_{\mathcal{A}})W^{\frac{1}{2}}\mathbf{X}^T\mathbf{X}W^{\frac{1}{2}}(\mathbf{I}-\widetilde{\mathbf{H}}_{\mathcal{A}})W^{\frac{1}{2}}\mathbf{X}^T\mathbf{D}\mathbf{X}_{\mathcal{A}^c}\mathbf{e}_j \\ &\leq \mathbf{e}_j^T\mathbf{X}_{\mathcal{A}^c}^T\mathbf{D}(\mathbf{X}W\mathbf{X}^T)^2\mathbf{D}\mathbf{X}_{\mathcal{A}^c}\mathbf{e}_j \\ &\leq (\lambda_{\max}^W)^2\mathbf{e}_j^T\mathbf{X}_{\mathcal{A}^c}^T\mathbf{D}(\mathbf{X}\mathbf{X}^T)^2\mathbf{D}\mathbf{X}_{\mathcal{A}^c}\mathbf{e}_j \\ &= (\lambda_{\max}^W)^2\mathbf{e}_j^T\mathbf{X}_{\mathcal{A}^c}^{oT}\mathbf{D}_{11}(\mathbf{X}^o\mathbf{X}^{oT})^2\mathbf{D}_{11}\mathbf{X}_{\mathcal{A}^c}^o\mathbf{e}_j.\end{aligned}$$

The last equality holds because we swap the row order in $\mathbf{D}$ to transform $\mathbf{D} = \begin{pmatrix}\mathbf{D}_{11} & 0 \\ 0 & 0\end{pmatrix}$, and re-arrange the same observation order in each other matrix according to $\mathbf{D}$, where $\mathbf{D}_{11} = \underset{i:\delta_i=1}{diag}(\frac{\delta_i}{\hat{w}(Y_i)})$, and $\mathbf{X}_{\mathcal{A}^c}^o, \mathbf{X}^o$ are the sub-matrixes formed by the rows in $\mathbf{X}_{\mathcal{A}^c}, \mathbf{X}$ where T_i is being observed, i.e. $\delta_i = 1$. Since

$$\mathbf{X}^o\mathbf{X}^{oT} = \mathbf{X}_{\mathcal{A}}^o\mathbf{X}_{\mathcal{A}}^{oT} + \mathbf{X}_{\mathcal{A}^c}^o\mathbf{X}_{\mathcal{A}^c}^{oT} \leq n(\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})\mathbf{I},$$

then

$$\begin{aligned}\|\mathbf{u}_j\|_2^2 &\leq n^2M^2(\lambda_{\max}^W)^2(\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2\mathbf{e}_j^T\mathbf{X}_{\mathcal{A}^c}^{oT}\mathbf{X}_{\mathcal{A}^c}^o\mathbf{e}_j \\ &\leq n^3M^2(\lambda_{\max}^W)^2\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c}(\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2.\end{aligned}$$

We conclude that

$$\delta_1 \leq 2(p+1-s)\exp\left(-\frac{a_1^2 n\lambda^2}{8\sigma^2 M^2(\lambda_{\max}^W)^2\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c}(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}\right).$$

Next, we derive the bound $\delta_2 = \Pr(\|\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}}\|_{\min} \leq a\lambda)$. Note that $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}} = \boldsymbol{\beta}_{\mathcal{A}}^* + \frac{1}{n}(\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\widetilde{\mathbf{X}}_{\mathcal{A}})^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\mathbf{X}\epsilon$, and then $\|\tilde{\boldsymbol{\beta}}^{\text{oracle}}\|_{\min} \geq \|\boldsymbol{\beta}_{\mathcal{A}}^*\|_{\min} - \|\frac{1}{n}(\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\widetilde{\mathbf{X}}_{\mathcal{A}})^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\mathbf{X}^T\epsilon\|_{\max}$. Thus, we have

$$\delta_2 \leq \Pr\left(\|\frac{1}{n}(\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\widetilde{\mathbf{X}}_{\mathcal{A}})^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\mathbf{X}^T\epsilon\|_{\max} \geq \|\boldsymbol{\beta}_{\mathcal{A}}^*\|_{\min} - a\lambda\right).$$

Denote $\mathbf{v}_j^T = \mathbf{e}_j^T\frac{1}{n}(\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\widetilde{\mathbf{X}}_{\mathcal{A}})^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}}^T W\mathbf{X}^T = \mathbf{e}_j^T(\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}W\mathbf{X}^T\mathbf{D}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}W\mathbf{X}^T$, then

$$\begin{aligned}
\|\mathbf{v}_j^T\|_2^2 &\leq \left(\frac{\lambda_{\max}^W}{\lambda_{\min}^W}\right)^2\|\mathbf{e}_j^T(\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{D}_{11}\mathbf{X}^o\mathbf{X}^{oT}\mathbf{D}_{11}\mathbf{X}^o_{\mathcal{A}})^{-1}\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{D}_{11}(\mathbf{X}^o\mathbf{X}^{oT})\|_2^2 \\
&\leq \left(\frac{\lambda_{\max}^W}{\lambda_{\min}^W}\right)^2\|n\frac{M}{m^2}(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})\mathbf{e}_j^T(\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{X}^o\mathbf{X}^{oT}\mathbf{X}^o_{\mathcal{A}})^{-1}\mathbf{X}^{oT}_{\mathcal{A}}\|_2^2 \\
&\leq n\left(\frac{\lambda_{\max}^W}{\lambda_{\min}^W}\right)^2\lambda_{\max}^{\mathcal{A}\mathcal{A}}\|n\frac{M}{m^2}(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})\mathbf{e}_j^T\left((\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{X}^o_{\mathcal{A}})^2+\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{X}^o_{\mathcal{A}^c}\mathbf{X}^{oT}_{\mathcal{A}^c}\mathbf{X}^o{}_{\mathcal{A}}\right)^{-1}\|_2^2 \\
&\leq \frac{1}{n}\frac{M^2}{m^4}\left(\frac{\lambda_{\max}^W}{\lambda_{\min}^W}\right)^2\lambda_{\max}^{\mathcal{A}\mathcal{A}}\frac{(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}{\lambda_{\min}^{\mathcal{A}\mathcal{A}\,4}}.
\end{aligned}$$

The last inequality holds because $(\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{X}^o_{\mathcal{A}})^2$ and $\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{X}^o_{\mathcal{A}^c}\mathbf{X}^{oT}_{\mathcal{A}^c}\mathbf{X}^o{}_{\mathcal{A}}$ are non-negative definite, and the minimum eigenvalue of their sum is bigger than the eigenvalue of individual. Again, by Hoeffding bound, we have

$$\begin{aligned}\delta_2 &\leq \Pr\left(\|\frac{1}{n}(\widetilde{\mathbf{X}}_{\mathcal{A}}^T W \widetilde{\mathbf{X}}_{\mathcal{A}})^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}}^T W \mathbf{X}^T \epsilon\|_{\max} \geq \|\boldsymbol{\beta}_{\mathcal{A}}^*\|_{\min} - a\lambda\right) \\ &\leq 2\sum_{j=1}^{s} \exp\left(-\frac{(\|\boldsymbol{\beta}_{\mathcal{A}}^*\|_{\min} - a\lambda)^2}{2\sigma^2\|\mathbf{v}_j\|_2^2}\right) \\ &\leq 2s\exp\left(-\frac{n\cdot m^4(\|\boldsymbol{\beta}_{\mathcal{A}}^*\|_{\min} - a\lambda)^2}{2\sigma^2 M^2}\frac{{\lambda_{\min}^{\mathcal{A}\mathcal{A}}}^4}{\lambda_{\max}^{\mathcal{A}\mathcal{A}}(\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}\left(\frac{\lambda_{\min}^W}{\lambda_{\max}^W}\right)^2\right).\end{aligned}$$

Finally, we derive the bound $\delta_0^{\text{lasso}} = \Pr(\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*\|_{\max} > a_0\lambda)$ using LASSO as the initial value. By the definition of the LASSO estimator,

$$\| W^{\frac{1}{2}}\tilde{y} - W^{\frac{1}{2}}\widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^{\text{lasso}}\|_2^2 + \lambda_{\text{lasso}}\|\hat{\boldsymbol{\beta}}^{\text{lasso}}\|_1 \ \leq \ \| W^{\frac{1}{2}}\tilde{y} - W^{\frac{1}{2}}\widetilde{\mathbf{X}}\boldsymbol{\beta}^*\|_2^2 + \lambda_{\text{lasso}}\|\boldsymbol{\beta}^*\|_1,$$

Since $\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}^* = \frac{1}{n}\mathbf{X}^T\epsilon$, we have

$$\| W^{\frac{1}{2}}\widetilde{\mathbf{X}}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{\text{lasso}}) + W^{\frac{1}{2}}\frac{1}{n}\mathbf{X}^T\epsilon\|_2^2 + \lambda_{\text{lasso}}\|\hat{\boldsymbol{\beta}}^{\text{lasso}}\|_1 \ \leq \ \|W^{\frac{1}{2}}\frac{1}{n}\mathbf{X}^T\epsilon\|_2^2 + \lambda_{\text{lasso}}\|\boldsymbol{\beta}_{\mathcal{A}}^*\|_1,$$

$$\begin{aligned}&(\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*)^T\widetilde{\mathbf{X}}^T W\widetilde{\mathbf{X}}(\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*) \\ \leq &\frac{2}{n}\epsilon^T\mathbf{X}W\widetilde{\mathbf{X}}(\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*) + \lambda_{\text{lasso}}(\|\boldsymbol{\beta}_{\mathcal{A}}^*\|_1 - \|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{lasso}}\|_1) - \lambda_{\text{lasso}}\|\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{\text{lasso}}\|_1 \\ \leq &\|\frac{2}{n}\epsilon^T\mathbf{X}W\widetilde{\mathbf{X}}\|_{\max}\cdot\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*\|_1 + \lambda_{\text{lasso}}(\|\boldsymbol{\beta}_{\mathcal{A}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{lasso}}\|_1) - \lambda_{\text{lasso}}\|\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{\text{lasso}}\|_1.\end{aligned}$$

Denote $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*, \hat{\boldsymbol{\delta}}_{\mathcal{A}} = \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{lasso}} - \boldsymbol{\beta}_{\mathcal{A}}^*, \hat{\boldsymbol{\delta}}_{\mathcal{A}^c} = \hat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{\text{lasso}} - 0 = \hat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{\text{lasso}}$, under the event $\{\|\frac{2}{n}\epsilon^T\mathbf{X}W\widetilde{\mathbf{X}}\|_{\max} \leq \lambda_{\text{lasso}}c\}$ for any $c \in (0,1)$, we have

$$0 \leq \hat{\boldsymbol{\delta}}^T\widetilde{\mathbf{X}}^T W\widetilde{\mathbf{X}}\hat{\boldsymbol{\delta}} \leq \lambda_{\text{lasso}}c(\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_1 + \|\hat{\boldsymbol{\delta}}_{\mathcal{A}^c}\|_1) + \lambda_{\text{lasso}}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_1 - \lambda_{\text{lasso}}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}^c}\|_1,$$

$$\|\hat{\boldsymbol{\delta}}_{\mathcal{A}^c}\|_1 \leq \frac{1+c}{1-c}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_1.$$

By the definition of the restricted eigenvalue κ, we have

$$\kappa\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_2^2 \leq \kappa\|\hat{\boldsymbol{\delta}}\|_2^2 \leq \hat{\boldsymbol{\delta}}^T \widetilde{\mathbf{X}}^T W \widetilde{\mathbf{X}} \hat{\boldsymbol{\delta}} \leq \lambda_{\text{lasso}} c(\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_1 + \|\hat{\boldsymbol{\delta}}_{\mathcal{A}^c}\|_1) + \lambda_{\text{lasso}}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_1$$
$$\leq \lambda_{\text{lasso}} \frac{1+c}{1-c}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_1 \leq \lambda_{\text{lasso}} \frac{1+c}{1-c}\sqrt{s}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_2.$$

Hence,

$$\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_2 \leq \frac{\lambda_{\text{lasso}}(1+c)\sqrt{s}}{(1-c)\kappa}.$$

Let $c = \frac{1}{2}$, since $\lambda \geq \frac{3\sqrt{s}\lambda_{\text{lasso}}}{a_0\kappa}$, it follows

$$\delta_0^{\text{lasso}} = \Pr(\|\hat{\boldsymbol{\delta}}\|_{\max} > a_0\lambda) \leq \Pr(\|\hat{\boldsymbol{\delta}}\|_2 > a_0\lambda) \leq \Pr(\|\hat{\boldsymbol{\delta}}\|_2 > \frac{3\lambda_{\text{lasso}}\sqrt{s}}{\kappa})$$
$$\leq \Pr(\|\frac{2}{n}\epsilon^T \mathbf{X} W \widetilde{\mathbf{X}}\|_{\max} > \frac{1}{2}\lambda_{\text{lasso}}) = \Pr(\|\mathbf{X}^T \mathbf{D} \mathbf{X} W \mathbf{X}^T \epsilon\|_{\max} > \frac{n^2\lambda_{\text{lasso}}}{4})$$
$$\leq 2\sum_{j=1}^{p+1} \Pr(|\mathbf{e}_j^T \mathbf{X}^T \mathbf{D} \mathbf{X} W \mathbf{X}^T \epsilon| > \frac{n^4\lambda_{\text{lasso}}^2}{16})$$
$$\leq 2(p+1)\exp\left(-\frac{n^4\lambda_{\text{lasso}}^2}{32\sigma^2\|\mathbf{e}_j^T \mathbf{X}^T \mathbf{D} \mathbf{X} W \mathbf{X}^T\|_2^2}\right).$$

Since

$$\mathbf{e}_j^T \mathbf{X}^T \mathbf{D} \mathbf{X} W \mathbf{X}^T \mathbf{X} W \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{e}_j$$
$$\leq (\lambda_{\max}^W)^2 M^2 \mathbf{e}_j^T (\mathbf{X}^{oT}\mathbf{X}^o)^3 \mathbf{e}_j \leq (\lambda_{\max}^W)^2 M^2 \lambda_{\max}^3\{\mathbf{X}^{oT}\mathbf{X}^o\}$$
$$= (\lambda_{\max}^W)^2 M^2 \lambda_{\max}^3\{\mathbf{X}^o\mathbf{X}^{oT}\} = (\lambda_{\max}^W)^2 M^2 \lambda_{\max}^3\{\mathbf{X}^o_{\mathcal{A}}\mathbf{X}^{oT}_{\mathcal{A}} + \mathbf{X}^o_{\mathcal{A}^c}\mathbf{X}^{oT}_{\mathcal{A}^c}\}$$
$$\leq n^3(\lambda_{\max}^W)^2 M^2 (\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^3,$$

thus

$$\delta_0^{\text{lasso}} \leq 2(p+1)\exp\left(-\frac{n\lambda_{\text{lasso}}^2}{32\sigma^2(\lambda_{\max}^W)^2 M^2 (\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^3}\right).$$

We then prove the second part of the theorem. By triangle inequality,

$$\begin{aligned}\Pr(\|\hat{\boldsymbol\beta}-\hat{\boldsymbol\beta}_{\mathcal{A}}^{\text{oracle}}\|_{\max}>\xi n^{-\theta}) \leq& \Pr(\|\hat{\boldsymbol\beta}-\tilde{\boldsymbol\beta}_{\mathcal{A}}^{\text{oracle}}\|_{\max}>\frac{1}{2}\xi n^{-\theta}) \\ &+\Pr(\|\tilde{\boldsymbol\beta}_{\mathcal{A}}^{\text{oracle}}-\hat{\boldsymbol\beta}_{\mathcal{A}}^{\text{oracle}}\|_{\max}>\frac{1}{2}\xi n^{-\theta})\end{aligned}$$

By Theorem 1, we have $\Pr(\hat{\boldsymbol\beta}\neq\tilde{\boldsymbol\beta}^{\text{oracle}})\leq\delta_0^{\text{lasso}}+\delta_1+\delta_2$. We only need to prove $\delta_3=\Pr(\|\tilde{\boldsymbol\beta}_{\mathcal{A}}^{\text{oracle}}-\hat{\boldsymbol\beta}_{\mathcal{A}}^{\text{oracle}}\|_{\max}>\frac{1}{2}\xi n^{-\theta})$ tending to 0, for any $\xi>0$. Note that by (2.4), $\hat{\boldsymbol\beta}_{\mathcal{A}}^{\text{oracle}}=(\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{y}$, and $\mathbf{X}_{\mathcal{A}}^T(\mathbf{D}\mathbf{y}-\mathbf{D}\mathbf{X}_{\mathcal{A}}\boldsymbol\beta_{\mathcal{A}}^*)=\mathbf{X}_{\mathcal{A}}^T\epsilon$, then $\hat{\boldsymbol\beta}_{\mathcal{A}}^{\text{oracle}}=\boldsymbol\beta_{\mathcal{A}}^*+(\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^T\epsilon$. Denote $\mathbf{w}_j^T=\mathbf{e}_j^T(\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^T$

$$\begin{aligned}\delta_3&\leq 2\sum_{j=1}^{s}\Pr(|(\mathbf{v}_j^T\epsilon-\mathbf{w}_j^T\epsilon|>\frac{1}{2}\xi n^{-\theta})\\ &\leq 2s\exp\left(-\frac{n^{-2\theta}\xi^2}{8\sigma^2\|\mathbf{v}_j^T-\mathbf{w}_j^T\|_2^2}\right)\\ &\leq 2s\exp\left(-\frac{n^{-2\theta}\xi^2}{16\sigma^2(\|\mathbf{v}_j^T\|_2^2+\|\mathbf{w}_j^T\|_2^2)}\right).\end{aligned}$$

Since

$$\begin{aligned}\|\mathbf{w}_j^T\|_2^2&=\mathbf{e}_j^T(\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{e}_j\\ &\leq\frac{1}{n}\frac{1}{m^2}\frac{\lambda_{\max}^{\mathcal{A}\mathcal{A}}}{\lambda_{\min}^{\mathcal{A}\mathcal{A}^2}},\end{aligned}$$

we have

$$\delta_3\leq 2s\exp\left(-\frac{n^{1-2\theta}\xi^2}{16\sigma^2}\frac{1}{\lambda_{\max}^{\mathcal{A}\mathcal{A}}}\left[m^2\lambda_{\min}^{\mathcal{A}\mathcal{A}^2}+\frac{M^4}{m^2}\frac{\lambda_{\min}^{\mathcal{A}\mathcal{A}^4}}{(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}\left(\frac{\lambda_{\min}^{W}}{\lambda_{\max}^{W}}\right)^2\right]\right).$$

S2 Simulation tables for Example 1.

S3 Simulation tables for Example 2.

Bibliography

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of statistics*, 42(3):819.

Foldes, A. and Rejto, L. (1981). Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. *The Annals of Statistics*, pages 122–129.

Table 1: Average numbers of correct and incorrect non-zero coefficients and average of mean squared errors from 1000 simulated datasets for Example 1, with their standard error shown in the parenthesis

			LASSO			SCAD			MS-SCAD		
error	p	censoring	C	I	MSE	C	I	MSE	C	I	MSE
unif	20	10%	2.00	11.64	0.011	2.00	13.28	0.013	2.00	0.29	0.003
			(0)	(2.98)	(0.005)	(0)	(2.07)	(0.005)	(0)	(0.71)	(0.004)
		30%	2.00	11.69	0.013	2.00	13.28	0.015	2.00	0.29	0.003
			(0)	(3.02)	(0.006)	(0)	(2.07)	(0.005)	(0)	(0.67)	(0.004)
		60%	2.00	11.51	0.021	2.00	13.14	0.023	2.00	0.38	0.005
			(0.04)	(3.12)	(0.011)	(0)	(2.2)	(0.009)	(0)	(0.82)	(0.007)
	100	10%	2.00	28.68	0.036	2.00	36.48	0.041	2.00	0.78	0.005
			(0.03)	(8.89)	(0.015)	(0)	(8.96)	(0.009)	(0)	(1.6)	(0.007)
		30%	2.00	28.37	0.038	2.00	36.35	0.045	2.00	0.76	0.006
			(0.05)	(9.53)	(0.016)	(0)	(9.41)	(0.01)	(0)	(1.48)	(0.007)
		60%	2.00	30.66	0.047	2.00	38.61	0.061	2.00	1.28	0.013
			(0.03)	(11.03)	(0.02)	(0)	(9.91)	(0.014)	(0)	(2.29)	(0.017)
	400	10%	2.00	54.94	0.055	2.00	72.06	0.056	2.00	2.07	0.011
			(0)	(17.96)	(0.018)	(0)	(15.72)	(0.007)	(0)	(2.97)	(0.012)
		30%	2.00	57.94	0.053	2.00	77.25	0.061	2.00	2.14	0.014
			(0)	(19.63)	(0.018)	(0)	(17.22)	(0.008)	(0)	(3.16)	(0.015)
		60%	2.00	106.30	0.063	2.00	117.13	0.072	2.00	4.00	0.031
			(0)	(40.09)	(0.023)	(0)	(33.27)	(0.01)	(0)	(7.24)	(0.054)
exp	20	10%	2.00	10.68	0.005	2.00	12.02	0.005	2.00	0.17	0.001
			(0)	(2.83)	(0.003)	(0)	(2.22)	(0.002)	(0)	(0.48)	(0.001)
		30%	2.00	10.69	0.006	2.00	12.11	0.006	2.00	0.18	0.001
			(0)	(2.87)	(0.004)	(0)	(2.21)	(0.003)	(0)	(0.55)	(0.001)
		60%	2.00	10.46	0.010	2.00	12.05	0.011	2.00	0.20	0.002
			(0)	(2.9)	(0.007)	(0)	(2.28)	(0.005)	(0)	(0.6)	(0.002)
	100	10%	2.00	24.56	0.020	2.00	29.66	0.019	2.00	0.30	0.001
			(0)	(7.69)	(0.01)	(0)	(9.37)	(0.008)	(0)	(0.89)	(0.002)
		30%	2.00	24.25	0.021	2.00	30.22	0.023	2.00	0.16	0.001
			(0.03)	(8.48)	(0.011)	(0)	(9.86)	(0.009)	(0)	(0.59)	(0.002)
		60%	2.00	25.90	0.026	2.00	34.02	0.035	2.00	0.20	0.002
			(0)	(10)	(0.014)	(0)	(10.31)	(0.012)	(0)	(0.78)	(0.003)
	400	10%	2.00	40.90	0.030	2.00	57.27	0.033	2.00	0.25	0.001
			(0)	(13.43)	(0.013)	(0)	(13.9)	(0.006)	(0)	(0.98)	(0.002)
		30%	2.00	42.33	0.030	2.00	62.34	0.037	2.00	0.24	0.001
			(0)	(14.83)	(0.014)	(0)	(15.35)	(0.007)	(0)	(0.86)	(0.002)
		60%	2.00	75.60	0.034	2.00	95.70	0.050	2.00	1.02	0.004
			(0)	(38.1)	(0.017)	(0)	(31.12)	(0.01)	(0)	(5.13)	(0.024)
normal	20	10%	2.00	11.57	0.013	2.00	13.33	0.014	2.00	0.20	0.003
			(0)	(3.01)	(0.006)	(0)	(2.01)	(0.005)	(0)	(0.56)	(0.003)
		30%	2.00	11.84	0.015	2.00	13.52	0.016	2.00	0.22	0.003
			(0)	(2.97)	(0.007)	(0)	(1.97)	(0.006)	(0)	(0.58)	(0.004)
		60%	2.00	11.52	0.022	2.00	13.23	0.025	2.00	0.29	0.005
			(0)	(3.21)	(0.011)	(0)	(2.18)	(0.011)	(0)	(0.72)	(0.007)
	100	10%	2.00	29.56	0.038	2.00	37.41	0.043	2.00	0.40	0.004
			(0.03)	(9.31)	(0.016)	(0)	(9.25)	(0.011)	(0)	(0.94)	(0.005)
		30%	2.00	29.41	0.040	2.00	37.22	0.048	2.00	0.56	0.005
			(0)	(9.49)	(0.016)	(0)	(9.21)	(0.012)	(0)	(1.32)	(0.007)
		60%	2.00	31.50	0.049	2.00	39.28	0.065	2.00	0.84	0.010
			(0)	(10.75)	(0.021)	(0)	(9.93)	(0.016)	(0)	(1.85)	(0.016)
	400	10%	2.00	56.87	0.058	2.00	73.03	0.058	2.00	1.02	0.007
			(0.03)	(18.16)	(0.019)	(0)	(15.66)	(0.008)	(0)	(2.2)	(0.01)
		30%	2.00	59.95	0.058	2.00	78.39	0.063	2.00	1.43	0.010
			(0.03)	(20.14)	(0.019)	(0)	(17.35)	(0.009)	(0)	(2.93)	(0.015)
		60%	2.00	103.52	0.069	2.00	116.18	0.075	2.00	3.61	0.026
			(0)	(38.9)	(0.025)	(0)	(32.38)	(0.012)	(0.03)	(8.56)	(0.037)

Table 2: Estimates of coefficients for Multi-Stage SCAD, their biases, standard errors, mean of asymptotic standard errors, and coverage probabilities for nominal 95% confidence intervals from 1000 simulated datasets for Example 1

			unif				exp				normal			
p	censoring		Bias	SE	ASE	CP	Bias	SE	ASE	CP	Bias	SE	ASE	CP
20	10%	b1	-0.0028	0.0525	0.0507	93.9	-0.0020	0.0311	0.0301	93.9	-0.0006	0.0565	0.0537	93.0
		b2	-0.0037	0.0925	0.0871	92.6	0.0003	0.0538	0.0521	94.0	-0.0025	0.0988	0.0924	93.7
	30%	b1	-0.0050	0.0573	0.0545	93.2	-0.0001	0.0350	0.0332	92.5	-0.0031	0.0606	0.0573	93.3
		b2	0.0000	0.0960	0.0941	94.4	-0.0036	0.0611	0.0576	93.6	-0.0047	0.1035	0.0981	92.5
	60%	b1	-0.0064	0.0706	0.0653	91.8	0.0011	0.0454	0.0429	93.8	-0.0014	0.0732	0.0686	93.1
		b2	-0.0033	0.1190	0.1136	92.9	-0.0043	0.0800	0.0740	92.7	-0.0041	0.1266	0.1182	92.4
100	10%	b1	-0.0022	0.0534	0.0496	92.3	-0.0016	0.0303	0.0297	93.9	-0.0050	0.0556	0.0531	92.8
		b2	-0.0068	0.0947	0.0856	90.4	-0.0009	0.0529	0.0516	94.2	-0.0007	0.0992	0.0911	91.8
	30%	b1	-0.0040	0.0565	0.0532	92.5	-0.0016	0.0349	0.0332	93.7	-0.0047	0.0603	0.0560	91.6
		b2	-0.0135	0.0997	0.0913	92.5	-0.0009	0.0593	0.0573	93.7	-0.0111	0.1047	0.0958	92.1
	60%	b1	-0.0059	0.0698	0.0626	90.5	-0.0011	0.0452	0.0430	94.2	-0.0061	0.0738	0.0664	91.8
		b2	-0.0197	0.1277	0.1083	88.7	0.0001	0.0755	0.0741	94.0	-0.0137	0.1254	0.1142	91.3
400	10%	b1	-0.0104	0.0542	0.0470	89.5	-0.0003	0.0312	0.0298	93.9	-0.0075	0.0558	0.0514	91.4
		b2	-0.0250	0.0967	0.0808	87.4	-0.0052	0.0544	0.0517	94.5	-0.0136	0.1016	0.0883	89.5
	30%	b1	-0.0124	0.0600	0.0503	87.5	-0.0032	0.0358	0.0331	93.4	-0.0085	0.0576	0.0542	91.8
		b2	-0.0226	0.1072	0.0867	85.8	-0.0039	0.0598	0.0574	93.2	-0.0278	0.1055	0.0929	88.8
	60%	b1	-0.0235	0.0762	0.0567	81.0	-0.0038	0.0476	0.0422	91.8	-0.0189	0.0780	0.0609	83.2
		b2	-0.0529	0.1575	0.0982	76.1	-0.0101	0.0873	0.0729	90.9	-0.0562	0.1551	0.1054	80.1

Table 3: Average numbers of correct and incorrect non-zero coefficients and average of mean squared errors from 1000 simulated datasets for Example 2, with their standard error shown in the parenthesis; AR(ρ) is the autoregressive correlation structure for predictors

			LASSO			SCAD			MS-SCAD		
	p	censoring	C	I	MSE	C	I	MSE	C	I	MSE
AR(0.5)	20	10%	3.00	14.30	0.006	3.00	14.39	0.007	3.00	0.29	0.002
			(0)	(2.19)	(0.002)	(0)	(1.43)	(0.002)	(0)	(0.91)	(0.002)
		30%	3.00	14.31	0.006	3.00	14.32	0.008	3.00	0.36	0.002
			(0)	(2.12)	(0.002)	(0)	(1.49)	(0.003)	(0)	(1)	(0.002)
		60%	3.00	14.32	0.011	3.00	14.37	0.014	3.00	0.64	0.004
			(0)	(2.1)	(0.005)	(0)	(1.45)	(0.005)	(0)	(1.44)	(0.004)
	100	10%	3.00	68.70	0.018	3.00	70.39	0.033	3.00	2.99	0.005
			(0)	(10.74)	(0.005)	(0)	(7.51)	(0.009)	(0)	(6.95)	(0.008)
		30%	3.00	66.87	0.022	3.00	68.68	0.044	3.00	5.27	0.009
			(0)	(10.16)	(0.007)	(0)	(7.82)	(0.015)	(0)	(8.82)	(0.013)
		60%	3.00	65.43	0.036	3.00	68.89	0.131	2.99	7.36	0.023
			(0)	(8.77)	(0.011)	(0)	(7.69)	(0.065)	(0.11)	(9.66)	(0.029)
	400	10%	3.00	230.92	0.030	3.00	263.35	0.437	3.00	3.98	0.006
			(0)	(29.47)	(0.005)	(0)	(24.64)	(0.096)	(0.08)	(8.22)	(0.011)
		30%	3.00	243.69	0.034	3.00	251.83	0.507	2.99	2.65	0.006
			(0)	(25.98)	(0.006)	(0)	(24.77)	(0.1)	(0.12)	(5.53)	(0.011)
		60%	3.00	213.96	0.036	3.00	218.36	0.551	2.91	1.98	0.015
			(0)	(25.11)	(0.01)	(0)	(25.66)	(0.091)	(0.34)	(4.59)	(0.051)
AR(0.8)	20	10%	3.00	7.64	0.004	3.00	8.21	0.006	3.00	0.45	0.002
			(0)	(3)	(0.002)	(0)	(2.98)	(0.003)	(0)	(1.08)	(0.002)
		30%	3.00	7.55	0.004	3.00	8.36	0.006	3.00	0.37	0.002
			(0)	(3.01)	(0.002)	(0)	(3)	(0.003)	(0.03)	(0.99)	(0.002)
		60%	3.00	7.81	0.007	3.00	8.70	0.01	3.00	0.54	0.004
			(0)	(3.02)	(0.004)	(0)	(2.86)	(0.005)	(0)	(1.16)	(0.004)
	100	10%	3.00	40.19	0.010	3.00	44.45	0.019	3.00	1.80	0.004
			(0)	(11.61)	(0.003)	(0)	(9.43)	(0.005)	(0.03)	(4.06)	(0.005)
		30%	3.00	39.78	0.012	3.00	44.58	0.023	3.00	2.15	0.005
			(0)	(11.59)	(0.004)	(0)	(9.63)	(0.006)	(0.03)	(4.24)	(0.006)
		60%	3.00	43.01	0.020	3.00	47.42	0.042	2.99	4.36	0.014
			(0)	(10.89)	(0.007)	(0)	(9.68)	(0.016)	(0.12)	(5.84)	(0.017)
	400	10%	3.00	154.67	0.022	3.00	190.94	0.14	2.99	6.57	0.010
			(0)	(30.32)	(0.004)	(0)	(27.58)	(0.047)	(0.08)	(11.15)	(0.014)
		30%	3.00	169.42	0.026	3.00	211.08	0.218	2.98	3.00	0.007
			(0)	(32.17)	(0.006)	(0)	(27.18)	(0.072)	(0.13)	(6.26)	(0.011)
		60%	3.00	187.90	0.037	3.00	194.92	0.383	2.86	1.80	0.012
			(0)	(27.73)	(0.01)	(0)	(27.81)	(0.134)	(0.39)	(2.95)	(0.017)

Table 4: Estimates of coefficients for Multi-Stage SCAD, their biases, standard errors, mean of asymptotic standard errors, and coverage probabilities for nominal 95% confidence intervals from 1000 simulated datasets for Example 2; AR(ρ) is the autoregressive correlation structure for predictors

			AR(0.5)				AR(0.8)			
p	censoring		Bias	SE	ASE	CP	Bias	SE	ASE	CP
20	10%	b1	-0.0015	0.0211	0.0199	92.8	-0.0006	0.0336	0.0299	92.0
		b2	-0.0001	0.0222	0.0200	91.1	0.0002	0.0353	0.0319	92.1
		b5	-0.0005	0.0195	0.0176	93.0	0.0003	0.0263	0.0215	89.5
	30%	b1	-0.0006	0.0239	0.0211	91.5	-0.0018	0.0346	0.0313	92.6
		b2	-0.0020	0.0241	0.0213	90.1	-0.0003	0.0390	0.0335	92.0
		b5	-0.0006	0.0203	0.0184	92.3	-0.0022	0.0283	0.0225	88.7
	60%	b1	-0.0015	0.0280	0.0259	91.8	-0.0022	0.0431	0.0373	89.8
		b2	0.0000	0.0289	0.0262	91.7	-0.0018	0.0478	0.0394	89.8
		b5	-0.0004	0.0259	0.0229	91.5	-0.0023	0.0359	0.0266	87.5
100	10%	b1	-0.0020	0.0211	0.0188	90.7	-0.0019	0.0338	0.0290	90.4
		b2	-0.0003	0.0228	0.0191	87.4	0.0003	0.0370	0.0307	90.6
		b5	-0.0017	0.0189	0.0166	89.4	-0.0023	0.0248	0.0203	87.3
	30%	b1	0.0008	0.0236	0.0196	88.6	0.0001	0.0350	0.0301	90.3
		b2	-0.0030	0.0244	0.0197	86.6	-0.0035	0.0373	0.0319	89.9
		b5	-0.0022	0.0220	0.0172	86.9	-0.0026	0.0276	0.0216	87.6
	60%	b1	-0.0025	0.0358	0.0222	78.4	0.0012	0.0496	0.0333	84.0
		b2	-0.0056	0.0384	0.0224	78.2	-0.0078	0.0574	0.0357	81.8
		b5	-0.0062	0.0376	0.0196	75.0	-0.0103	0.0426	0.0249	79.6
400	10%	b1	0.0028	0.0242	0.0186	88.5	0.0024	0.0362	0.0268	84.7
		b2	-0.0033	0.0278	0.0186	88.6	-0.0028	0.0406	0.0285	85.0
		b5	-0.0018	0.0218	0.0163	87.5	-0.0060	0.0301	0.0193	82.7
	30%	b1	-0.0002	0.0247	0.0201	90.0	0.0005	0.0387	0.0293	87.5
		b2	-0.0029	0.0303	0.0201	88.9	-0.0004	0.0436	0.0312	87.4
		b5	-0.0030	0.0316	0.0174	86.4	-0.0055	0.0409	0.0204	85.5
	60%	b1	-0.0066	0.0580	0.0246	83.7	0.0012	0.0675	0.0353	83.0
		b2	-0.0041	0.0568	0.0250	83.3	-0.0064	0.0832	0.0370	82.6
		b5	-0.0158	0.0657	0.0208	82.9	-0.0271	0.0861	0.0233	79.5