Web-based Supplementary Materials for Drawing inferences for High-dimensional Linear Models: A Selection-assisted Partial Regression and Smoothing Approach by

Zhe Fei[1], Ji Zhu[2], Moulinath Banerjee[2], and Yi Li[1]
1. Department of Biostatistics, University of Michigan
2. Department of Statistics, University of Michigan

October 29, 2018

1 Web Appendix A

Main proofs to Theorems 1-3.

Proof of Theorem 1. Our estimator for β_j^0 by the one-time SPARE is

$$\tilde{\beta}_j = \left\{ (X_{S\cup j}^{1}{}^{\mathrm{T}} X_{S\cup j}^{1})^{-1} X_{S\cup j}^{1}{}^{\mathrm{T}} Y^1 \right\}_j . \tag{A.1}$$

Here $D_1 = (X^1, Y^1)$ with sample size $\lfloor n/2 \rfloor$, for notational simplicity, we denote $m = \lfloor n/2 \rfloor$ within this proof.

By (A3), with probability at least $1 - o(m^{-c_2-1})$, the selection $S \supset S_{0,n}$. Since the two halves of data D_1 and D_2 are mutually exclusive, $(X^1, Y^1) \perp S$. Thus given $S \supset S_{0,n}$ and

X^1, the OLS estimator $\tilde{\beta}^1 = ({X^1_{S\cup j}}^T X^1_{S\cup j})^{-1} {X^1_{S\cup j}}^T Y^1$ is unbiased,

$$
\begin{aligned}
&\mathbf{E}\left(\tilde{\beta}^1 \middle| S, X^1\right) \\
=&\mathbf{E}\left(({X^1_{S\cup j}}^T X^1_{S\cup j})^{-1} {X^1_{S\cup j}}^T X^1 \beta^0 \middle| S, X^1\right) + \mathbf{E}\left(({X^1_{S\cup j}}^T X^1_{S\cup j})^{-1} {X^1_{S\cup j}}^T X^1 \boldsymbol{\varepsilon}^1 \middle| S, X^1\right) \\
=&\mathbf{E}\left(({X^1_{S\cup j}}^T X^1_{S\cup j})^{-1} {X^1_{S\cup j}}^T X^1_{S\cup j} \beta^0_{S\cup j} \middle| S, X^1\right) + \mathbf{E}\left(\boldsymbol{\varepsilon}^1 \middle| S, X^1\right) \\
=&\beta^0_{S\cup j}.
\end{aligned} \tag{A.2}
$$

In addition, $\mathrm{Var}\left(\tilde{\beta}^1 \middle| S, X^1\right) = \sigma^2 \Sigma^{-1}_{S\cup j}/m$, which is bounded by assumption (A1). Thus,

$$\sqrt{m}(\tilde{\beta}^1 - \beta^0_{S\cup j})\Big| S, X^1 \xrightarrow{d} N(0, \sigma^2 \Sigma^{-1}_{S\cup j}). \tag{A.3}$$

Furthermore,

$$\sqrt{m}(\tilde{\beta}_j - \beta^0_j)\Big| S, X^1 \xrightarrow{d} N(0, \tilde{\sigma}^2_j), \tag{A.4}$$

where $\tilde{\sigma}^2_j = \sigma^2 \left(\Sigma^{-1}_{S\cup j}\right)_{jj}$.

Next we show the uniform convergence of $\sqrt{m}(\tilde{\beta}_j - \beta^0_j)/\tilde{\sigma}_j$ with respect to j, S and X^1. From the partial regression formulation of $\tilde{\beta}_j$, if $S \supset S_{0,n}$,

$$\tilde{\beta}_j - \beta^0_j = \frac{{X^1_j}^{\mathrm{T}}(I_m - H^1_{S\setminus j})\boldsymbol{\varepsilon}^1}{{X^1_j}^{\mathrm{T}}(I_m - H^1_{S\setminus j})X^1_j} = \frac{m}{{X^1_j}^{\mathrm{T}}(I_m - H^1_{S\setminus j})X^1_j} \frac{{X^1_j}^{\mathrm{T}}(I_m - H^1_{S\setminus j})\boldsymbol{\varepsilon}^1}{m}. \tag{A.5}$$

By Lemma (1),

$$\frac{m}{{X^1_j}^{\mathrm{T}}(I_m - H^1_{S\setminus j})X^1_j} = \left(\widehat{\Sigma}^{-1}_{S\cup j}\right)_{jj} \to \left(\Sigma^{-1}_{S\cup j}\right)_{jj}, \tag{A.6}$$

and $\forall j, S$, $\left|\frac{m}{{X^1_j}^{\mathrm{T}}(I_m - H^1_{S\setminus j})X^1_j}\right| \le 2/c_{\min}$. Moreover, the second term of the right hand side in (A.5) is the mean of i.i.d. $\tilde{x}^1_{ij}\varepsilon^1_i$'s, where $(\tilde{x}^1_{ij})_{i=1,..,m} = X^1_j(I_m - H^1_{S\setminus j})$. Since $\mathbf{E}|\boldsymbol{\varepsilon}_i|^3 \le \rho_0$ and $X^1_j(I_m - H^1_{S\setminus j})$ is the projection vector of X^1_j,

$$\mathbf{E}|X^1_j(I_m - H^1_{S\setminus j})|^3_\infty \le \mathbf{E}|X^1_j|^3_\infty \le \rho_1. \tag{A.7}$$

By the Berry-Esseen Theorem, $\forall j,\ X$ and $S \supset S_{0,n}$,

$$|F_n(x) - \Phi(x)| \leq \left(\frac{2}{c_{\min}}\right)^3 \frac{C\rho_0\rho_1}{\tilde{\sigma}_j^3\sqrt{m}} \leq \frac{8c_{\max}^{3/2}C\rho_0\rho_1}{c_{\min}^3\sigma^3\sqrt{m}}, \tag{A.8}$$

where $F_n(x)$ is the CDF of $\sqrt{m}(\tilde{\beta}_j - \beta_j^0)/\tilde{\sigma}_j$ and $\Phi(x)$ is the CDF of standard normal. Thus as $m \to \infty$, with probability at least $1 - o(m^{-c_2-1})$,

$$\sqrt{m}(\tilde{\beta}_j - \beta_j^0)/\tilde{\sigma}_j \to N(0,1). \tag{A.9}$$

□

Proof of Theorem 2. We first introduce the *oracle* SPARE estimators of β_j^0's, i.e. the ones we would compute if we knew the true active set $S_{0,n}$,

$$\hat{\beta}_j^0 = \left\{(X_{S_{0,n}\cup j}{}^T X_{S_{0,n}\cup j})^{-1} X_{S_{0,n}\cup j}{}^T Y\right\}_j \tag{A.10}$$

$$\hat{\beta}_{j,S_{0,n}}^b = \left\{(X_{S_{0,n}\cup j}^b{}^T X_{S_{0,n}\cup j}^b)^{-1} X_{S_{0,n}\cup j}^b{}^T Y^b\right\}_j, \tag{A.11}$$

which are estimations on the original data (X, Y) and the bootstrap half data D_1^b, respectively. Since $\hat{\beta}_j^0$ is the least square corresponding to X_j when regressing Y on $X_{S_{0,n}\cup j}$, we have for each j

$$W_j^0 = \sqrt{n}(\hat{\beta}_j^0 - \beta_j^0)/\sigma_j \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty, \tag{A.12}$$

where $\sigma_j^2 = \sigma^2\left(\Sigma_{S_{0,n}\cup j}^{-1}\right)_{jj}$ that corresponds to subscript j. By Cauchy's interlacing theorem (Proposition 3), $\sigma^2/c_{\max} \leq \sigma_j^2 \leq \sigma^2/c_{\min}$, and thus it is bounded away from zero and infinity.

Now we consider the behavior of the selections S^b's from D_2^b's. For each $b = 1, 2, ..., B$, the subsample D_2^b consists of $m_b \geq n/2$ distinct observations from the original data that are not drawn in the bootstrap half dataset D_1^b. In other words, D_2^b can be regarded as a sample of m_b i.i.d. observations from the population distribution. In addition, since m_b is independent of the observations, with a conditional argument on m_b, the following holds for each b by (B3),

$$\begin{aligned}
&\mathbf{P}(S^b = S_{0,n}) \\
=&\int \mathbf{P}(S^b = S_{0,n}|m_b = m)\mathrm{d}\mathbf{P}(m) \\
\geq&\int \left\{1 - o(m^{-c_2-1})\right\}\mathrm{d}\mathbf{P}(m) \\
\geq&1 - o\{(n/2)^{-c_2-1}\} \\
=&1 - o(n^{-c_2-1}).
\end{aligned} \tag{A.13}$$

Next, we decompose $\hat{\beta}_j$ into two parts:

$$\begin{aligned}\hat{\beta}_j =&\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_j^b\\ =&\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{j,S_{0,n}}^b+\frac{1}{B}\sum_{b:S^b\neq S_{0,n}}\Big(\hat{\beta}_j^b-\hat{\beta}_{j,S_{0,n}}^b\Big),\end{aligned}\tag{A.14}$$

and equivalently

$$\begin{aligned}&\sqrt{n}(\hat{\beta}_j-\beta_j^0)\\ =&\sqrt{n}\Big(\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{j,S_{0,n}}^b-\beta_j^0\Big)+\frac{\sqrt{n}}{B}\sum_{b:S^b\neq S_{0,n}}\Big(\hat{\beta}_j^b-\hat{\beta}_{j,S_{0,n}}^b\Big)\\ \doteq&Z_j^0+\Delta_j.\end{aligned}\tag{A.15}$$

To show $\Delta_j=o_p(1)$, we write

$$\Delta_j=\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(S^b\neq S_{0,n})\sqrt{n}\Big(\hat{\beta}_j^b-\hat{\beta}_{j,S_{0,n}}^b\Big);\tag{A.16}$$

$$\Delta_j=\frac{1}{B}\sum_{b=1}^{B}\delta_b;\quad \delta_b\doteq\mathbf{1}(S^b\neq S_{0,n})\sqrt{n}\Big(\hat{\beta}_j^b-\hat{\beta}_{j,S_{0,n}}^b\Big).\tag{A.17}$$

By Corollary (2),

$$\begin{aligned}\mathbf{E}\delta_b =&\mathbf{P}(S^b\neq S_{0,n})\mathbf{E}\sqrt{n}\Big(\hat{\beta}_j^b-\hat{\beta}_{j,S_{0,n}}^b\Big)\\ =&o\Big(n^{-c_2-1}2C_\beta n^{c_1+\frac{1}{2}}\Big)\\ =&o\Big(n^{-c_2+c_1-\frac{1}{2}}\Big)\\ \to&0\quad\text{as}\quad n\to\infty.\end{aligned}\tag{A.18}$$

Similarly,

$$\begin{aligned}\mathbf{Var}\delta_b =&\mathbf{P}(S^b\neq S_{0,n})\mathbf{E}n\Big(\hat{\beta}_j^b-\hat{\beta}_{j,S_{0,n}}^b\Big)^2\\ =&o\Big(n^{-c_2-1}4C_\beta^2 n^{2c_1+1}\Big)\\ =&o(n^{-c_2+2c_1})\\ \to&0\quad\text{as}\quad n\to\infty.\end{aligned}\tag{A.19}$$

Thus $\delta_b = o_p(1)$ for all $b \in [B]$. Furthermore, since $\mathbf{E}\Delta_j = \mathbf{E}\delta_b$ and $\mathbf{Var}\Delta_j \leq \mathbf{Var}\delta_b$, we have $\Delta_j = o_p(1)$.

Next, we show the convergence of Z_j^0. Notice that

$$Z_j^0/\sigma_j = W_j^0 + \sqrt{n}\Big(\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{j,S_{0,n}}^b - \hat{\beta}_j^0\Big)/\sigma_j \doteq W_j^0 + T_n^B/\sigma_j. \tag{A.20}$$

By (A.12), we are only left to show $T_n^B = o_p(1)$. Define $t_{n,b} = \sqrt{n}(\hat{\beta}_{j,S_{0,n}}^b - \hat{\beta}_j^0)$, then $T_n^B = \sqrt{n}(\frac{1}{B}\sum_{b=1}^B \hat{\beta}_{j,S_{0,n}}^b - \hat{\beta}_j^0) = \frac{1}{B}\sum_{b=1}^B t_{n,b}$. Recall that $\hat{\beta}_{j,S_{0,n}}^b$ is the bootstrap statistic of $\hat{\beta}_j^0$, so its conditional mean is $\hat{\beta}_j^0$ and conditional variance is $\hat{\sigma}^2\Big\{(X_{S_{0,n}\cup j}^T X_{S_{0,n}\cup j})^{-1}\Big\}_{jj} = \hat{\sigma}^2\Big(\widehat{\Sigma}_{S_{0,n}\cup j}^{-1}\Big)_{jj}/n \doteq \hat{\sigma}_j^2/n$, where $\hat{\sigma}^2 = \|(\mathrm{I}_n - H_{S_{0,n}})Y\|_2^2/n$ (Freedman et al. (1981)). Thus, conditional on the data, $\{t_{n,b}\}_{b=1,2,..,B}$ are i.i.d. with

$$\mathbf{E}\big(t_{n,b}|(X^{(n)},Y^{(n)})\big) = 0, \quad \mathbf{Var}\big(t_{n,b}|(X^{(n)},Y^{(n)})\big) = \hat{\sigma}_j^2 = \hat{\sigma}^2\Big(\widehat{\Sigma}_{S_{0,n}\cup j}^{-1}\Big)_{jj}. \tag{A.21}$$

We now argue that with probability going to 1, $\hat{\sigma}_j^2$'s, $j = 1,2,..,p$, are bounded. First, $\mathbf{P}(\hat{\sigma}^2 < 2\sigma^2) \to 1$ as $n \to \infty$. Then,

$$\Big(\widehat{\Sigma}_{S_{0,n}\cup j}^{-1}\Big)_{jj} \leq \lambda_{\max}(\widehat{\Sigma}_{S_{0,n}\cup j}^{-1}) = 1/\lambda_{\min}(\widehat{\Sigma}_{S_{0,n}\cup j}), \tag{A.22}$$

whenever $\lambda_{\min}(\widehat{\Sigma}_{S_{0,n}\cup j}) > 0$. Assumption (B3) implies $|S_{0,n}|/n \leq \eta$. By Lemma (4) from Vershynin (2010) and Lemma (5), letting $\epsilon = c_{\min}/2$ and $t^2 = c_{\min}^2\eta/C$ for some constant C only depending on the sub-Gaussian norm $\|\mathbf{x}_i\|_{\psi_2}$, we have that with probability at least $1 - 2\exp(-c_{\min}^2\eta n^{\gamma_0}/C)$

$$\lambda_{\min}(\widehat{\Sigma}_{S_{0,n}\cup j}) \geq \lambda_{\min}(\Sigma_{S_{0,n}\cup j}) - c_{\min}/2 \geq \lambda_{\min}(\Sigma) - c_{\min}/2 \geq c_{\min}/2, \tag{A.23}$$

where the second inequality follows the interlacing property of the eigenvalues. Combining (A.22) and (A.23), $\Big(\widehat{\Sigma}_{S_{0,n}\cup j}^{-1}\Big)_{jj} \leq 2/c_{\min}$ with probability going to 1 exponentially fast in n, and consequently $\hat{\sigma}_j^2 < 4\sigma^2/c_{\min}$. Now define

$$\Omega_n = \{(X^{(n)},Y^{(n)}) = (\mathbf{x}_i, y_i)_{i=1,2,..,n} : \hat{\sigma}_j^2 < 4\sigma^2/c_{\min}, \forall j = 1,2,...,p\}. \tag{A.24}$$

Since $p = O(n^{\gamma_1})$ for some $\gamma_1 > 1$, $\mathbf{P}\{(X^{(n)}, Y^{(n)}) \in \Omega_n\} \to 1$ as $n \to \infty$. Thus $\forall (X^{(n)}, Y^{(n)}) \in \Omega_n$, $\mathbf{Var}\{t_{n,b}|(X^{(n)}, Y^{(n)})\} \leq 4\sigma^2/c_{\min}$. Furthermore,

$$\mathbf{Var}\{T_n^B|(X^{(n)}, Y^{(n)})\} = \frac{1}{B^2}\sum_{b=1}^{B}\mathbf{Var}\{t_{n,b}|(X^{(n)}, Y^{(n)})\} \leq \frac{4\sigma^2}{Bc_{\min}} \tag{A.25}$$

Thus, $\forall \delta, \zeta > 0$, $\exists N_0, B_0 > 0$ such that $\forall n > N_0, B > B_0$,

$$\begin{aligned}
&\mathbf{P}(|T_n^B| \geq \delta) \\
\leq &\int_{\Omega_n} \mathbf{P}\left\{|T_n^B| \geq \delta \big| (X^{(n)}, Y^{(n)})\right\} \mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\} \\
\leq &\int_{\Omega_n} \frac{\mathbf{Var}\left\{T_n^B|(X^{(n)}, Y^{(n)})\right\}}{\delta^2} \mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\} \\
\leq &\frac{4\sigma^2}{B_0\delta^2 c_{\min}} \int_{\Omega_n} \mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\} \\
\leq &\zeta/2 + \zeta/2 \\
\leq &\zeta.
\end{aligned} \tag{A.26}$$

Finally, combining this with (A.12), we have

$$Z_j^0/\sigma_j = W_j^0 + T_n^B/\sigma_j \xrightarrow{d} N(0, 1) \quad \text{as} \quad B, n \to \infty. \tag{A.27}$$

□

Proof of Theorem 3. Follow the previous proof, we replace the arguments in j with those in $S^{(1)}$. The *oracle* estimators are

$$\hat{\beta}^0_{S^{(1)}} = \Big((X_{S_{0,n}\cup S^{(1)}}{}^T X_{S_{0,n}\cup S^{(1)}})^{-1} X_{S_{0,n}\cup S^{(1)}}{}^T Y\Big)_{S^{(1)}} \tag{A.28}$$

$$\hat{\beta}^b_{S^{(1)},S_{0,n}} = \Big((X^b_{S_{0,n}\cup S^{(1)}}{}^T X^b_{S_{0,n}\cup S^{(1)}})^{-1} X^b_{S_{0,n}\cup S^{(1)}}{}^T Y^b\Big)_{S^{(1)}}. \tag{A.29}$$

Notice that $|S^{(1)}| = p_1 = O(1)$, as $n \to \infty$, $|S_{0,n} \cup S^{(1)}| = O\big(|S_{0,n}|\big) = o(n)$, so that the above quantities are well-defined. Next

$$W^{(1)} = \sqrt{n}\{\Sigma^{(1)}\}^{-1}(\hat{\beta}^0_{S^{(1)}} - \beta^0_{S^{(1)}}) \xrightarrow{d} N(0, \mathrm{I}_{p_1}) \quad \text{as} \quad n \to \infty, \tag{A.30}$$

where $\Sigma^{(1)} = \sigma^2\Big(\Sigma^{-1}_{S_{0,n}\cup S^{(1)}}\Big)_{S^{(1)}}$. Similar to (A.15), we decompose $\sqrt{n}(\hat{\beta}_{S^{(1)}} - \beta^0_{S^{(1)}})$ into three parts:

$$\begin{aligned}&\sqrt{n}(\hat{\beta}_{S^{(1)}} - \beta^0_{S^{(1)}})\\ \doteq& Z^{(1)} + \Delta_0^{(1)} + \Delta_1^{(1)}.\end{aligned} \tag{A.31}$$

For the sake of space, we prefer not to write out these quantities, but it is straightforward analog that $\Delta_0^{(1)} = \Delta_1^{(1)} = o_p(\mathbf{1}_{p_1})$ and ${\Sigma^{(1)}}^{-1} Z^{(1)} - W^{(1)} = o_p(\mathbf{1}_{p_1})$ as well, which completes the proof. □

2 Web Appendix B

Technical details on useful definitions, lemmas and related proofs.

Lemma 1. Assume $X = (X_1, ..., X_p) = (x_1^{\mathrm{T}}, ..., x_n^{\mathrm{T}})^{\mathrm{T}}$ where x_i's are i.i.d. copies of a sub-Gaussian random vector in $\mathbf{R}^p$ with covariance matrix $\Sigma_{p\times p}$, with

$$0 < c_{\min} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le c_{\max} < \infty.$$

For any subset $S \subset \{1, 2, .., p\}$ with $|S| \le \eta n$, $0 < \eta < 1$, and $\forall j \in S$, with probability at least $1 - 2\exp(-\frac{\varepsilon^2\eta}{C_K} n)$,

$$\frac{c_{\min}}{2} \le \frac{1}{n} X_j^{\mathrm{T}}(I_n - H_{S\setminus j})X_j \le c_{\max} + \frac{1 + c_{\min}}{2} \tag{B.1}$$

where $\varepsilon = \min(\frac{1}{2}, \frac{c_{\min}}{2})$ and C_K is the constant depends only on the sub-Gaussian norm $K = \|x_i\|_{\psi_2}$.

Corollary 2. Given model (1) and assumptions (A1,A2), consider the partial regression estimator on (X, Y) given subset S. If $|S| \le \eta n$, $0 < \eta < 1$, then with probability at least $1 - 2\exp(-\frac{\varepsilon^2\eta}{C_K} n)$,

$$\hat{\beta}_j \le C_\beta n^{c_1}, \tag{B.2}$$

where C_β depends on $c_{\min}, c_{\max}, c_\beta$.

Proposition 3 (Cauchy interlacing theorem). Let A be a symmetric $n\times n$ matrix. The $m\times m$ matrix B, where $m \le n$, is called a compression of A if there exists an orthogonal projection P onto a subspace of dimension m such that $P^{\mathrm{T}}AP = B$. The Cauchy interlacing theorem states:

if the eigenvalues of A are $\lambda_1 \le ... \le \lambda_n$, and those of B are $\nu_1 \le ... \le \nu_m$, then for all $j < m + 1$,

$$\lambda_j \le \nu_j \le \lambda_{n-m+j}$$

Proposition 4 (Corollary 5.50 in Vershynin (2010)). Consider a $n \times q$ matrix X whose rows $\mathbf{x}_i$'s are i.i.d. samples from a sub-Gaussian distribution in R^q with covariance matrix Σ, and let $\epsilon \in (0,1), t \geq 1$. Denote the sample covariance matrix as $\widehat{\Sigma}_n = X^{\mathrm{T}}X/n$ Then with probability at least $1 - 2\exp(-t^2 q)$ one has

$$\textit{If} \quad n \geq C(t/\epsilon)^2 q \quad \textit{then} \quad \|\widehat{\Sigma}_n - \Sigma\| \leq \epsilon. \tag{B.3}$$

Here $C = C_K$ depends only on the sub-Gaussian norm $K = \|\mathbf{x}_i\|_{\psi_2}$ of a random vector taken from this distribution.

Definition 1. The sub-Gaussian norm of a random variable V is defined as

$$\|V\|_{\psi_2} = \sup_{k \geq 1} k^{-1/2}(E|V|^k)^{1/k} \tag{B.4}$$

then the sub-Gaussian norm of a random vector V in R^q is defined as

$$\|V\|_{\psi_2} = \sup_{x \in S^{q-1}} \|V^{\mathrm{T}}x\|_{\psi_2} \tag{B.5}$$

Remark 1. Assume $V_0 = (v_1, v_2, ..., v_q)$ is a sub-Gaussian random vector in R^q, and $V_1 = (v_1, v_2, ..., v_r), r < q$ is the sub-vector of V_0. By taking $x = (x_1, .., x_r, 0, .., 0) \in S^{q-1}$, we have $\|V_1\|_{\psi_2} \leq \|V_0\|_{\psi_2}$.

Corollary 5. For two $n \times n$ positive definite matrices Σ_1 and Σ_2, if $\|\Sigma_1 - \Sigma_2\| \leq \epsilon$, then

$$\begin{aligned} \lambda_{\min}(\Sigma_2) &\geq \lambda_{\min}(\Sigma_1) - \epsilon \\ \lambda_{\max}(\Sigma_2) &\leq \lambda_{\max}(\Sigma_1) + \epsilon. \end{aligned} \tag{B.6}$$

Proof. On one hand, $\forall n-$vector X with $\|X\|_2 = 1$,

$$\begin{aligned} \epsilon \geq& \|\Sigma_1 - \Sigma_2\| \\ \geq& \|(\Sigma_1 - \Sigma_2)X\|_2 \\ \geq& \|\Sigma_1 X\|_2 - \|\Sigma_2 X\|_2 \end{aligned} \tag{B.7}$$

then take X to be the eigenvector for $\lambda_{\min}(\Sigma_2)$, we have

$$\begin{aligned} \lambda_{\min}(\Sigma_2) =& \|\Sigma_2 X\|_2 \\ \geq& \|\Sigma_1 X\|_2 - \epsilon \\ \geq& \lambda_{\min}(\Sigma_1) - \epsilon. \end{aligned} \tag{B.8}$$

On the other hand,

$$\begin{aligned}\lambda_{\max}(\Sigma_2) =& \|\Sigma_2\| \\ \leq& \|\Sigma_1\| + \|\Sigma_2 - \Sigma_1\| \\ \leq& \|\Sigma_1\| + \epsilon \\ =& \lambda_{\max}(\Sigma_1) + \epsilon\end{aligned} \tag{B.9}$$

□

Proof of lemma (1). Note that

$$\frac{n}{X_j^T(I_n - H_{S\backslash j})X_j}$$

is the $(j,j)^{\text{th}}$ entry of $\widehat{\Sigma}_S^{-1}$, where $\widehat{\Sigma}_S = (X_S^T X_S)/n$ is the sample covariance matrix corresponds to subset S. Therefore

$$\frac{1}{\lambda_{\max}(\widehat{\Sigma}_S)} \leq \frac{n}{X_j^T(I_n - H_{S\backslash j})X_j} \leq \frac{1}{\lambda_{\min}(\widehat{\Sigma}_S)}. \tag{B.10}$$

Refer to Corollary 5.50 in Vershynin (2010) and choose $\varepsilon = \min(\frac{1}{2}, \frac{c_{\min}}{2})$. Then with probability at least $1 - 2\exp(-\frac{\varepsilon^2 \eta}{C_K} n)$,

$$\|\widehat{\Sigma}_S - \Sigma_S\| \leq \varepsilon. \tag{B.11}$$

By Corollary (5) and Cauchy interlacing theorem,

$$\lambda_{\min}(\widehat{\Sigma}_S) \geq \lambda_{\min}(\Sigma_S) - \varepsilon \geq \lambda_{\min}(\Sigma) - \varepsilon \geq c_{\min}/2, \tag{B.12}$$

and

$$\lambda_{\max}(\widehat{\Sigma}_S) \leq \lambda_{\max}(\Sigma_S) + \varepsilon \leq \lambda_{\max}(\Sigma) + \varepsilon \leq c_{\max} + (1 + c_{\min})/2. \tag{B.13}$$

Thus, with high probability,

$$\frac{c_{\min}}{2} \leq \frac{1}{n} X_j^T(I_n - H_{S\backslash j})X_j \leq c_{\max} + \frac{1 + c_{\min}}{2} \tag{B.14}$$

□

Proof of Corollary (2). From Lemma (1), we can bound $\hat{\beta}_j$ as below:

$$\begin{aligned}\hat{\beta}_j =& \frac{X_j^{\mathrm{T}}(I - H_{S\setminus j})Y}{X_j^{\mathrm{T}}(I - H_{S\setminus j})X_j} \\ =& \frac{n}{X_j^{\mathrm{T}}(I - H_{S\setminus j})X_j} \frac{X_j^{\mathrm{T}}(I - H_{S\setminus j})X_{S_{0,n}}\beta^0_{S_{0,n}}}{n} \\ \leq& \frac{2}{c_{\min}} \frac{c_\beta \sum_{k\in S_{0,n}} |X_j^{\mathrm{T}}(I - H_{S\setminus j})X_k|}{n} \\ \leq& \frac{2}{c_{\min}} c_\beta \big(c_{\max} + \frac{1 + c_{\min}}{2}\big) n^{c_1}.\end{aligned} \tag{B.15}$$

Let $C_\beta = \frac{2c_\beta}{c_{\min}}\big(c_{\max} + \frac{1+c_{\min}}{2}\big)$, we complete the proof. □

References

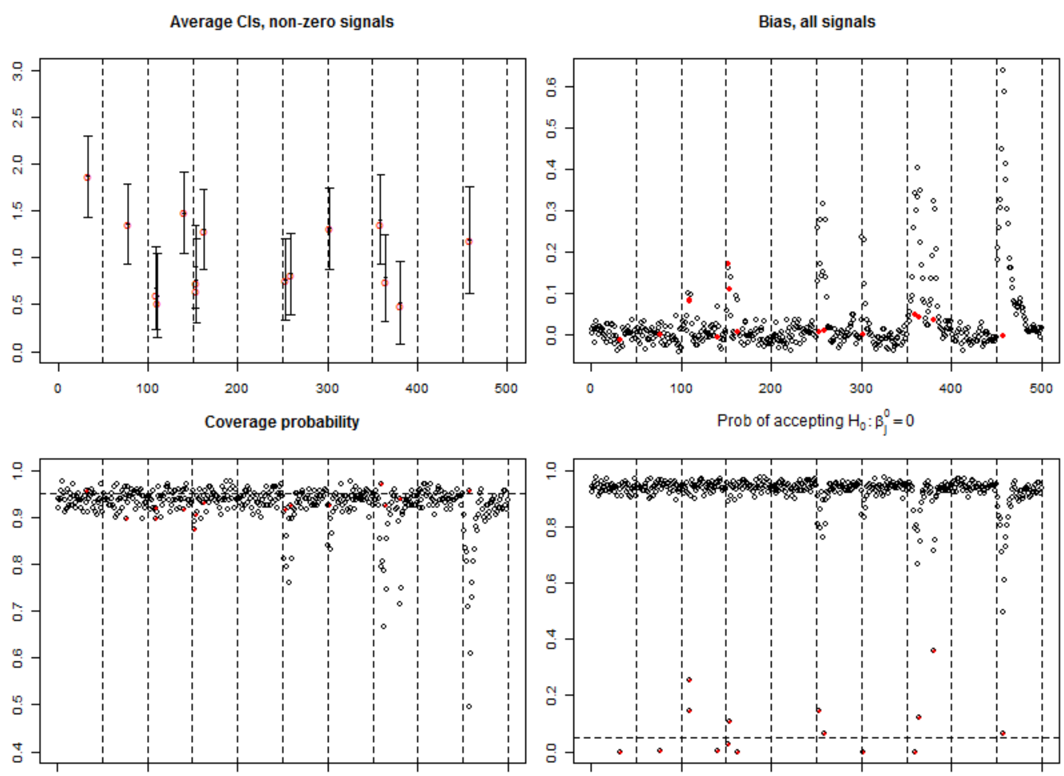
Freedman, D. A. et al. (1981). Bootstrapping regression models. *The Annals of Statistics 9*(6), 1218–1228.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

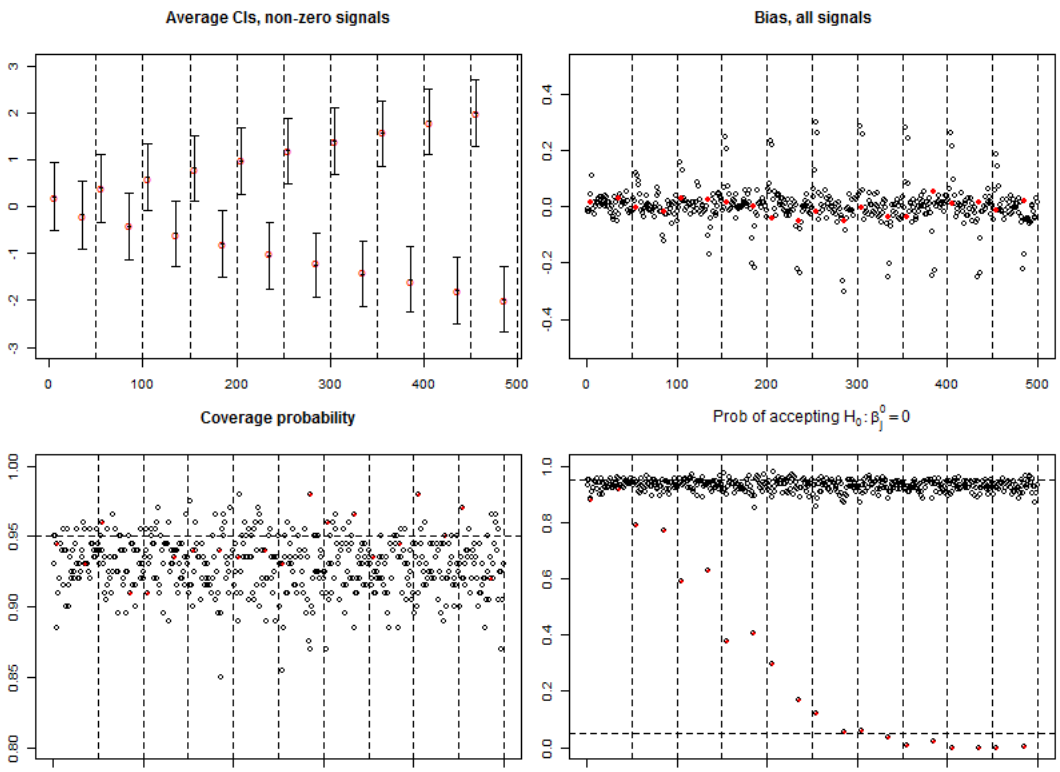
Web Table 1: Comparisons of SPARES and one-time SPARE based on 200 replications. Bias (SE) is displayed in each cell. LSE refers to least square estimation as if $S_{0,n}$ were known.

Index	β_j^0	SPARES	One-time SPARE	LSE
199	1.00	0.03(0.16)	-0.02(0.26)	0.03(0.16)
243	-1.00	-0.02(0.16)	0.03(0.26)	-0.02(0.16)
256	1.00	-0.002(0.16)	-0.007(0.26)	-0.002(0.16)
0's	0.00	0.000(0.16)	-0.001(0.26)	

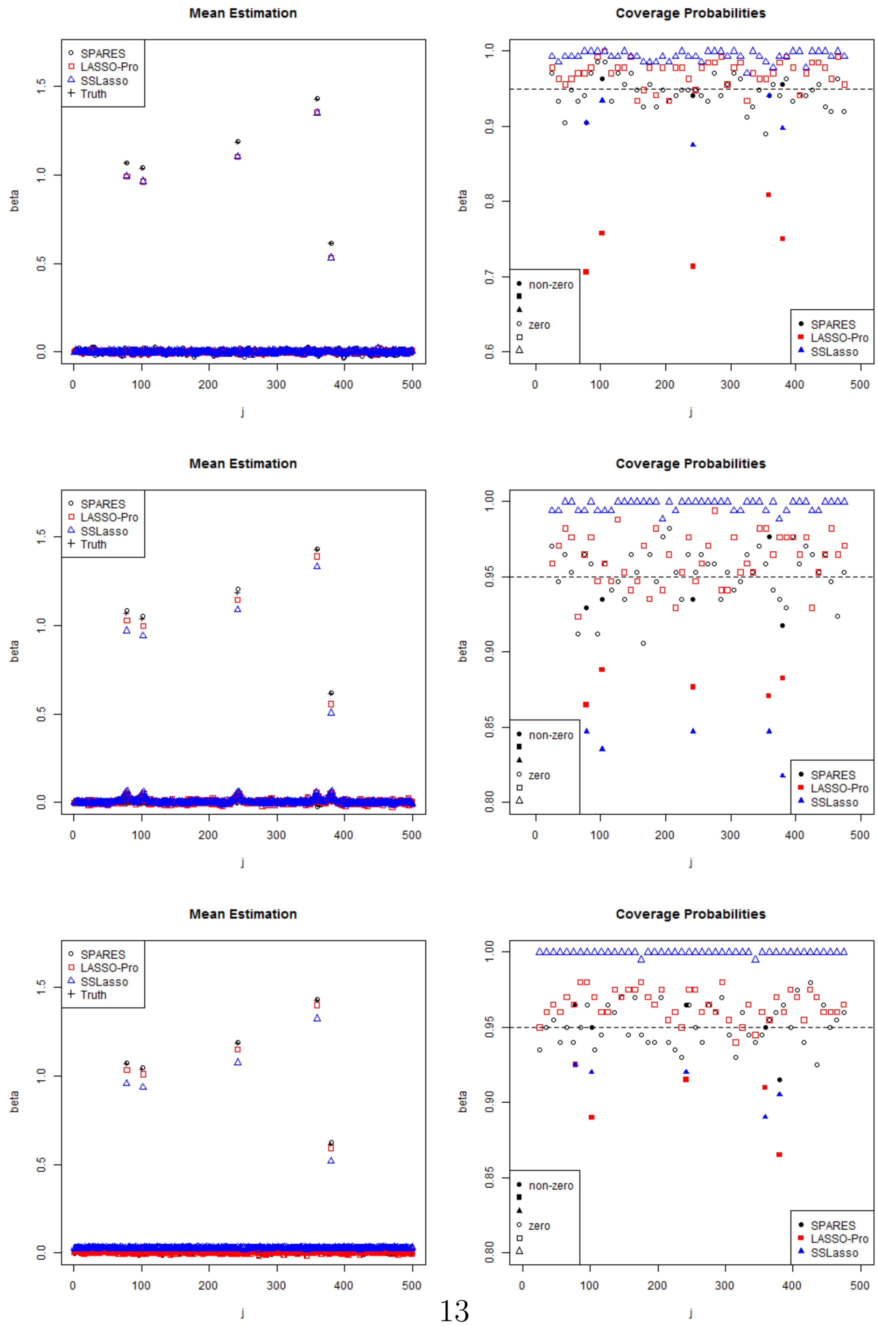
Web Figure 1: Performance of SPARES under simulation example 2.1. X-axis is the variable index. **Topleft:** Average estimates and average CIs V.S. true signals. **Topright:** Bias of SPARES estimates for each j, red dots are non-zero signals, dashed lines indicate blocks of the predictors. **Bottomleft:** Coverage probability of β^0 for each j w.r.t. 0.95 norminal level. **Bottomright:** Empirical probability of not rejecting $H_0 : \beta_j^0 = 0$.



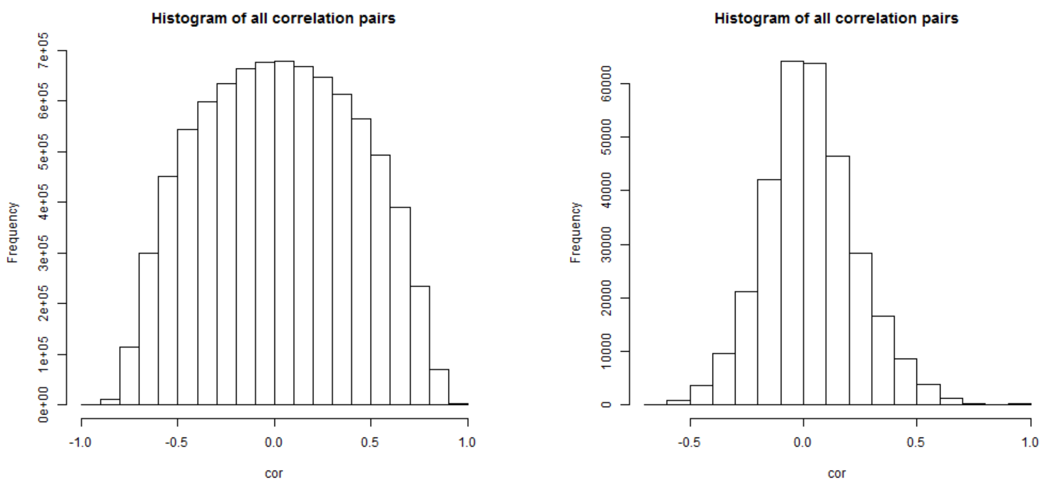
Web Figure 2: Performance of SPARES under simulation examples 2.2.



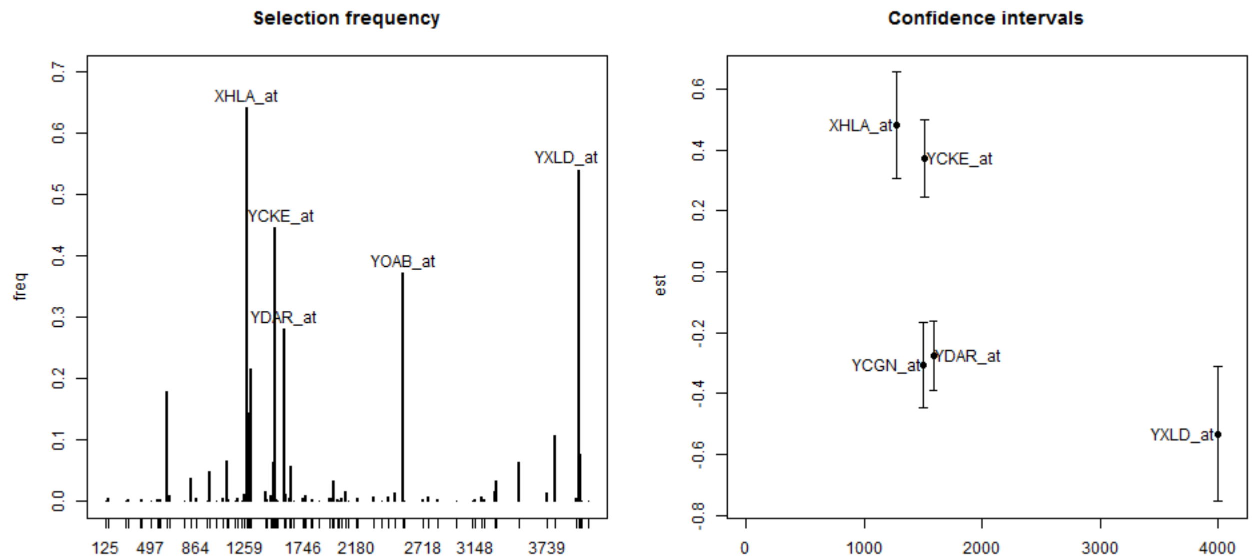
Web Figure 3: Comparisons of SPARES with LASSO-Pro and SSLASSO under simulation example 4. Left panels: Mean estimates from each method and the true signals. Right panels: Coverage probabilities for each $j \in S_{0,n}$ and 20 representatives of $j \notin S_{0,n}$.



Web Figure 4: Correlation among predictors: left panel - riboflavin data; right panel - multiple myeloma data.



Web Figure 5: Results of the riboflavin genomic data analysis. Left panel: selection frequency of each gene; Right panel: confidence intervals of the top five most significant genes.



Web Figure 6: Results of the Multiple Myeloma genomic data analysis. Left panel: selection frequency of each gene; Right panel: confidence intervals of the top two most significant genes.