Measures of Uncertainty for Shrinkage Model Selection

Yuanyuan Li and Jiming Jiang

Department of Statistics, University of California, Davis, U.S.A.

Supplementary Material

S1 EO when coefficients paths cross zero

As λ decreases, some non-zero coefficients could be zero again, which happens when some variables are highly correlated so that a later entering variable changes the coefficient path of an early entering one. In this situation, the order of those highly correlated variables in EO actually doesn't matter too much since their orders could just be formed by chance. To protect the nested property of NMCS, we have modified the steps of constructing the EO by separating it into two steps. First, we find the EO of the variables in the selected model $\hat{M}$, and the orders are defined by their first time of becoming non-zero when λ decreases from a large value to $\hat{\lambda}$; then we decide the entering order of the rest variables based on a grid of λ values decreasing from $\hat{\lambda}$ to 0, and add these variables to EO in order. Note that when λ is zero, Lasso estimates will be the same as the least square

estimates so that all variables will be active.

Overall, this modification doesn't change the original EO (defined by the initial entering order of each variable as λ decreases) most of the time. It can be seen that only when a variable enters the active set and stays inactive until λ decreases to the selected $\hat{\lambda}$, the order of this variable changes (increases) to the order where it re-enters the model when λ further decreases from $\hat{\lambda}$ to 0. Such an order change event rarely happens; for example, it did not happen at all in our simulation studies and real data analyses, although, in theory, it can still happen. Specifically, the order change will never happen if Lasso paths are monotone. Efron, Hastie and Tibshirani (2004) provides a necessary and sufficient condition for the monotonicity of the Lasso paths: let X_A denote a subset of A columns of the design matrix X, each multiplied by a set of arbitrary signs $s_1, s_2, ...s_{|A|}$, and let S_A be a diagonal matrix of the s_j values,

$$S_A(X_A^T X_A)^{-1} S_A \mathbf{1} \geq 0, \ \forall A \in 1, \cdots, p, S_A. \tag{S1.1}$$

where the inequality is understood as element-wise. This means that for all subsets of predictors, the inverse covariance matrix is always diagonally dominant, i.e., each diagonal element in the inverse covariance matrix is at least as large as the sum of the other elements in its row. This is a weak assumption about the correlation of the covariates in X. See Hastie et al.

(2007) for more discussions about the monotonicity of the Lasso paths.

To summarize, the order change of EO brings some complexity of obtaining EO empirically. We modified the EO computation in this section to accommodate this special case, but still keeps the necessary structure of EO and NMCS. It can be seen that the theoretical property of NMCS in Theorem 1 is still valid even when order changes happen. In most cases when covariates are not highly correlated, the order change will not happen, and the modified EO is the same as original EO, which is defined by the first time a variable enters model.

S2 Proof of Proposition 1

Under the conditions above, the solution to the LASSO problem (2.2) is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)\left(|\hat{\beta}_j^0| - \frac{\lambda}{2n}\right)^+, \tag{S2.1}$$

$j = 1, \cdots, p$, where $\hat{\beta}^0 = (\hat{\beta}_1^0, \cdots, \hat{\beta}_p^0)'$ is the ordinary least square estimator of β, and $x^+ = \max(x, 0)$. Without loss of generality, let $\sigma^2 = 1$, $\{\hat{\beta}_i^0\}_{i=1,\cdots,p}$ are independent random variables with distribution $N(\beta_i, n^{-1})$. We next prove that the NMCS based on Lasso satisfies assumption *A2*.

By Portnoy (1984), if $p(\log p)/n \to 0$, $||\hat{\beta}^0 - \beta||^2 = O_{\text{P}}(p/n)$. Hence, for any $\varepsilon > 0$, there is a constant $M > 0$ and constant $N > 0$ such that for

$n \geq N$,

$$\begin{aligned} 1-\varepsilon &\leq \mathrm{P}\Big(||\hat{\beta}^0-\beta||^2 \leq \frac{Mp}{n}\Big) \leq \mathrm{P}\Big(\max_{0\leq j\leq p}|\hat{\beta}_j^0-\beta_j| \leq \sqrt{\frac{Mp}{n}}\Big) \\ &= \mathrm{P}\Big(\bigcap_{0\leq j\leq p}\Big\{|\hat{\beta}_j^0-\beta_j| \leq \sqrt{\frac{Mp}{n}}\Big\}\Big) \end{aligned} \tag{S2.2}$$

Let $z_1 = \min_{i=1,\cdots,q}|\hat{\beta}_i^0|$, $z_2 = \max_{q+1,\cdots,p}|\hat{\beta}_i^0|$. Since $\sqrt{Mp/n}$ $< 1/2\min_{i=1,\cdots,q}|\beta_i|$ for large n, we get $\mathrm{P}(z_1 > z_2) \to 1$ as $n \to \infty$. By (S2.1), we can get case-by-case selection results according to the value of λ. If $\lambda \geq 2nz_1$, an underfitting is chosen; $2nz_1 > \lambda \geq 2nz_2$, the optimal model $(\beta_1, \cdots, \beta_q \neq 0)$ is chosen; $2nz_2 > \lambda$, an overfitting model will be chosen. Denotes the left side of (3.11) in assumption *A2* as $\triangle$, and $\psi_{\text{opt}} = (\beta_{(q)}^T, 0^T)^T$, where $\beta_{(q)}$ is q-vector with nonzero coefficients.

For $w = 0$, $j = 0$, any fixed vector $\tilde{\psi}_{\text{opt}} = (\tilde{\beta}_{(q)}^T, 0^T)^T$, we have $\mathrm{P}(M_{\text{opt}} = \hat{M}^*|M_{\text{opt}}, \tilde{\psi}_{\text{opt}}) = \mathrm{P}(2nz_1 > \lambda \geq 2nz_2\big|M_{\text{opt}}, \tilde{\psi}_{\text{opt}}) = \mathrm{P}(2nz_1 > \lambda|\big|M_{\text{opt}}, \tilde{\psi}_{\text{opt}})$

$\times \mathrm{P}(\lambda \geq 2nz_2 \big| M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}})$.

$$\begin{aligned}
\mathrm{P}(2nz_1 > \lambda \big| M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}) &= \prod_{l=1}^{q} \Big(\mathrm{P}(\sqrt{n}(\hat{\beta}_i^0 - \tilde{\beta}_i) > \lambda/(2\sqrt{n}) - \sqrt{n}\tilde{\beta}_i) \\
&\quad + \mathrm{P}(\sqrt{n}(\hat{\beta}_q^0 - \tilde{\beta}_q) < -\lambda/(2\sqrt{n}) - \sqrt{n}\tilde{\beta}_q) \Big) \\
&= \prod_{l=1}^{q} \Big(1 - \Phi(\lambda/(2\sqrt{n}) - \sqrt{n}\tilde{\beta}_i) + \Phi(-\lambda/(2\sqrt{n}) \\
&\quad - \sqrt{n}\tilde{\beta}_i) \Big) \\
&= 1 - \Phi_q\Big(\frac{\lambda}{2\sqrt{n}} - \sqrt{n}\tilde{\beta}_{(q)}) + \Phi_q(-\lambda/(2\sqrt{n}) \\
&\quad - \sqrt{n}\tilde{\beta}_{(q)} \Big), \qquad \text{(S2.3)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}(\lambda \geq 2nz_2 \big| M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}) &= \mathrm{P}(\lambda/(2\sqrt{n}) \geq \sqrt{n}|\hat{\beta}_{q+1}^0| \big| M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}) \cdots \\
&\quad \mathrm{P}(\lambda/(2\sqrt{n}) \geq \sqrt{n}|\hat{\beta}_p^0| \big| M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}) \\
&= \prod_{l=q+1}^{p} \mathrm{P}(-\lambda/(2\sqrt{n}) \leq \sqrt{n}\hat{\beta}_l^0 \leq \lambda/(2\sqrt{n})) \\
&= \prod_{l=q+1}^{p} \big(\Phi(\lambda/(2\sqrt{n})) - \Phi(-\lambda/(2\sqrt{n})) \big) \\
&\leq 1. \qquad \text{(S2.4)}
\end{aligned}$$

Then we have

$$\begin{aligned}
\triangle &\leq |\Phi_q(\lambda/(2\sqrt{n}) - \sqrt{n}\tilde{\beta}_{(q)}) - \Phi_q(\lambda/(2\sqrt{n}) - \sqrt{n}\beta_{(q)})| \\
&\quad + |\Phi_q(-\lambda/(2\sqrt{n}) - \sqrt{n}\tilde{\beta}_{(q)}) - \Phi_q(-\lambda/(2\sqrt{n}) - \sqrt{n}\beta_{(q)})| \\
&= |d_+| + |d_-|. \qquad \text{(S2.5)}
\end{aligned}$$

By Taylor expansion, $d_\pm = \sqrt{n}\phi_q(\pm\lambda/(2\sqrt{n}) - \sqrt{n}\beta_{(q)\pm})^T(\tilde{\beta}_{(q)} - \beta_{(q)})$, where

$\phi(\cdot)$ is the pdf of $N(0,1)$ and applied element-wise, and $\beta_{(q)\pm}$ is between $\beta_{(q)}$ and $\tilde{\beta}_{(q)}$. By $\min_{i=1,\cdots,q}|\beta_i| \geq M_0 n^{\frac{c_1-1}{2}}$, and $\lambda = O(\sqrt{n})$. For large n, we get $|\pm\lambda/(2\sqrt{n}) - \sqrt{n}\beta^{(i)}_{(q)\pm}| \geq M_1 n^{\frac{c_1}{2}} > \sqrt{\log n}$, M_1 is a positive constant. Thus, $\phi_q(\pm\lambda/(2\sqrt{n}) - \sqrt{n}\beta^{(i)}_{(q)\pm}) = \frac{1}{\sqrt{(2\pi)^q}} \exp\left\{ - \frac{\left(\pm\lambda/(2\sqrt{n}) - \sqrt{n}\beta^{(i)}_{(q)\pm}\right)^2}{2} \right\} \leq \frac{1}{\sqrt{(2\pi)^q n}}$ for large n. It follows that $|d_\pm| \leq |\tilde{\beta}_{(q)} - \beta_{(q)}|/\sqrt{(2\pi)^q} \leq |\tilde{\beta} - \beta|/\sqrt{(2\pi)^q}$, assumption *A2* is satisfied.

For any $0 < w \leq 2p$, $j = 0, \cdots, w$, we have

$$
\begin{aligned}
& \mathrm{P}\Big[M_{\mathrm{opt}} \in (\hat{M}^*_{-w+j}, \hat{M}^*_j) | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\Big] \\
= & \sum_{h=-w+j}^{j} \mathrm{P}\big(M_{\mathrm{opt}} = \hat{M}^*_h | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big) \\
= & \sum_{h=-w+j}^{j} \Big[\mathrm{P}\big(M_{\mathrm{opt}} = \hat{M}^*_h, \hat{M}^* = M_{\{\beta=\mathbf{0}\}} | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big) \\
& + \mathrm{P}\big(M_{\mathrm{opt}} = \hat{M}^*_h, \hat{M}^* = M_{\{\beta_1 \neq 0\}} | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big) + \cdots \\
& + \mathrm{P}\big(M_{\mathrm{opt}} = \hat{M}^*_h, \hat{M}^* = M_{\{\beta_1, \cdots, \beta_p \neq 0\}} | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big)\Big] \\
= & \sum_{h=-w+j}^{j} \Big[\mathrm{P}\big(\hat{M}^* = M_{(\beta=\mathbf{0})} | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big) \times \\
& \mathrm{P}\big(M_{\mathrm{opt}} = \hat{M}^*_h | \hat{M}^* = M_{\{\beta=\mathbf{0}\}}, M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big) \\
& + \cdots + \mathrm{P}\big(\hat{M}^* = M_{\{\beta_1, \cdots, \beta_p \neq 0\}} | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big) \times \\
& \mathrm{P}\big(M_{\mathrm{opt}} = \hat{M}^*_h | \hat{M}^* = M_{\{\beta_1, \cdots, \beta_p \neq 0\}}, M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big)\Big] \\
= & \ a_0 \mathrm{P}\big(\hat{M}^* = M_{\{\beta=\mathbf{0}\}} | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big) + \cdots \\
& + a_q \mathrm{P}\big(\hat{M}^* = M_{\{\beta_1, \cdots, \beta_q \neq 0\}} | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big) + \cdots \\
& + a_p \mathrm{P}\big(\hat{M}^* = M_{\{\beta_1, \cdots, \beta_p \neq 0\}} | M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}\big), \qquad \text{(S2.6)}
\end{aligned}
$$

where $a_0, \cdots, a_p$ are some constants with values either 0 or 1. The first $q+1$ probabilities in (S2.6) have a similar form to (S2.3), while the last $p - q$ probabilities are all constants that do not contain $\tilde{\psi}_{\mathrm{opt}}$ similar as (S2.4). Hence, by repeating previous proof, assumption *A2* is satisfied for every $w \geq 0$, $0 \leq j \leq w$.

S3 Proof of Theorem 1

Let $\mathrm{P}^*(\cdot) = \mathrm{P}(\cdot|\hat{M}, \hat{\psi})$. Given $\hat{M}, \hat{\psi}$, by the weak law of large numbers, we have

$$\frac{1}{B}\sum_{b=1}^{B} 1_{(\hat{M}^*_{x,[b]}=\hat{M})} \xrightarrow{\mathrm{P}} \mathrm{P}^*(\hat{M}^*_x = \hat{M}), \tag{S3.1}$$

as $B \to \infty$ for any $-p \le x \le p$. Define $f(w) = \underset{0\le j\le w}{\operatorname{argmax}} \sum_{i=0}^{w} \mathrm{P}^*(\hat{M}^*_{-i+j} = \hat{M})$, and recall that $f^*(w) = \underset{0\le j\le w}{\operatorname{argmax}} \sum_{i=0}^{w} B^{-1} \sum_{b=1}^{B} 1_{(\hat{M}^*_{-i+j,[b]}=\hat{M})}$. For any given non-negative integer w, $f^*(w)$ converge in probability to $f(w)$, as B increases. Furthermore, it is easy to see that there exits a unique w such that

$$\mathrm{CP}(w-1) = \sum_{i=0}^{w-1} \mathrm{P}^*(\hat{M}^*_{-i+f(w-1)} = \hat{M}) < 1-\alpha \tag{S3.2}$$

$$\mathrm{CP}(w) = \sum_{i=0}^{w} \mathrm{P}^*(\hat{M}^*_{-i+f(w)} = \hat{M}) \ge 1-\alpha. \tag{S3.3}$$

Next we prove that w^* converges in probability to this w. If $w^* - w > 0$ for all B, we have $w \le w^* - 1$, and

$$\begin{aligned}\sum_{i=0}^{w} \mathrm{P}^*(\hat{M}^*_{-i+f(w)} = \hat{M}) &\le \sum_{i=0}^{w^*-1} \mathrm{P}^*(\hat{M}^*_{-i+f(w^*-1)} = \hat{M}) \\ &= \sum_{i=0}^{w^*-1} \frac{1}{B}\sum_{b=1}^{B} 1_{(\hat{M}^*_{-i+f^*(w^*-1),[b]}=\hat{M})} + o_{\mathrm{P}}(1). \end{aligned} \tag{S3.4}$$

As B goes to infinity, the $o_{\mathrm{P}}(1)$ in (S3.4) converges to zero in probability. Thus, with probability tending to one, the right side of (S3.4) is smaller

than $1-\alpha$ by the definition of w^*-1, which contradicts (S3.3). Similarly, if $w^*-w<0$ for all B, then $w-1\geq w^*$, and

$$\sum_{i=0}^{w-1}\mathrm{P}^*(\hat{M}^*_{-i+f(w-1)}=\hat{M})\geq\sum_{i=0}^{w^*}\mathrm{P}^*(\hat{M}^*_{-i+f(w^*)}=\hat{M})$$
$$=\sum_{i=0}^{w^*}\frac{1}{B}\sum_{b=1}^{B}1_{(\hat{M}^*_{-i+f^*(w^*),[b]}=\hat{M})}+o_{\mathrm{P}}(1). \quad \text{(S3.5)}$$

By a similar argument, with probability tending to one, the right side of (S3.5) is at least $1-\alpha$ according to the definition of w^*, which again results in a contradictory to (S3.2). Thus, w^*, j^* $[=f^*(w^*)]$ converge to w and j $[=f(w)]$ as B goes to infinity, in probability with respect to P^*. It follows that we have

$$\begin{aligned}1-\alpha &\leq \sum_{i=0}^{w^*}\frac{1}{B}\sum_{b=1}^{B}1_{(\hat{M}^*_{-i+j^*,[b]}=\hat{M})}\\ &= \sum_{i=0}^{w}\mathrm{P}^*(\hat{M}^*_{-i+j}=\hat{M})+o_{\mathrm{P}}(1)\\ &\leq \mathrm{P}(M_{\mathrm{opt}}\in(\hat{M}^*_{-w+j},\hat{M}^*_j)|M_{\mathrm{opt}},\hat{\psi}_{\mathrm{opt}})1_{(\hat{M}=M_{\mathrm{opt}})}\\ &+ 1_{(\hat{M}\neq M_{\mathrm{opt}})}+o_{\mathrm{P}}(1) \quad \text{(S3.6)}\end{aligned}$$

where $o_{\mathrm{P}}(1)$ converges to 0 with respect to P^* as B goes to infinity.

Next, by *A2*, we have

$$\begin{aligned}&\mathrm{P}(M_{\mathrm{opt}}\in(\hat{M}^*_{-w+j},\hat{M}^*_j)|M_{\mathrm{opt}},\hat{\psi}_{\mathrm{opt}})\\ \leq\ &\mathrm{P}(M_{\mathrm{opt}}\in(\hat{M}^*_{-w+j},\hat{M}^*_j)|M_{\mathrm{opt}},\psi_{\mathrm{opt}})+c|\hat{\psi}_{\mathrm{opt}}-\psi_{\mathrm{opt}}|, \quad \text{(S3.7)}\end{aligned}$$

Combining (S3.6) and (S3.7), we have

$$1-\alpha \leq \Big\{\mathrm{P}\big(M_{\mathrm{opt}} \in (\hat{M}^*_{-w+j}, \hat{M}^*_j)\big) + c|\hat{\psi}_{\mathrm{opt}} - \psi_{\mathrm{opt}}|\Big\}1_{(\hat{M}=M_{\mathrm{opt}})} + 1_{(\hat{M}\neq M_{\mathrm{opt}})} + o_{\mathrm{P}}(1), \tag{S3.8}$$

Note that the $o_{\mathrm{P}}(1)$ on the right side of (S3.8) is bounded, because we are dealing with the average of indicator functions. Therefore, by the dominated convergence theorem (e.g., Jiang (2010, Theorem 2.16)), we have $\mathrm{E}\{o_{\mathrm{P}}(1)\} = o(1)$, where the E is with respect to the joint distribution of y and y^*, the bootstrapped samples. We now take expectation on both sides of (S3.8), again with respect to the joint distribution of y and y^*, we have

$$\begin{aligned} 1-\alpha \leq\ & \mathrm{P}\big(M_{\mathrm{opt}} \in (\hat{M}^*_{-w+j}, \hat{M}^*_j)\big)\mathrm{P}(\hat{M} = M_{\mathrm{opt}}) \\ & + c\mathrm{E}\Big\{|\hat{\psi}_{\mathrm{opt}} - \psi_{\mathrm{opt}}|1_{(\hat{M}=M_{\mathrm{opt}})}\Big\} \\ & + \mathrm{P}(\hat{M} \neq M_{\mathrm{opt}}) + o(1). \end{aligned} \tag{S3.9}$$

The inequality in Theorem 2 then follows.

S4 Proof of Proposition 2

Finite $2k$th moment. Assume $\varepsilon_i, i = 1, \cdots, n$ are i.i.d. random variables with finite $2k$'th moment $E(\varepsilon_i)^{2k} < \infty$ for an integer $k > 0$, and design matrix satisfies Strong Irrepresentable Condition and (3.19)-(3.22). By Theorem 3 of Zhao and Yu (2006), Lasso has strong sign consistency for

$p_n = o(n^{(c_2-c_1)k})$. In particular, for any λ_n satisfies $\lambda_n/\sqrt{n} = o(n^{\frac{c_2-c_1}{2}})$ and $p_n^{-1}(\lambda_n/\sqrt{n})^{2k} \to \infty$, one has $\mathrm{P}(\hat{M} \neq M_{\mathrm{opt}}|M_{\mathrm{opt}}, \psi_{\mathrm{opt}}) = O(p_n n^k/\lambda_n^{2k})$, where $0 \leq c_1 < c_2 \leq 1$ are some constants. The selection consistency assumption *B1* is satisfied. Since the estimated coefficients of Lasso are continous functions of λ and $0 < P_* < 1$, assumptions *B2* and *B4* follows obviously.

Assume $p_n n^k/\lambda_n^{2k} = cn^{-a}$ for simplicity, where $a > 0$ is a constant that can be estimated using a common method, e.g., MLE, and c is a normalizing constant. We have

$$\frac{\log\{\mathrm{P}(\hat{M}^* \neq M_{\mathrm{opt}}|M_{\mathrm{opt}}, \hat{\boldsymbol{\psi}}_{\mathrm{opt}})\}}{\log\{\mathrm{P}(\hat{M}^* \neq M_{\mathrm{opt}}|M_{\mathrm{opt}}, \boldsymbol{\psi}_{\mathrm{opt}})\}} = \frac{\log \hat{c} - \hat{a}\log n}{\log c - a \log n} \tag{S4.1}$$

$$\xrightarrow{\mathrm{P}} 1. \tag{S4.2}$$

This is because $\log n = o(n^{1/2})$, $(\hat{a} - a)\log n = O_{\mathrm{P}}(n^{-1/2})o(n^{1/2}) = o_{\mathrm{P}}(1)$. Hence, assumption *B3* is satisfied.

S5 Proof of Proposition 3

Gaussian noise. Assume $\varepsilon_i, i = 1, \cdots, n$ are i.i.d. Gaussian random variables, and the design matrix satisfies the strong Irrepresentable Condition and (3.19)-(3.22), if there exists $0 \leq c_3 < c_2 - c_1$ for which $p_n = O(e^{n^{c_3}})$, by Theorem 4 of Zhao and Yu (2006), Lasso has strong sign consistency, i.e.,

for $\lambda \propto n^{\frac{1+c_4}{2}}$, with $c_3 < c_4 < c_2 - c_1$, $\mathrm{P}(\hat{M} \neq M_{\text{opt}}|M_{\text{opt}}, \psi_{\text{opt}}) = o(e^{-n^{c_3}})$. Assumptions *B1, B2, B4* are satisfied.

Assume $\mathrm{P}(\hat{M} \neq M_{\text{opt}}|M_{\text{opt}}, \psi_{\text{opt}}) = c_1 e^{-n^{c_2} - a_n}$, where a_n is positive and goes to infinity at a slower rate than n^{c_2}, both c_2 and a_n can be estimated. It follows that

$$\begin{aligned} \frac{\log\{\mathrm{P}(\hat{M}^* \neq M_{\text{opt}}|M_{\text{opt}}, \hat{\psi}_{\text{opt}})\}}{\log\{\mathrm{P}(\hat{M}^* \neq M_{\text{opt}}|M_{\text{opt}}, \psi_{\text{opt}})\}} &= \frac{-n^{\hat{c}_2} - \hat{a}_n + \log c_1}{-n^{c_2} - a_n + \log c_1} \\ &\approx n^{\hat{c}_2 - c_2} \\ &= e^{(\hat{c}_2 - c_2)\log n}, \end{aligned} \tag{S5.1}$$

Since $\log n = o(n^{1/2})$, $(\hat{c}_2 - c_2)\log n = O_{\mathrm{P}}(n^{-1/2})o(n^{1/2}) = o_{\mathrm{P}}(1)$. Hence, as $n \to \infty$, (S5.1) converges to 1; therefore, assumption *B3* is satisfied.

S6 Tables

Bibliography

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression.*Annals of Statistics* 32(2), 407–499.]

Hastie, T., Taylor, J., Tibshirani, R., and Walther, G. (2007). Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics, Electron. J. Statist.* 1(none), 1-29.

Table 1: Comparison on LM

		NMCS			MCB	
Case	C.L.(%)	P^*	CP	AW	CP	AW
1	95	0.965	1.000	3.845	1.000	9.090
1	90	0.965	1.000	2.905	1.000	3.455
1	80	0.965	0.995	1.995	1.000	0.790
1	70	0.965	0.990	1.285	1.000	0.155
1	60	0.965	0.990	0.800	0.995	0.045
1	50	0.965	0.985	0.035	0.990	0.040
1	40	0.965	0.965	0.010	0.980	0.015
1	30	0.965	0.965	0.001	0.970	0.005
2	95	0.850	1.000	18.010	1.000	44.170
2	90	0.850	1.000	15.065	1.000	25.140
2	80	0.850	1.000	9.850	1.000	7.730
2	70	0.850	0.985	4.220	1.000	1.495
2	60	0.850	0.985	1.395	1.000	0.500
2	50	0.850	0.960	0.780	0.980	0.300
2	40	0.850	0.895	0.370	0.905	0.150
2	30	0.850	0.870	0.180	0.875	0.065
3	95	0.875	0.995	15.765	1.000	34.550
3	90	0.875	0.995	13.675	1.000	19.265
3	80	0.875	0.995	10.330	1.000	7.200
3	70	0.875	0.985	5.935	1.000	3.465
3	60	0.875	0.975	3.440	1.000	1.700
3	50	0.875	0.980	2.665	0.990	0.845
3	40	0.875	0.965	1.320	0.960	0.440
3	30	0.875	0.935	0.990	0.945	0.230

Table 2: NMCS for logistic regression

Case	C.L.(%)	SIS_rate	P^*	CP	AW
1	95	1.000	0.595	0.990	33.040
1	90	1.000	0.595	0.990	21.325
1	80	1.000	0.595	0.995	8.035
1	70	1.000	0.595	0.980	3.785
1	60	1.000	0.595	0.940	1.990
1	50	1.000	0.595	0.860	1.240
1	40	1.000	0.595	0.750	0.700
1	30	1.000	0.595	0.595	0.305
2	95	0.980	0.480	0.970	53.060
2	90	0.980	0.480	0.955	37.505
2	80	0.980	0.480	0.955	17.065
2	70	0.980	0.480	0.940	10.045
2	60	0.980	0.480	0.905	5.600
2	50	0.980	0.480	0.855	2.840
2	40	0.980	0.480	0.765	1.640
2	30	0.980	0.480	0.600	0.950
3	95	0.970	0.365	0.905	74.840
3	90	0.970	0.365	0.905	63.785
3	80	0.970	0.365	0.905	42.570
3	70	0.970	0.365	0.905	29.970
3	60	0.970	0.365	0.900	20.080
3	50	0.970	0.365	0.880	15.325
3	40	0.970	0.365	0.820	11.285
3	30	0.970	0.365	0.730	6.435

Table 3: NMCS for GAM

C.L.(%)	P^*	CP	AW
95	0.24	0.99	23.59
90	0.24	0.97	22.49
80	0.24	0.96	20.77
70	0.24	0.94	18.82
60	0.24	0.91	17.03
50	0.24	0.92	15.28
40	0.24	0.88	13.36
30	0.24	0.87	11.26

Table 4: Performance of LogP measure on LM

Penalty	n	p	B	$\mathrm{E}(\widehat{\mathrm{LogP}})$	LogP	P^*	%RB	CV
Lasso	50	18	500	-0.096	-0.186	0.170	-48.3	0.94
SCAD	50	18	500	-1.343	-1.561	0.790	-13.9	0.45
ALasso	50	18	500	-1.571	-1.609	0.800	-2.4	0.52
Lasso	100	31	500	-0.177	-0.439	0.355	-59.7	0.86
SCAD	100	31	500	-2.148	-2.254	0.895	-4.7	0.31
ALasso	100	31	500	-2.482	-2.303	0.900	7.8	0.46
Lasso	200	53	500	-0.256	-0.486	0.385	-47.4	0.74
SCAD	200	53	500	-2.727	-2.659	0.930	2.6	0.24
ALasso	200	53	500	-4.516	-3.912	0.980	15.4	0.43
Lasso	500	105	500	-0.402	-0.589	0.445	-31.8	0.64
SCAD	500	105	500	-3.298	-2.733	0.935	20.7	0.23
ALasso	500	105	500	-7.901	-5.298	0.995	49.1	0.25
Lasso	500	105	1000	-0.412	-0.669	0.488	-38.5	0.64
SCAD	500	105	1000	-3.361	-3.124	0.956	7.6	0.19
ALasso	500	105	1000	-7.675	-5.298	0.995	44.9	0.26

Table 5: Comparison on NDD1 data

Criterion	Method	$\hat{M}$	LogP	C.L.	Width	LBM	UBM
ALasso-BIC	NMCS	1, 5, 80	-0.028	0.95	22	NULL	1, 4, 5, 8, 10, 12, 15, 25,...
				0.9	11	NULL	1, 4, 5, 12, 15, 26, 47, 54...
				0.8	6	NULL	1, 4, 5, 12, 80, 96
				0.7	4	NULL	1, 4, 5, 80
	MCB	1, 5, 80		0.95	65	NULL	1, 2, 3, 4, 5, 8, 10, 11, 12, ...
				0.9	64	1	1, 2, 3, 4, 5, 8, 10, 11, 12, ...
				0.8	63	1, 5	1, 2, 3, 4, 5, 8, 10, 11, 12, ...
				0.7	62	1, 5, 80	1, 2, 3, 4, 5, 8, 10, 11, 12, ...
ALasso-CV	NMCS	1, 4, 5, 80	-0.030	0.95	21	NULL	1, 4, 5, 10, 12, 15, 16, 25,...
				0.9	10	NULL	1, 4, 5, 12, 15, 26, 47, 54, 80, 96
				0.8	5	NULL	1, 4, 5, 12, 80
				0.7	3	NULL	1, 5, 80
	MCB	1, 5, 80		0.95	50	NULL	1, 2, 3, 4, 5, 8, 10, 11, 12, ...
				0.9	40	NULL	1, 2, 3, 4, 5, 8, 10, 11, 12, ...
				0.8	25	1	1, 2, 3, 4, 5, 10, 11, 12, ...
				0.7	12	1	1, 2, 3, 4, 5,19,44,47 ...

Table 6: NMCS for South African Heart Disease Data

Penalty	Tune	$\hat{M}$	LogP	C.L.	Width	LBM	UBM
Lasso	AIC	1, 2, 3, 5, 6, 7, 9	-0.139	0.95	6	2, 5, 9	1:9
				0.9	6	2, 5, 9	1:9
				0.8	5	2, 3, 5, 9	1:9
				0.7	4	2, 3, 5, 6, 9	1:9
	BIC:	1, 2, 3, 5, 6, 9	-0.228	0.95	8	9	1:9
				0.9	7	5, 9	1:9
				0.8	7	5, 9	1:9
				0.7	5	2, 3, 5, 9	1:9
	CV	2, 3, 5, 6, 9	-0.521	0.95	7	NULL	1, 2, 3, 5, 6, 7, 9
				0.9	6	NULL	1, 2, 3, 5, 6, 7, 9
				0.8	4	2, 5, 9	1, 2, 3, 5, 6, 7, 9
				0.7	3	2, 5, 9	1, 2, 3, 5, 6, 9
ALasso	AIC	2, 3, 5, 6, 9	-0.446	0.95	6	9	1, 2, 3, 5, 6, 7, 9
				0.9	5	9	1, 2, 3, 5, 6, 9
				0.8	3	5, 9	2, 3, 5, 6, 9
				0.7	2	2, 5, 9	2, 3, 5, 6, 9
	BIC	2, 3, 5, 6, 9	-0.729	0.95	6	2, 5, 9	1:9
				0.9	4	2, 3, 5, 9	1, 2, 3, 5, 6, 7, 9
				0.8	3	2, 3, 5, 9	1, 2, 3, 5, 6, 7, 9
				0.7	2	2, 3, 5, 9	1, 2, 3, 5, 6, 9
	CV	2, 5, 9	-0.465	0.95	2	2, 5, 9	2, 3, 5, 6, 9
				0.9	2	2, 5, 9	2, 3, 5, 6, 9
				0.8	1	2, 5, 9	2, 3, 5, 9
				0.7	1	2, 5, 9	2, 3, 5, 9

Jiang, J. (2010), Large Sample Techniques for Statistics, *Springer*, New York.

Portnoy, S.(1984), Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *Ann. Statist.* 12 (4) 1298 - 1309.

Zhao, P. and Yu, B (2006), On Model Selection Consistency of Lasso, *Journal of Machine Learning Research* 7, 2541–2563.