Journal of Machine Learning Research 14 (2013) 1-9 Submitted 8/12; Revised 7/13; Published 11/13

Consistent Selection of Tuning Parameters via Variable Selection Stability (Supplementary Material)

Wei Sun SUN244@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Junhui Wang JUNHUI@UIC.EDU
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA

Yixin Fang YIXIN.FANG@NYUMC.ORG
Departments of Population Health and Environmental Medicine
New York University
New York, NY 10016, USA

Editor: Xiaotong Shen

In this supplementary material, we provide Lemmas 2 and 3 and their proofs.

Suppose that $x_1,\ldots,x_n$ are i.i.d. from a probability distribution with mean 0 and finite covariance matrix $C=(C_{jk})$.

Assumption S1: Assume that x_1 has finite fourth moment, that is, $E(x_{1i}x_{1j}x_{1k}x_{1l})$ is finite for any $1\le i,j,k,l\le p$.

Lemma 2 *Suppose that Assumption S1 is met. Assumptions 1 and 2 are satisfied by the lasso regression and the SCAD with $r_n=n^{-1/2}$ and $s_n=o(1)$ under the assumptions in Zhao and Yu (2006) or Fan and Li (2001), and by the adaptive lasso with $r_n=n^{-1}$ and $s_n=n^{-1/2}$ under the assumptions in Zou (2006), on the random splitting subsamples generated in Algorithm 1.*

Proof of Lemma 2: First, for random variables $w_i\sim Bern(1/2)$; $i=1,\ldots,n$ that are independent with x_i's and satisfy $\sum_{i=1}^n w_i=\lfloor n/2\rfloor$, we show that $\frac{2}{n}\sum_{i=1}^n w_i x_i x_i^T \xrightarrow{p} C$.

For fixed p, it suffices to show the componentwise convergence in probability,

$$S_n=\frac{1}{n}\sum_{i=1}^n 2w_i x_{ij}x_{ik}\xrightarrow{p} C_{jk}. \tag{1}$$

Note that $E(w_i)=E(w_i^2)=1/2$, and thus

$$E(S_n)=\frac{1}{n}\sum_{i=1}^n E(2w_i x_{ij}x_{ik})=\frac{1}{n}\sum_{i=1}^n E(2w_i)E(x_{ij}x_{ik})=C_{jk},$$

following the independence between w_i and x_i.

In addition,

$$\begin{aligned}
\text{var}(S_n) &= E(S_n^2) - E(S_n)^2 = E\left(\left(\frac{1}{n}\sum_{i=1}^{n} 2w_i x_{ij}x_{ik}\right)^2\right) - C_{jk}^2 \\
&= \frac{4}{n^2}\left(E\Big(\sum_{i=1}^{n} w_i^2 x_{ij}^2 x_{ik}^2\Big) + \sum_{i\neq l} E(w_i w_l)E(x_{ij}x_{ik})E(x_{lj}x_{lk})\right) - C_{jk}^2 \\
&= \frac{4}{n^2}E\Big(\sum_{i=1}^{n} w_i^2 x_{ij}^2 x_{ik}^2\Big) + \frac{4(n-1)C_{jk}^2}{n}\,\text{cov}(w_1, w_2) - \frac{C_{jk}^2}{n} \\
&\leq \frac{4}{n^2}E\Big(\sum_{i=1}^{n} w_i^2 x_{ij}^2 x_{ik}^2\Big) = \frac{2}{n}E(x_{ij}^2 x_{ik}^2) \to 0,
\end{aligned}$$

where the inequalities follow from the fact $\text{cov}(w_1, w_2) < 0$, $E(x_{ij}x_{ik}) = E(x_{lj}x_{lk}) = C_{jk}$, and $E(w_i^2) = 1/2$, and the convergence is due to the finite fourth moment of x_i. The Chebyshev's inequality immediately implies that $\frac{2}{n}\sum_{i=1}^{n} w_i x_i x_i^T \xrightarrow{p} C$.

Next we prove Lemma 2 for (i) the lasso regression, (ii) the adaptive lasso, and (iii) the SCAD, respectively.

(i): The lasso regression. When the original assumption $\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \to C$ is replaced by *Assumption S1*, the proof follows from the above convergence in probability statement and slight modification of some existing results in literature. Specifically, the existence of r_n and s_n for selection consistency in Assumption 1 can be verified as in Section 2.1 of Zhao and Yu (2006). The condition (3) in Assumption 1 is a direct result from Assumption 2, which will be shown after we verify conditions in Assumption 2 based on Lemma C.2 in Bach (2009).

Denote the permutated subsample $(w_i x_1, \ldots, w_n x_n)$ as $\mathbf{Z} = (z_1, \ldots, z_m)^T$ with $m = \lfloor n/2 \rfloor$. Denote $Q = \frac{1}{m}\mathbf{Z}^T\mathbf{Z} = \frac{1}{m}\sum_{i=1}^{m} z_i z_i^T$, $\lambda_{\min}(Q)$ as the minimal eigenvalue of Q, $q = \mathbf{Z}^T\varepsilon/m$, the true coefficient as β^*, and a sign pattern $s = \{1, 0, -1\}^p$ such that for any $j \in \{1, \ldots, p\}$, $s_j = \text{sign}(\beta_j)$. For simplicity, we denote $J = \widehat{\mathcal{A}}_{\lambda_m}$, $\mathbf{J} = \mathcal{A}_T$, and s_J as the sign pattern of variables indexed by J. Let $M(\beta) = \min_{j\in\{1,\ldots,p\}, \beta_j \neq 0} |\beta_j|$ as the smallest magnitude of non-zero elements in β, and $\|C\|_\infty$ as the largest magnitude of all the elements in matrix C. According to Lemma C.2 in Bach (2009), when the selected active set is over-fitting such that $s_{\mathbf{J}} = \text{sign}(\beta_{\mathbf{J}})$ and $J \supset \mathbf{J}$, we have that s is selected if and only if

$$\|Q_{J^c,J}Q_{J,J}^{-1}q_J - q_{J^c} - \lambda_m Q_{J^c,J}Q_{J,J}^{-1}s_J\|_\infty \leq \lambda_m; \tag{2}$$

$$\text{sign}\Big(\beta_{\mathbf{J}}^* + (Q_{J^c,J}Q_{J,J}^{-1}q_J - \lambda_m Q_{J,J}^{-1}s_J)_{\mathbf{J}}\Big) = \text{sign}(\beta_{\mathbf{J}}^*); \tag{3}$$

$$\text{sign}\Big(Q_{J,J}^{-1}q_J - \lambda_m Q_{J,J}^{-1}s_J\Big)_{J\cap\mathbf{J}^c} = s_{J\cap\mathbf{J}^c}. \tag{4}$$

Therefore, for a particular over-fitting sign pattern $\tilde{s}$ with jth noise variable selected in the active set J, we have $\{j \in J\} \supseteq \{\text{sign}(\hat{\beta}_n) = \tilde{s}\}$, where $\{\text{sign}(\hat{\beta}_n) = \tilde{s}\}$ is equivalent to the conditions of (2)-(4) with $s = \tilde{s}$. For short, we denote $\{(3)^c\}$ as the complement of condition in (3), and $\{(2),(4)\}^c$ as the complement of conditions in (2) and (4), respectively. When $\sqrt{m}\lambda_m \leq \lambda_0 \in (0, \infty)$, Proposition 2.4

in Bach (2009) leads to

$$\begin{aligned}
&\bigcup_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0} \{(3)^c\} \\
\subseteq\ & \bigcup_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0} \left\{\sqrt{m}\lambda_m > \frac{\sqrt{m}M(\beta^*)\lambda_{\min}(Q)}{2\sqrt{p}},\ \text{or}\ \|(Q_{J,J}^{-1}q_J)_{\mathbf{J}}\|_2 > \frac{M(\beta^*)}{2}\right\} \\
\subseteq\ & \left\{\lambda_0 > \frac{\sqrt{m}M(\beta^*)\lambda_{\min}(Q)}{2\sqrt{p}},\ \text{or}\ \|(Q_{J,J}^{-1}q_J)_{\mathbf{J}}\|_2 > \frac{M(\beta^*)}{2}\right\}, \qquad (5)
\end{aligned}$$

with the right hand side in (5) having probability tending to 0, and as $m\to\infty$

$$\bigcup_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0} \{(2),(4)\}^c \to \{v\notin C(\tilde{s},\lambda_0)\}, \qquad (6)$$

where v is normal with zero mean and covariance matrix Q, and $C(\tilde{s},\lambda_0)$ is a convex set and its complement also have non-empty interior, and hence $P(v\notin C(\tilde{s},\lambda_0))$ is strictly within $(0,1)$ for any fixed λ_0. Therefore, as $m\to\infty$, combining (5) and (6) leads to

$$\begin{aligned}
&P\Big(\bigcap_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0} \{j\in\widehat{\mathcal{A}}_{\lambda_m}\}\Big) \\
\ge\ & P\Big(\bigcap_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0} \{\text{sign}(\hat{\beta}_n)=\tilde{s}\}\Big) \\
=\ & P\Big(\bigcap_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0} \{(2),(4)\}\cap\{(3)\}\Big) \\
\ge\ & 1-P\Big(\bigcup_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0} \{(2),(4)\}^c\Big) - P\Big(\bigcup_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0} \{(3)\}^c\Big) \\
\to\ & 1-P(v\notin C(\tilde{s},\lambda_0))\in(0,1),
\end{aligned}$$

and hence condition (5) in Assumption 2 is verified. In addition, as shown in Proposition 1 in Bach (2008), when λ_0 converges to 0, $P(v\notin C(\tilde{s},\lambda_0))\to 0$, and hence $c_1(\lambda_0)=1-P(v\notin C(\tilde{s},\lambda_0))\to 1$. The condition (6) in Assumption 2 can be proved by defining the particular sign pattern $\tilde{s}$ to be the one with jth noise variable not selected in the active set J, then $\{j\notin J\}\supseteq\{\text{sign}(\hat{\beta}_n)=\tilde{s}\}$. All the proof can be derived following similar approach as above. Therefore, for any $j\in\mathcal{A}_T^c$, we have $P(\cap_{r_n^{-1}\lambda_n\ge\lambda_0}\{j\notin\widehat{\mathcal{A}}_{\lambda_n}\})\ge c_2(\lambda_0)$ with $c_2(\lambda_0)\to 1$ as $\lambda_0\to\infty$. In addition, after a slight modification of Proposition 2.5 in Bach (2009), we can show that uniformly over λ_m such that $\sqrt{m}\lambda_m\le\lambda_0$, all the important variable will be selected with probability tending to 1, which verifies condition (4) in Assumption 2. This ends the verification of Assumption 2 for the lasso regression.

Finally we show condition (3) in Assumption 1 for the lasso regression. Note that

$$\begin{aligned}
&\bigcap_{\lambda_0 m^{-1/2}\le\lambda_m\le\lambda_m^*} \Big\{\widehat{\mathcal{A}}_{\lambda_m}=\mathcal{A}_T\Big\} \\
=\ & \bigcap_{\lambda_0 m^{-1/2}\le\lambda_m\le\lambda_m^*} \left\{\bigcap_{j\in\mathcal{A}_T}\{j\in\widehat{\mathcal{A}}_{\lambda_m}\}\right\}\bigcap\left\{\bigcap_{j_1\in\mathcal{A}_T^c}\{j_1\notin\widehat{\mathcal{A}}_{\lambda_m}\}\right\} \\
\supset\ & \Big\{\bigcap_{\lambda_m\le\lambda_m^*;j\in\mathcal{A}_T}\{j\in\widehat{\mathcal{A}}_{\lambda_m}\}\Big\}\bigcap\Big\{\bigcap_{\lambda_m\ge\lambda_0 m^{-1/2};j_1\in\mathcal{A}_T^c}\{j_1\notin\widehat{\mathcal{A}}_{\lambda_m}\}\Big\}.
\end{aligned}$$

Following the similar strategy in the proof of conditions (4) and (6), the selection consistency in Zhao and Yu (2006) and Proposition 2.5 in Bach (2009) imply that all the important variables will be included uniformly over $\lambda_m \le \lambda_m^*$, and all the noisy variables will be excluded in the active set $\widehat{\mathcal{A}}_{\lambda_m}$ uniformly over $\lambda_m \ge \lambda_0 m^{-1/2}$. Therefore, when n is sufficiently large,

$$\begin{aligned} & P\Big(\bigcap_{\lambda_0 m^{-1/2} \le \lambda_m \le \lambda_m^*} \Big\{\widehat{\mathcal{A}}_{\lambda_m} = \mathcal{A}_T\Big\}\Big) \\ \ge\ & P\Big(\bigcap_{\lambda_m \le \lambda_m^*; j \in \mathcal{A}_T} \{j \in \widehat{\mathcal{A}}_{\lambda_m}\}\Big) + P\Big(\bigcap_{\lambda_0 m^{-1/2} \le \lambda_m; j_1 \in \mathcal{A}_T^c} \{j_1 \notin \widehat{\mathcal{A}}_{\lambda_m}\}\Big) - 1 \\ \ge\ & c_2(\lambda_0) - \zeta_n. \end{aligned}$$

Since $\zeta_n \to 0$ and $\lim_{\lambda_0 \to \infty} c_2(\lambda_0) = 1$ as shown above, letting $c_0(\lambda_0) = 1 - c_2(\lambda_0)/2$ leads to (3) in Assumption 1. This ends the verification for lasso regression.

(ii): The adaptive lasso. When the original assumption $\frac{1}{n}\sum_{i=1}^n x_i x_i^T \to C$ is replaced by *Assumption S1*, the selection consistency established in Zou (2006) when $n\lambda_n \to \infty$ and $\sqrt{n}\lambda_n \to 0$ is still valid with the above convergence in probability statement. In specific, we also denote the permutated subsample $(w_i x_1, \ldots, w_n x_n)$ as $\mathbf{Z} = (z_1, \ldots, z_m)^T$ with $m = \lfloor n/2 \rfloor$. It is shown above that $\frac{1}{m}\mathbf{Z}^T\mathbf{Z} = \frac{1}{m}\sum_{i=1}^m z_i z_i^T \xrightarrow{p} C$. Denote β^* as the true coefficient, $\beta = \beta^* + \frac{u}{\sqrt{m}}$, and

$$\Psi_m(u) = \Big\|\mathbf{y} - \sum_{j=1}^p \mathbf{z}_{(j)}\Big(\beta_j^* + \frac{u_j}{\sqrt{m}}\Big)\Big\|^2 + m\lambda_m \sum_{j=1}^p \frac{\big|\beta_j^* + \frac{u_j}{\sqrt{m}}\big|}{|\hat{\beta}_j^L|},$$

where $\hat{\beta}_j^L$ is the estimator from the lasso regression. Let $\hat{u}_m = \arg\min \Psi_m(u)$, $\hat{\beta}_m = \beta^* + \frac{\hat{u}_m}{\sqrt{m}}$, and $V_m(u) = \Psi_m(u) - \Psi_m(0)$ with

$$V_m(u) = u^T\Big(\frac{\mathbf{Z}^T\mathbf{Z}}{m}\Big)u - \frac{2\varepsilon^T\mathbf{Z}}{\sqrt{m}}u + \sqrt{m}\lambda_m \sum_{j=1}^p \frac{\sqrt{m}\Big(\big|\beta_j^* + \frac{u_j}{\sqrt{m}}\big| - |\beta_j^*|\Big)}{|\hat{\beta}_j^L|}.$$

Note that $\frac{\mathbf{Z}^T\mathbf{Z}}{m} \xrightarrow{p} \mathbf{C}$, $\frac{\varepsilon^T\mathbf{Z}}{\sqrt{m}} \xrightarrow{d} W^T \sim N(0, \sigma^2\mathbf{C})$ from the central limit theorem. Similar to the fixed design case in Zou (2006), we can show that with probability tending to 1, the asymptotic normality of $\hat{u}_m$ holds on $\mathcal{A}_T$ and $\hat{u}_m \to 0$ on $\mathcal{A}_T^c$. In addition, for any $j \notin \mathcal{A}_T$, it's sufficient to show $P(j \in \widehat{\mathcal{A}}_{\lambda_m}) \to 0$. Note that when $j \in \widehat{\mathcal{A}}_{\lambda_m}$, the Karush-Kuhn-Tucker (KKT) conditions imply that $2\mathbf{z}_{(j)}^T(\mathbf{y} - \mathbf{Z}\hat{\beta}_n) = m\frac{\lambda_m}{|\hat{\beta}_j^L|}$, where $\sqrt{m}\frac{\lambda_m}{|\hat{\beta}_j^L|} \xrightarrow{p} \infty$, and

$$\frac{2\mathbf{z}_{(j)}^T(\mathbf{y} - \mathbf{Z}\hat{\beta}_m)}{\sqrt{m}} = \frac{\mathbf{z}_{(j)}^T\mathbf{Z}\sqrt{m}(\beta^* - \hat{\beta}_m)}{m} + \frac{2\mathbf{z}_{(j)}^T\varepsilon}{\sqrt{m}}.$$

Note that $\frac{\mathbf{Z}^T\mathbf{Z}}{m} \xrightarrow{p} \mathbf{C}$ and $\sqrt{m}(\beta^* - \hat{\beta}_m)$ is asymptotic normal as shown above, Slutsky's theorem implies that both $2\mathbf{z}_{(j)}^T\mathbf{Z}\sqrt{m}(\beta^* - \hat{\beta}_m)/m$ and $2\mathbf{z}_{(j)}^T\varepsilon/\sqrt{m}$ converge in distribution to normal. Therefore, $P(j \in \widehat{\mathcal{A}}_{\lambda_m}) \to 0$.

Next we verify Assumption 2 for the permutated subsample $\mathbf{Z} = (z_1, \ldots, z_m)^T$. When $\lambda_m \preceq m^{-1}$, we have $\lambda_m \prec m^{-1/2}$, and hence the asymptotic normality of $\hat{\beta}_m$ still holds for any satisfied λ_m (Zou,

2006). This implies condition (4) in Assumption 2 directly. It then suffices to consider the event $j \notin \widehat{\mathcal{A}}_{\lambda_m}$ for any $j \in \mathcal{A}_T^c$. Note that when $j \notin \widehat{\mathcal{A}}_{\lambda_m}$, the Karush-Kuhn-Tucker (KKT) conditions imply that

$$\left|2\mathbf{z}_{(j)}^T(\mathbf{y}-\mathbf{Z}\hat{\beta}_n)\right| \leq m\frac{\lambda_m}{|\hat{\beta}_j^L|}.$$

In addition,

$$\frac{2\mathbf{z}_{(j)}^T(\mathbf{y}-\mathbf{Z}\hat{\beta}_m)}{\sqrt{m}} = \frac{2\mathbf{z}_{(j)}^T\mathbf{Z}\sqrt{m}(\beta^*-\hat{\beta}_m)}{m} + \frac{2\mathbf{z}_{(j)}^T\varepsilon}{\sqrt{m}}.$$

By the asymptotic normality of $\hat{\beta}_m$ and $\frac{\mathbf{Z}^T\mathbf{Z}}{m} \xrightarrow{p} \mathbf{C}$, the Slutsky's theorem implies that $2\mathbf{z}_{(j)}^T\mathbf{Z}\sqrt{m}(\beta^*-\hat{\beta}_m)/m \xrightarrow{d} N(0,\Delta_1)$ for some Δ_1, and $2\mathbf{z}_{(j)}^T\varepsilon/\sqrt{m} \xrightarrow{d} N(0,\Delta_2)$ for some Δ_2. In addition, when $m\lambda_m \leq \lambda_0$ for some $\lambda_0 \in (0,\infty)$, we have

$$\begin{aligned}\left\{\bigcup_{\lambda_m:m\lambda_m\leq\lambda_0}\{j\notin\widehat{\mathcal{A}}_{\lambda_m}\}\right\} &\subseteq \left\{\bigcup_{\lambda_m:m\lambda_m\leq\lambda_0}\left\{\left|2\mathbf{z}_{(j)}^T(\mathbf{y}-\mathbf{Z}\hat{\beta}_m)\right| \leq \frac{m\lambda_m}{|\hat{\beta}_j^L|}\right\}\right\} \\ &\subseteq \left\{\left|2\mathbf{z}_{(j)}^T(\mathbf{y}-\mathbf{Z}\hat{\beta}_m)\right| \leq \frac{\lambda_0}{|\hat{\beta}_j^L|}\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}P\Big(\bigcup_{\lambda_m:m\lambda_m\leq\lambda_0}\{j\notin\widehat{\mathcal{A}}_{\lambda_m}\}\Big) &\leq P\Big(\left|2\mathbf{z}_{(j)}^T(\mathbf{y}-\mathbf{Z}\hat{\beta}_m)\right| \leq \frac{\lambda_0}{|\hat{\beta}_j^L|}\Big) \\ &= P\Big(\Big|\frac{2\mathbf{z}_{(j)}^T\mathbf{Z}\sqrt{m}(\beta^*-\hat{\beta}_m)}{m} + \frac{2\mathbf{z}_{(j)}^T\varepsilon}{\sqrt{m}}\Big| \leq \frac{\lambda_0}{|\sqrt{m}\hat{\beta}_j^L|}\Big).\end{aligned}$$

The DCT theorem implies that for sufficiently large m,

$$P\Big(\bigcup_{\lambda_m:m\lambda_m\leq\lambda_0}\{j\notin\widehat{\mathcal{A}}_{\lambda_m}\}\Big) \leq 1-c_0(\lambda_0),$$

with $c_1(\lambda_0)$ being a strictly positive constant in $(0,1)$, and hence

$$P\Big(\bigcap_{\lambda_m:m\lambda_m\leq\lambda_0}\{j\in\widehat{\mathcal{A}}_{\lambda_n}\}\Big) \geq c_0(\lambda_0).$$

In addition, the condition (6) in Assumption 2 can be verified by slightly modifying the proof in the lasso regression. Specifically, we define s to be the one with jth variable not selected in the active set J. Then we only need to replace λ_m with $\lambda_m/\widehat{\beta}_j^L$ for $j \in J$ in (5) and (6). When $m\lambda_m \leq \lambda_0$, we have that (5) is replaced with

$$\begin{aligned}&\bigcup_{\lambda_m:m\lambda_m\leq\lambda_0}\{(3)\}^c \\ \subseteq\ &\bigcup_{\lambda_m:m\lambda_m\leq\lambda_0}\left\{\frac{m\lambda_m}{M(\sqrt{m}\widehat{\beta}_J^L)} > \frac{\sqrt{m}M(\beta^*)\lambda_{\min}(Q)}{2\sqrt{p}},\ \text{or}\ \|(Q_{J,J}^{-1}q_J)_{\mathbf{J}}\|_2 > \frac{M(\beta^*)}{2}\right\} \\ \subseteq\ &\left\{\lambda_0 > \frac{\sqrt{m}M(\beta^*)\lambda_{\min}(Q)M(\sqrt{m}\widehat{\beta}_J^L)}{2\sqrt{p}},\ \text{or}\ \|(Q_{J,J}^{-1}q_J)_{\mathbf{J}}\|_2 > \frac{M(\beta^*)}{2}\right\}, \qquad (7)\end{aligned}$$

with the right hand side in (7) still having probability tending to 0 since $\sqrt{m}\widehat{\beta}_j^L = O_p(1)$ for any $j \in J$. In addition, (6) is replaced with

$$\bigcup_{\lambda_m : m\lambda_m \le \lambda_0} \{(2),(4)\}^c \to \{v \notin \mathcal{C}(s,\lambda_0)\}.$$

Therefore, we still have

$$P\Big(\bigcap_{\lambda_m : m\lambda_m \le \lambda_0} \{ j \notin \widehat{\mathcal{A}}_{\lambda_m} \} \Big) \ge c_1(\lambda_0).$$

This ends the proof of Assumption 2 for the adaptive lasso.

(iii): The SCAD. Fan and Li (2001) showed that the SCAD is selection consistent under the random design when $\sqrt{m}\lambda_m \to \infty$ and $\lambda_m \to 0$. In addition, condition (3) in Assumption 1 follows after the verification of Assumption 2 by similar approach as in the proof of lasso regression case.

Next, we show Assumption 2 for SCAD. It then suffices to consider the event $j \notin \widehat{\mathcal{A}}_{\lambda_m}$ for any $j \in \mathcal{A}_T^c$. In fact, the SCAD minimizes

$$Q(\beta) = \Big\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{z}_{(j)} \beta_j \Big\|^2 + m \sum_{j=1}^{p} p_{\lambda_m}(|\beta_j|),$$

where the penalty term satisfies $p'_\lambda(\theta) = \lambda\big(I(\theta \le \lambda) + \frac{(\gamma\lambda - \theta)_+}{(\gamma-1)\lambda} I(\theta > \lambda)\big)$ for some $\gamma > 2$ and $\theta > 0$. For any $\beta \in \{\beta : \beta - \hat{\beta}_m = O_P(m^{-1/2})\}$, then

$$\begin{aligned} \frac{\partial Q(\beta)}{\partial \beta_j} &= -2\mathbf{z}_{(j)}^T(\mathbf{y} - \mathbf{Z}\beta) + m p'_{\lambda_m}(|\beta_j|)\mathrm{sgn}(\beta_j) \\ &= -m\lambda_m \left(\frac{\frac{2\mathbf{z}_{(j)}^T \mathbf{Z}\sqrt{m}(\beta^* - \beta)}{m} + \frac{2\mathbf{z}_{(j)}^T \varepsilon}{\sqrt{m}}}{\sqrt{m}\lambda_m} - \frac{p'_{\lambda_m}(|\beta_j|)\mathrm{sgn}(\beta_j)}{\lambda_m} \right), \end{aligned}$$

where $\frac{\mathbf{Z}^T\mathbf{Z}}{m} \xrightarrow{p} \mathbf{C}$, $\|\sqrt{m}(\beta^* - \beta)\| \le \|\sqrt{m}(\beta^* - \hat{\beta}_m)\| + \|\sqrt{m}(\hat{\beta}_m - \beta)\|$ is bounded in probability due to that fact that $\hat{\beta}_m$ is a $\sqrt{m}$-consistent estimate of β^* by Theorem 1 of Fan and Li (2001), and $2\mathbf{z}_{(j)}^T\varepsilon/\sqrt{m} \xrightarrow{d} N(0,\Delta_3)$ for some Δ_3. Here condition (4) can be verified by similar approach as in lasso regression case since we have the asymptotic normality of $\hat{\beta}_m$. In addition, $p'_{\lambda_m}(|\beta_j|)/\lambda_m = I(\theta \le \lambda_m) + \frac{(\gamma\lambda_m - \theta)_+}{(\gamma-1)\lambda_m} I(\theta > \lambda_m) \le 1$. Therefore, we have

$$\begin{aligned} & \bigcap_{\lambda_m : \sqrt{m}\lambda_m \le \lambda_0} \left\{ \left| \frac{\frac{2\mathbf{z}_{(j)}^T \mathbf{Z}\sqrt{m}(\beta^* - \beta)}{m} + \frac{2\mathbf{z}_{(j)}^T \varepsilon}{\sqrt{m}}}{\sqrt{m}\lambda_m} \right| > \left| \frac{p'_{\lambda_m}(|\beta_j|)\mathrm{sgn}(\beta_j)}{\lambda_m} \right| \right\} \\ = & \bigcap_{\lambda_m : \sqrt{m}\lambda_m \le \lambda_0} \left\{ \left| \frac{2\mathbf{z}_{(j)}^T \mathbf{Z}\sqrt{m}(\beta^* - \beta)}{m} + \frac{2\mathbf{z}_{(j)}^T \varepsilon}{\sqrt{m}} \right| > \sqrt{m}\lambda_m \frac{p'_{\lambda_m}(|\beta_j|)}{\lambda_m} \right\} \\ \supseteq & \left\{ \left| \frac{2\mathbf{z}_{(j)}^T \mathbf{Z}\sqrt{m}(\beta^* - \beta)}{m} + \frac{2\mathbf{z}_{(j)}^T \varepsilon}{\sqrt{m}} \right| > \lambda_0 \right\}, \end{aligned}$$

and hence

$$\begin{aligned} & P\left(\bigcap_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0}\left\{\left|\frac{2\mathbf{z}_{(j)}^T\mathbf{Z}\sqrt{m}(\beta^*-\beta)}{m}+\frac{2\mathbf{z}_{(j)}^T\varepsilon}{\sqrt{m}}\right|>\left|\frac{p'_{\lambda_m}(|\beta_j|)\mathrm{sgn}(\beta_j)}{\lambda_m}\right|\right\}\right) \\ \ge\ & P\left(\left|\frac{2\mathbf{z}_{(j)}^T\mathbf{Z}\sqrt{m}(\beta^*-\beta)}{m}+\frac{2\mathbf{z}_{(j)}^T\varepsilon}{\sqrt{m}}\right|>\lambda_0\right). \end{aligned}$$

Therefore, there exists a positive probability $c_2(\lambda_0)\in(0,1)$, uniformly on λ_m,

$$\begin{aligned} \frac{\partial Q(\beta)}{\partial\beta_j} &< 0 \text{ when } 0<\beta_j<Mm^{-1/2}; \\ \frac{\partial Q(\beta)}{\partial\beta_j} &> 0 \text{ when } -Mm^{-1/2}<\beta_j<0, \end{aligned}$$

with M sufficient large such that $P\big(\sup_{\|u\|=M}Q\big(\beta^*+(m^{-1/2}+a_m)u\big)>Q(\beta^*)\big)\to 1$ and $a_m=\max\{p'_{\lambda_m}(|\beta_j^*|):\beta_j^*\neq 0\}$, which implies that for sufficiently large m $P(\cap_{\lambda_m:\sqrt{m}\lambda_m\le\lambda_0}\{\hat\beta_j\neq 0\})\ge c_3(\lambda_0)$ with $c_3(\lambda_0)$ strictly positive for fixed λ_0, and $c_3(\lambda_0)$ converges to 1 as $\lambda_0\to 0$. Therefore, condition (5) in Assumption 2 is verified. By similar approach, we can replace $\sqrt{m}\lambda_m\le\lambda_0$ with $\sqrt{m}\lambda_m\ge\tilde\lambda_0$, and bound the probability of the event $|2\mathbf{z}_{(j)}^T\mathbf{Z}\sqrt{m}(\beta^*-\beta)/m+2\mathbf{z}_{(j)}^T\varepsilon/\sqrt{m}|>\tilde\lambda_0$. Then condition (6) in Assumption 2 can be verified. Therefore, Assumptions 2 is satisfied by the SCAD with $r_m=m^{-1/2}$ and $s_m=o(1)$. ■

Remark: The convergence in (1) is valid under the fixed design with Assumption *S*2.

Assumption S2: Assume that $\frac{1}{n}\sum_{i=1}^n x_i x_i^T\to C$ with C positive definite, and $\frac{1}{n}\sum_{i=1}^n x_{ij}^2x_{ik}^2$ is finite for any $1\le j,k\le p$.

Proof of Remark: For random variable w_i as defined above, we can also show that

$$S_n=\frac{1}{n}\sum_{i=1}^n 2w_ix_{ij}x_{ik}\xrightarrow{p}C_{jk}. \tag{8}$$

Note that $E(S_n)=\frac{1}{n}\sum_{i=1}^n 2E(w_i)x_{ij}x_{ik}=\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}\to C_{jk}$ and

$$\begin{aligned} & \mathrm{var}(S_n)=E\Big(\Big(\frac{1}{n}\sum_{i=1}^n 2w_ix_{ij}x_{ik}\Big)^2\Big)-C_{jk}^2 \\ =\ & \frac{4}{n^2}\Big(\sum_{i=1}^n E(w_i^2)x_{ij}^2x_{ik}^2+\sum_{i\neq l}E(w_iw_l)x_{ij}x_{ik}x_{lj}x_{lk}\Big)-C_{jk}^2 \\ =\ & \frac{4}{n^2}\Big(\frac{1}{2}\sum_{i=1}^n x_{ij}^2x_{ik}^2+\sum_{i\neq l}\mathrm{cov}(w_iw_l)x_{ij}x_{ik}x_{lj}x_{lk}\Big)+\Big(\frac{1}{n^2}\sum_{i\neq l}x_{ij}x_{ik}x_{lj}x_{lk}-C_{jk}^2\Big) \\ \to\ & 0, \end{aligned}$$

following from Assumption *S*2 and the fact that $\mathrm{cov}(w_i,w_l)<0$ for $i\neq l$, $\sum_{i\neq l}\frac{x_{ij}x_{ik}}{n}\frac{x_{lj}x_{lk}}{n}\to C_{jk}^2\ge 0$. Then the Chebyshev's inequality implies that

$$P\Big(\Big|S_n-\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}\Big|\ge\varepsilon\Big)\le\frac{\mathrm{var}(S_n)}{\varepsilon^2}\to 0,\ \text{as } n\to\infty.$$

This together with the fact $\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} \to C_{jk}$ imply (8). ■

Lemma 3 *Suppose that Assumption S1 is met and $p_n = o(n^{1/4})$. Assumptions 1a and 2a are satisfied by the lasso regression with $r_n = n^{-1/2}$ and $s_n = o(1)$ under the assumptions (A1)-(A4) in Bach (2009) on the random splitting subsamples generated in Algorithm 1.*

Proof of Lemma 3: According to the similar techniques in the proof of Lemma 2, we only need to validate conditions (8)-(10) in Assumption 2a, then condition (7) in Assumption 1a is a direct result from Assumption 2a as shown in Lemma 2.

For a particular sign pattern $\tilde{s}$ with jth noise variable selected in the active set J, we have $\{j \in J\} \supseteq \{\text{sign}(\hat{\beta}_n) = \tilde{s}\}$, where $\{\text{sign}(\hat{\beta}_n) = \tilde{s}\}$ is equivalent to the conditions (2)-(4) with $s = \tilde{s}$. We first show the probability of (5) also tends to zero for diverging p_n. According to Proposition 2.4 in Bach (2009), Berry-Esseen inequality implies that

$$P(S3) \geq 1 - 2p_n e^{-C_1 \frac{n}{p_n}},$$

where C_1 is a positive constant independent of n and p_n. Therefore, from (5),

$$P\Big(\|(Q_{J,J}^{-1} q_J)_{\mathbf{J}}\|_2 > \frac{M(\beta^*)}{2}\Big) \leq 2p_n e^{-C_1 \frac{n}{p_n}}.$$

In addition, $\sqrt{n}M(\beta^*)\lambda_{\min}(Q)/(2\sqrt{p_n})$ is unbounded and hence the right hand side in (5) having probability tending to 0 when $p_n = o(n)$. Next, we can show that (6) is also valid for diverging p_n based on Proposition 2.4 in Bach (2009). Specifically,

$$P\left(\bigcup_{\lambda_n:\sqrt{n}\lambda_n \leq \lambda_0} \{(2),(4)\}^c\right) - P(\{v \notin \mathcal{C}(\tilde{s},\lambda_0)\}) \leq C_2 \frac{p_n^2}{n^{1/2}}, \tag{9}$$

where C_2 is a constant independent of n and p_n, and $\mathcal{C}(\tilde{s},\lambda_0)$ is defined in (6). Here $P(\{v \notin \mathcal{C}(\tilde{s},\lambda_0)\})$ is strictly within $(0,1)$ for any fixed λ_0, and when λ_0 converges to 0, $P(v \notin \mathcal{C}(\tilde{s},\lambda_0)) \to 0$. Therefore, for $p_n = o(n^{1/4})$ and $j \in \mathcal{A}_T^c$, as $n \to \infty$, combining (5) and (9) leads to

$$P\Big(\bigcap_{\lambda_n:\sqrt{n}\lambda_n \leq \lambda_0} \{j \in \widehat{\mathcal{A}}_{\lambda_n}\}\Big) \to 1 - P(v \notin \mathcal{C}(\tilde{s},\lambda_0)),$$

and hence condition (9) in Assumption 2a is verified. The condition (10) in Assumption 2a can be proved by defining the particular sign pattern $\tilde{s}$ to be the one with jth noise variable not selected in the active set J, then $\{j \notin J\} \supseteq \{\text{sign}(\hat{\beta}_n) = \tilde{s}\}$. Then all the proof can be derived following similar approach. In addition, Proposition 2.5 in Bach (2009) implies that uniformly over λ_n with $\sqrt{n}\lambda_n \leq \lambda_0$, all the important variable will be selected with probability tending to 1, which verifies (8) in Assumption 2a. Specifically, for any $j \in \mathcal{A}_T$,

$$P\Big(\bigcap_{\sqrt{n}\lambda_n \leq \lambda_0} \{j \in \widehat{\mathcal{A}}_{\lambda_n}\}\Big) \geq 1 - 2p_n e^{-C_3 \frac{n}{p_n}},$$

where C_3 is a positive constant independent of n and p_n. Therefore, let $\zeta_n = 2p_n e^{-C_3 \frac{n}{p_n}}$ satisfying condition (8). This ends the verification of Assumption 2a for the lasso regression. Finally, the condition (7) in Assumption 1a is also a direct result from Assumption 2a as shown in Lemma 2. This ends the verification for the lasso regression in Lemma 3. ■

References

F.R. Bach. Bolasso: Model consistent Lasso estimation through the bootstrap. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

F.R. Bach. Model-consistent sparse estimation through the bootstrap. In *Technical Report, Laboratoire d'Informatique de l'Ecole Normale Superieure, Paris*, 2009.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360, 2001.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541-2563, 2006.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429, 2006.