Weighted Lasso Bootstrap

Tun Lee Ng Michael A. Newton

August 18, 2018

1 Introduction

Consider the following linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \tag{1}$$

for $i = 1, \ldots, n$, and $\{\epsilon_i\}$ are independent and identically distributed (iid) random variables with mean 0 and finite variance σ^2. We assume that p is fixed. Without loss of generality, the covariates are centered to have mean 0, so that $\hat{\beta}_0 = \bar{Y}$, and Y_i in (1) can be replaced by $Y_i - \bar{Y}$. Again, without loss of generality, we assume that $\bar{Y} = 0$. Then, (1) can be expressed as the following

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \tag{2}$$

where Y_i is the centered response, $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$ is the $p \times 1$ centered covariate vector and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are the regression parameters.

Let $\boldsymbol{\beta}_0$ be the true values of the regression parameters $\boldsymbol{\beta}$. The model is assumed to be sparse, ie. some of the elements of $\boldsymbol{\beta}_0$ are exactly zero corresponding to predictors that are irrelevant to the response.

The Lasso estimator is defined to be the minimizer of the l_1-penalized least square objective function,

$$\widehat{\boldsymbol{\beta}}_n := \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \tag{3}$$

for a given penalty or regularization parameter λ_n. The Lasso estimator was first introduced by Tibshirani (1996). Knight and Fu (2000) obtained the asymptotic distribution of the Lasso estimator and showed that the Lasso is weakly consistent under some mild regularity condition. Chatterjee and Lahiri (2011) studied strong consistency of the Lasso estimator under a slightly more stringent regularity condition.

Following the idea by Newton and Raftery (1994), for a given set of responses $\boldsymbol{y} = (y_1, \ldots, y_n)'$, we define the weighted Lasso estimator as follows:

$$\widehat{\boldsymbol{\beta}}_n^w := \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \widetilde{w}_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda_n \widetilde{w}_{n+1} \sum_{j=1}^p |\beta_j|. \tag{4}$$

Here, $\widetilde{\boldsymbol{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_{n+1})$ are random weights drawn from

$$\left(\frac{W_1}{\sum_{i=1}^{n+1} W_i}, \ldots, \frac{W_{n+1}}{\sum_{i=1}^{n+1} W_i} \right) = \left(\frac{W_1}{(n+1)\overline{W}}, \ldots, \frac{W_{n+1}}{(n+1)\overline{W}} \right),$$

where $W_1, \ldots, W_n \overset{iid}{\sim} \exp(1)$ and $W_{n+1} = 1$ a.s. Note that the random weights $\widetilde{\boldsymbol{w}}$ are generated independently of the data $\boldsymbol{y}$, and are similar in structure to a Dirichlet weight vector as expounded by Newton and Raftery (1994).

Hence, for a given set of data $\boldsymbol{y}$, (4) can be expressed as

$$\widehat{\boldsymbol{\beta}}_n^w = \arg\min_{\boldsymbol{\beta}} \frac{1}{\overline{W}} \left\{ \frac{1}{n} \sum_{i=1}^n W_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \frac{\lambda_n}{n} \sum_{j=1}^p |\beta_j| \right\}. \tag{5}$$

For any given set of data, the sampling distribution of $\left\{ \widehat{\boldsymbol{\beta}}_{n,k}^w \right\}_{k=1}^K$ is induced by the randomly drawn weights $\{\widetilde{\boldsymbol{w}}_k\}_{k=1}^K$.

2 Asymptotics for WLB

Need to define the proper probability space...
Need to define "*convergence in conditional probability*" (denoted with $\xrightarrow{\text{c.p.}}$)...
Need to define "*convergence in conditional distribution*" (denoted with $\xrightarrow{\text{c.d.}}$)...

Need restrictions on the topology of parameter space ($\boldsymbol{\beta} \in \Theta$ for open, convex subset Θ of $\mathcal{R}^p$)...

Theorem 2.1. ***(Conditional Consistency)*** *Consider the linear regression model in* (2)*, where the error terms* $\{\epsilon_i\}$ *are iid with mean 0 and variance* σ^2*. Suppose that*

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|_2^2 = O(1), \qquad \textit{and} \qquad \frac{1}{n}\mathop{max}_{1\leq i\leq n}\|\boldsymbol{x}_i\|_2^2 \to 0,$$

and there exists a non-singular matrix C *such that*

$$X'X = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' \to C \quad \textit{as} \quad n \to \infty.$$

If $\dfrac{\lambda_n}{n} \to \lambda_0 \in [0,\infty)$*, then*

$$\left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right)\Bigg|\,\textit{data} \xrightarrow{c.p.} \arg\min Z,$$

where

$$Z(\boldsymbol{u}) = \boldsymbol{u}'C\boldsymbol{u} + \lambda_0\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1.$$

Hence, if $\lambda_0 = 0$, then $\widehat{\boldsymbol{\beta}}_n^w \xrightarrow{\text{c.p.}} \boldsymbol{\beta}_0$, ie. The weighted Lasso estimator is conditionally consistent if the penalty term vanishes in the limit.

Theorem 2.2. ***(Asymptotic Conditional Distribution)*** *Consider the linear regression model in* (2)*, where the error terms* $\{\epsilon_i\}$ *are iid with mean 0 and variance* σ^2*. Suppose that*

$$\mathbb{E}(\epsilon_i)^4 < \infty, \qquad \textit{and} \qquad \mathop{max}_{1\leq i\leq n}\|\boldsymbol{x}_i\|_2^2 = \mathcal{O}(1),$$

and there exists a non-singular matrix C *such that*

$$X'X = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' \to C \quad \textit{as} \quad n \to \infty.$$

If $\dfrac{\lambda_n}{\sqrt{n}} \to \lambda_0 \in [0, \infty)$*, then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{OLS}\right)\Bigg| data \xrightarrow{c.d.} \arg\min(V),$$

where $\widehat{\boldsymbol{\beta}}_n^{OLS}$ *are the ordinary least squares (OLS) estimators for the linear regression model in* (2),

$$V(\boldsymbol{u}) = -2\boldsymbol{u}'\boldsymbol{Z} + \boldsymbol{u}'C\boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} \left[u_j\ sgn(\beta_{0,j}) \mathbb{1}_{\{\beta_{0,j} \neq 0\}} + |u_j| \mathbb{1}_{\{\beta_{0,j}=0\}} \right],$$

and $\boldsymbol{Z} \sim N\left(\mathbf{0}, \sigma^2 C\right)$*.*

Hence, if $\lambda_0 = 0$, then $\arg\min_{\boldsymbol{u}} V(\boldsymbol{u}) = C^{-1}\boldsymbol{Z} \sim N\left(\mathbf{0}, \sigma^2 C^{-1}\right)$, which corresponds to the asymptotic normality result of the OLS estimator.

3 Appendix

Here are the proofs for the theorems in this paper.

Lemma 3.1. *For* $W_1, W_2, \ldots \overset{iid}{\sim} \exp(1)$*, let* $D_n = diag(W_1, \ldots, W_n)$*. If*

$$\frac{1}{n} \underset{1 \le i \le n}{max} \|\boldsymbol{x}_i\|_2^2 \to 0 \qquad and \qquad \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{x}_i = \mathcal{O}(1),$$

and there exists a non-singular matrix C *such that*

$$X'X = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' \to C \quad as \quad n \to \infty,$$

then

$$\frac{1}{n}X'D_nX \xrightarrow{p} C \tag{6}$$

Proof. Note that $X'X$ is symmetric and $\dfrac{1}{n}X'X \to C$, so C is symmetric. Therefore, by Lemma 1 of Strawderman (1994), we prove (6) by showing that for every $t > 0$,

$$P\left(\left\|\frac{1}{n}X'D_nX - C\right\|_\infty \ge t\right) \to 0 \qquad \text{as} \quad n \to \infty,$$

where $\|.\|_\infty$ refers to the spectral norm of a matrix. By Proposition 6.2 of Mackey et al. (2014),

$$\begin{aligned}
&P\left(\left\|\frac{1}{n}X'D_nX - C\right\|_\infty \geq t\right)\\
&\leq \frac{1}{t^2}\mathbb{E}\left\|\frac{1}{n}X'D_nX - C\right\|_2^2\\
&= \frac{1}{t^2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n W_i\boldsymbol{x}_i\boldsymbol{x}_i' - C\right\|_2^2\\
&= \frac{1}{t^2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n (W_i-1)\boldsymbol{x}_i\boldsymbol{x}_i' + \frac{1}{n}X'X - C\right\|_2^2\\
&= \frac{1}{t^2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n (W_i-1)\boldsymbol{x}_i\boldsymbol{x}_i'\right\|_2^2 + \frac{1}{t^2}\left\|\frac{1}{n}X'X - C\right\|_2^2\\
&= \frac{1}{n^2t^2}\sum_{i=1}^n \|\boldsymbol{x}_i\boldsymbol{x}_i'\|_2^2 + \frac{1}{t^2}\left\|\frac{1}{n}X'X - C\right\|_2^2\\
&= \frac{1}{n^2t^2}\sum_{i=1}^n (\boldsymbol{x}_i'\boldsymbol{x}_i)^2 + \frac{1}{t^2}\left\|\frac{1}{n}X'X - C\right\|_2^2\\
&\leq \frac{1}{nt^2}\max_{1\leq i\leq n}\boldsymbol{x}_i'\boldsymbol{x}_i \times \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i + \frac{1}{t^2}\left\|\frac{1}{n}X'X - C\right\|_2^2\\
&\to 0 \qquad \text{as} \quad n \to \infty,
\end{aligned}$$

where the last step follows from our assumptions. □

Lemma 3.2. *Let the error terms $\{\epsilon_i\}$ be iid with mean 0 and variance σ^2. Suppose that*

$$\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{x}_i\|_2^2 = \mathcal{O}(1), \qquad \textit{and} \qquad \frac{1}{n}\underset{1\leq i\leq n}{max}\|\boldsymbol{x}_i\|_2^2 \to 0,$$

Then,

$$\left(\frac{1}{n}X'D_n\boldsymbol{\epsilon}\right)\bigg|\, data \xrightarrow{c.p.} \mathbf{0}. \tag{7}$$

Proof. First, by Jensen's inequality, we have

$$\mathbb{E}|X| = \mathbb{E}(\sqrt{X^2}) \leq \sqrt{\mathbb{E}(X^2)} \leq \mathbb{E}X^2 = \sigma^2 < \infty.$$

Since $\mathbb{E}(\epsilon_i) = 0$ and we assumed

$$\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2^2 = \mathcal{O}(1),$$

then, by Lemma 3.1 of Chatterjee and Lahiri (2011),

$$\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \boldsymbol{x}_i \xrightarrow{\text{a.s.}} \boldsymbol{0}.$$

In addition, note that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \|\epsilon_i \boldsymbol{x}_i\|_2 &= \frac{1}{n}\sum_{i=1}^{n} |\epsilon_i| \|\boldsymbol{x}_i\|_2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 + \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2^2.\end{aligned}$$

The strong law of large number ensures that

$$\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \xrightarrow{\text{a.s.}} \mathbb{E}(\epsilon_1^2) = \sigma^2.$$

So,

$$\frac{1}{n}\sum_{i=1}^{n} \|\epsilon_i \boldsymbol{x}_i\|_2 = \mathcal{O}(1) \quad \text{a.s.} \tag{8}$$

Furthermore, note that

$$\begin{aligned}\frac{1}{n}\max_{1\leq i\leq n} \|\epsilon_i \boldsymbol{x}_i\|_2 &= \frac{1}{n}\max_{1\leq i\leq n} |\epsilon_i| \|\boldsymbol{x}_i\|_2 \\ &\leq \frac{1}{n}\max_{1\leq i\leq n} |\epsilon_i| \times \frac{1}{n}\max_{1\leq i\leq n} \|\boldsymbol{x}_i\|_2 \\ &\xrightarrow{\text{a.s.}} 0, \end{aligned} \tag{9}$$

where the last line follows from the fact that $\mathbb{E}|\epsilon_1| < \infty$, so $\dfrac{1}{n}\max_{1\le i\le n}|\epsilon_i| \xrightarrow{\text{a.s.}} 0$ by Lemma 14 of Newton (1991).

Conditional on data, $\boldsymbol{\epsilon}$ is fixed albeit unobservable. Hence, for every $t > 0$, we deploy the multi-dimensional Chebyshev inequality

$$P\left(\left\|\frac{1}{n}X'D_n\boldsymbol{\epsilon} - \mathbf{0}\right\|_2 \ge t \middle| \text{data}\right) \le \frac{1}{t^2}\mathbb{E}_{\boldsymbol{W}}\left\{\left\|\frac{1}{n}X'D_n\boldsymbol{\epsilon}\right\|_2^2 \middle| \text{data}\right\}$$

to show that

$$P\left(\left\|\frac{1}{n}X'D_n\boldsymbol{\epsilon} - \mathbf{0}\right\|_2 \ge t \middle| \text{data}\right) \to 0 \qquad \text{as} \qquad n \to \infty$$

due to the constraints in (8) and (9). Then, by Lemma 3 of Newton (1991),

$$\left(\frac{1}{n}X'D_n\boldsymbol{\epsilon}\right)\Bigg|\text{data} \xrightarrow{\text{c.p.}} \mathbf{0}.$$

□

Now we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. From (5), conditional on data,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^w &= \arg\min_{\boldsymbol{\beta}} \frac{1}{\overline{W}}\left\{\frac{1}{n}(Y - X\boldsymbol{\beta})'D_n(Y - X\boldsymbol{\beta}) + \frac{\lambda_n}{n}\|\boldsymbol{\beta}\|_1\right\} \\ &= \arg\min_{\boldsymbol{\beta}} \frac{1}{\overline{W}}\left\{\frac{1}{n}[\boldsymbol{\epsilon} - X(\boldsymbol{\beta} - \boldsymbol{\beta}_0)]'D_n[\boldsymbol{\epsilon} - X(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] + \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0 + \boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0) &= \arg\min_{\boldsymbol{u}} \frac{1}{\overline{W}}\left\{\frac{1}{n}(\boldsymbol{\epsilon} - X\boldsymbol{u})'D_n(\boldsymbol{\epsilon} - X\boldsymbol{u}) + \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1\right\} \\ &= \arg\min_{\boldsymbol{u}} \frac{1}{\overline{W}}\left\{\frac{1}{n}[-2\boldsymbol{u}'(X'D_n\boldsymbol{\epsilon}) + \boldsymbol{u}'(X'D_nX)\boldsymbol{u}] + \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1\right\}.\end{aligned}$$

The strong law of large numbers ensures that $\overline{W} \xrightarrow{\text{a.s.}} 1$. Let

$$Z_n(\boldsymbol{u}) := -2\boldsymbol{u}'\left(\frac{X'D_n\boldsymbol{\epsilon}}{n}\right) + \boldsymbol{u}'\left(\frac{X'D_nX}{n}\right)\boldsymbol{u} + \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1.$$

By Lemma 3.1, we have

$$\frac{1}{n}X'D_nX \xrightarrow{\text{p}} C.$$

By Lemma 3.2, we have

$$\left(\frac{1}{n}X'D_n\boldsymbol{\epsilon}\right)\Big|\text{data} \xrightarrow{\text{c.p.}} \mathbf{0}.$$

Hence, if $\frac{\lambda_n}{n} \to \lambda_0 \in [0, \infty)$, then by Lemma 2 of Newton (1991),

$$Z_n(\boldsymbol{u})\big|\text{data} \xrightarrow{\text{c.p.}} Z(\boldsymbol{u}) \equiv \boldsymbol{u}'C\boldsymbol{u} + \lambda_0\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|.$$

Note that $Z_n(\boldsymbol{u})$ is a sequence of random convex functions of $\boldsymbol{u}$. Hence, by the Convexity Lemma (Pollard, 1991), for a compact set $K \subset \Theta$, where Θ is itself a convex, open subset of $\mathcal{R}^p$,

$$\sup_{\boldsymbol{u}\in K\subset\Theta} |Z_n(\boldsymbol{u}) - Z(\boldsymbol{u})|\Big|\text{data} \xrightarrow{\text{c.p.}} 0.$$

Also, note that $\left(\widehat{\boldsymbol{\beta}}_n^w\big|\text{data}\right) = O_p(1)$. Therefore,

$$\begin{aligned}&\left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right)\Big|\text{data}\\ &= \arg\min_{\boldsymbol{u}}\left\{\frac{1}{\overline{W}}Z_n(\boldsymbol{u})\Big|\text{data}\right\}\\ &\xrightarrow{\text{c.p.}} \arg\min_{\boldsymbol{u}} Z(\boldsymbol{u}).\end{aligned}$$

It follows that if $\lambda_0 = 0$, then $\arg\min_{\boldsymbol{u}} Z(\boldsymbol{u}) = \mathbf{0}$, i.e. $\widehat{\boldsymbol{\beta}}_n^w \xrightarrow{\text{c.p.}} \boldsymbol{\beta}_0$. □

Lemma 3.3. *Consider two sequences $\{V_n\}$ and $\{U_n\}$ and two other random variables V and U, all defined on the same product space $(\Omega, \mathcal{F})$. If*

$$V_n \xrightarrow{c.p.} V \qquad \textit{and} \qquad U_n \xrightarrow{c.d.} U,$$

then

$$V_nU_n \xrightarrow{c.d.} VU \qquad \textit{and} \qquad V_n + U_n \xrightarrow{c.d.} V + U.$$

Proof. For each fixed infinite sequence of data, the results follow from properties of convergence in distribution due to Slutsky's theorem. □

Lemma 3.4. *Suppose that*

$$\underset{1\leq i\leq n}{max}\|\boldsymbol{x}_i\|_2^2 = \mathcal{O}(1),$$

and there exists a non-singular matrix C such that

$$X'X = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' \to C \quad as \quad n\to\infty,$$

then

$$\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\Bigg| data \xrightarrow{c.p.} \sigma^2 C,$$

where $\{e_i\}$ are the OLS residuals.

Proof. Without loss of generality, we first consider the univariate case. Since

$$\widehat{Y}_i^{\text{OLS}} = x_i\widehat{\beta}_n^{\text{OLS}} = x_i\frac{\sum_{i=1}^n x_iY_i}{\sum_{i=1}^n x_i^2},$$

then, under the OLS approach,

$$\begin{aligned}
e_i^2 &= \left(Y_i - \widehat{Y}_i^{\text{OLS}}\right)^2\\
&= Y_i^2 + x_i^2\left(\widehat{\beta}_n^{\text{OLS}}\right)^2 - 2\widehat{\beta}_n^{\text{OLS}}x_iY_i\\
&= (x_i\beta_0+\epsilon_i)^2 + x_i^2\left(\widehat{\beta}_n^{\text{OLS}}\right)^2 - 2\widehat{\beta}_n^{\text{OLS}}x_i(x_i\beta_0+\epsilon_i)\\
&= \epsilon_i^2 + x_i^2\left[\beta_0^2 + \left(\widehat{\beta}_n^{\text{OLS}}\right)^2 - 2\left(\beta_0\widehat{\beta}_n^{\text{OLS}}\right)\right] - 2x_i\epsilon_i\left(\beta_0\widehat{\beta}_n^{\text{OLS}}\right),
\end{aligned}$$

which leads us to

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} x_i^2e_i^2\\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2\epsilon_i^2 + \left[\beta_0^2 + \left(\widehat{\beta}_n^{\text{OLS}}\right)^2 - 2\left(\beta_0\widehat{\beta}_n^{\text{OLS}}\right)\right]\left(\frac{1}{n}\sum_{i=1}^{n} x_i^4\right) - 2\left(\beta_0\widehat{\beta}_n^{\text{OLS}}\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i^3\epsilon_i\right).
\end{aligned}$$

We have assumed that

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \to c \quad \text{for some } c>0.$$

From the assumption $\max_{1\le i\le n} x_i^2 = \mathcal{O}(1)$, we have

$$\frac{1}{n}\sum_{i=1}^n x_i^4 = \mathcal{O}(1).$$

and the Strong Law of Large Numbers gives us

$$\frac{1}{n}\sum_{i=1}^n x_i\epsilon_i \xrightarrow{\text{a.s.}} 0. \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^n x_i^3\epsilon_i \xrightarrow{\text{a.s.}} 0.$$

Therefore,

$$\widehat{\beta}_n^{\text{OLS}} \xrightarrow{\text{a.s.}} \frac{\beta_0 c}{c} = \beta_0,$$

that is, the OLS estimator is strongly consistent. Again, the Strong Law of Large Numbers ensures that

$$\frac{1}{n}\sum_{i=1}^n x_i^2(\epsilon_i - \sigma^2) \xrightarrow{\text{a.s.}} 0,$$

while

$$\sigma^2 \times \frac{1}{n}\sum_{i=1}^n x_i^2 \to \sigma^2 c.$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^n x_i^2\epsilon_i^2 \xrightarrow{\text{a.s.}} \sigma^2 c.$$

Finally, piecing the terms together, we have

$$\frac{1}{n}\sum_{i=1}^n x_i^2 e_i^2 \xrightarrow{\text{a.s.}} \sigma^2 c,$$

and hence

$$\left(\frac{1}{n}\sum_{i=1}^n x_i^2 e_i^2\right)\Bigg|\text{data} \xrightarrow{\text{c.p.}} \sigma^2 c.$$

A sketch of proof is also provided for the multivariate case. First, note that the assumption

$$\frac{1}{n}X'X \to C$$

implies that $X'X = \mathcal{O}(n^{-1})$. Therefore, by Theorem 1 of Lai et al. (1978),

$$\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}_0.$$

Then,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' \\
&= \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' \\
&+ \frac{2}{n}\sum_{i=1}^{n} \epsilon_i \boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} - \boldsymbol{\beta}_0\right) \boldsymbol{x}_i \boldsymbol{x}_i' \\
&+ \frac{1}{n}\sum_{i=1}^{n} \left\{ \boldsymbol{x}_i' \left[\boldsymbol{\beta}_0 \boldsymbol{\beta}_0' - 2\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \boldsymbol{\beta}_0' + \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right)'\right] \boldsymbol{x}_i \right\} \boldsymbol{x}_i \boldsymbol{x}_i'.
\end{aligned}$$

With our assumptions, the Strong Law of Large Numbers ensures that the first term converges to $\sigma^2 C$ with probability 1 whereas the other two terms converges to zero matrix almost surely. Therefore,

$$\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right) \Bigg| \text{data} \xrightarrow{\text{c.p.}} \sigma^2 C.$$

□

Lemma 3.5. *Suppose that*

$$\underset{1 \le i \le n}{max} \|\boldsymbol{x}_i\|_2^2 = \mathcal{O}(1) \qquad \textit{and} \qquad \mathbb{E}(\epsilon_i) < \infty.$$

Then,

$$\left(\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n^{OLS}\right) \Bigg| \textit{data} \xrightarrow{c.d.} N\left(\boldsymbol{0}, \sigma^2 C\right).$$

Proof. First, note that

$$\begin{aligned}
&\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n^{\text{OLS}} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{1/2} \times \left(\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1/2} \times \sum_{i=1}^{n} e_i \boldsymbol{x}_i W_i
\end{aligned}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{1/2} \times \left(\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1/2} \times \sum_{i=1}^{n} e_i \boldsymbol{x}_i (W_i - 1),$$

where the last equality follows from the fact that

$$\begin{aligned}\sum_{i=1}^{n} e_i \boldsymbol{x}_i &= X' \boldsymbol{e}_n^{\text{OLS}} \\ &= X'Y - X'X(X'X)^{-1}X'Y \\ &= \mathbf{0}.\end{aligned}$$

By Lemma 3.4,

$$\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{1/2} \Bigg| \text{data} \xrightarrow{\text{c.p.}} \sigma C^{1/2}.$$

Without loss of generality, we will continue our proof for the univariate case. We shall show that the Lindeberg's Central Limit Theorem gives

$$\left(\frac{\sum_{i=1}^{n} e_i x_i (W_i - 1)}{\sqrt{\sum_{i=1}^{n} e_i^2 x_i^2}}\right) \Bigg| \text{data} \xrightarrow{\text{c.d.}} N(0,1)$$

by verifying the following Liapounov's sufficient condition

$$\frac{\sum_{i=1}^{n} \mathbb{E}\left[e_i^4 x_i^4 (W_i - 1)^4 | \text{data}\right]}{\left(Var\left[\sum_{i=1}^{n} e_i x_i (W_i - 1) | \text{data}\right]\right)^2} \to 0 \quad \text{as } n \to \infty.$$

With our assumptions that $\max\limits_{1 \le i \le n} x_i^2 = \mathcal{O}(1)$ and $\mathbb{E}(\epsilon_i) < \infty$, we can use similar technique in Lemma 3.4 to show that

$$\sum_{i=1}^{n} e_i^4 x_i^4 = \mathcal{O}(n) \text{ a.s.}$$

Since $\mathbb{E}\left[(W_i - 1)^4\right] = 9$,

$$\sum_{i=1}^{n} \mathbb{E}\left[e_i^4 x_i^4 (W_i - 1)^4 | \text{data}\right] = \mathcal{O}(n)$$

On the other hand, by Lemma 3.4,

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 e_i^2\right)^2 \xrightarrow{\text{a.s.}} \sigma^4 c^2,$$

which implies that

$$\left(\sum_{i=1}^{n} x_i^2 e_i^2\right)^2 = \mathcal{O}(n^2) \text{ a.s.}$$

Hence,

$$\begin{aligned}
&\left(Var\left[\sum_{i=1}^{n} e_i x_i (W_i - 1)\middle|\text{data}\right]\right)^2 \\
&= \left(\sum_{i=1}^{n} x_i^2 e_i^2\right)^2 \middle|\text{data} \\
&= \mathcal{O}(n^2)
\end{aligned}$$

Therefore, conditional on data,

$$\sum_{i=1}^{n} e_i^4 x_i^4 \mathbb{E}\left[(W_i - 1)^4\right] = o\left[\left(\sum_{i=1}^{n} e_i^2 x_i^2\right)^2\right],$$

that is, the Liapounov's sufficient condition is satisfied. Finally, we apply Lemma 3.3 to obtain

$$\left(\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n^{\text{OLS}}\right)\middle|\text{data} \xrightarrow{\text{c.d.}} N\left(\mathbf{0}, \sigma^2 C\right).$$

□

Now we are ready to prove Theorem 2.2.

Proof of Theorem 2.2. Define

$$Q_n(\boldsymbol{z}) := \left\| D_n^{\frac{1}{2}} (\boldsymbol{y} - X\boldsymbol{z}) \right\|_2^2 + \lambda_n \|\boldsymbol{z}\|_1,$$

which leads to

$$\begin{aligned}
Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) &= \left\| D_n^{\frac{1}{2}} \left[Y - X\left(\widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right)\right] \right\|_2^2 + \lambda_n \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u} \right\|_1 \\
&= \left\| D_n^{\frac{1}{2}} \left(\boldsymbol{e}_n^{\text{OLS}} - \frac{1}{\sqrt{n}} X\boldsymbol{u}\right) \right\|_2^2 + \lambda_n \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u} \right\|_1,
\end{aligned}$$

and

$$\begin{aligned}Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}\right) &= \left\| D_n^{\frac{1}{2}}\left(Y - X\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}\right)\right\|_2^2 + \lambda_n \left\|\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}\right\|_1 \\ &= \left\| D_n^{\frac{1}{2}}\boldsymbol{e}_n^{\mathrm{OLS}}\right\|_2^2 + \lambda_n \left\|\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}\right\|_1 .\end{aligned}$$

Now define

$$V_n(\boldsymbol{u}) := Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) - Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}\right),$$

such that

$$\arg\min_{\boldsymbol{u}} V_n(\boldsymbol{u}) = \arg\min_{\boldsymbol{u}} Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) = \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{w} - \widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}\right).$$

Notice that $V_n(\boldsymbol{u})$ can be simplified into

$$-2\boldsymbol{u}'\left(\frac{X'D_n\boldsymbol{e}_n^{\mathrm{OLS}}}{\sqrt{n}}\right) + \boldsymbol{u}'\left(\frac{X'D_nX}{n}\right)\boldsymbol{u} + \lambda_n\left\{\left\|\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right\|_1 - \left\|\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}\right\|_1\right\}.$$

By Lemma 3.1,

$$\frac{1}{n}X'D_nX \xrightarrow{\mathrm{p}} C.$$

By Lemma 3.5,

$$\frac{1}{\sqrt{n}}X'D_n\boldsymbol{e}_n^{\mathrm{OLS}} \xrightarrow{\mathrm{c.d.}} N\left(\mathbf{0}, \sigma^2 C\right).$$

For the penalty term,

$$\begin{aligned}&\lambda_n\left\{\left\|\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} + \frac{1}{\sqrt{n}}\boldsymbol{u}\right\|_1 - \left\|\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}\right\|_1\right\} \\ &= \frac{\lambda_n}{\sqrt{n}}\sum_{j=1}^{p}\left\{\left|\sqrt{n}\widehat{\beta}_{n,j}^{\mathrm{OLS}} + \mu_j\right| - \left|\sqrt{n}\widehat{\beta}_{n,j}^{\mathrm{OLS}}\right|\right\} \\ &:= \frac{\lambda_n}{\sqrt{n}}\sum_{j=1}^{p} p_n(u_j).\end{aligned}$$

We assumed, for Theorem 2.2, that

$$\frac{\lambda_n}{\sqrt{n}} \to \lambda_0 \in [0, \infty).$$

When $\widehat{\beta}_{n,j}^{\text{OLS}} = 0$, $p_n(u_j) = |u_j|$. Also, for large n, $\sqrt{n}\widehat{\beta}_{n,j}^{\text{OLS}}$ dominates u_j. Hence, it is easy to verify that, for large n, we have

$$p_n(u_j) = u_j \text{sgn}\left(\widehat{\beta}_{n,j}^{\text{OLS}}\right) \mathbb{1}_{\left\{\widehat{\beta}_{n,j}^{\text{OLS}} \neq 0\right\}} + |u_j| \mathbb{1}_{\left\{\widehat{\beta}_{n,j}^{\text{OLS}} = 0\right\}}$$

based on the following observations

- when $\widehat{\beta}_{n,j}^{\text{OLS}} > 0$ and $u_j > 0$, then $p_n(u_j) = u_j$;
- when $\widehat{\beta}_{n,j}^{\text{OLS}} > 0$ and $u_j < 0$, then $p_n(u_j) = u_j$;
- when $\widehat{\beta}_{n,j}^{\text{OLS}} < 0$ and $u_j > 0$, then $p_n(u_j) = -u_j$;
- when $\widehat{\beta}_{n,j}^{\text{OLS}} < 0$ and $u_j < 0$, then $p_n(u_j) = -u_j$.

On the other hand, we have shown that $\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}$ is strongly consistent. Therefore, conditional on data,

$$\begin{aligned} &\lambda_n \left\{ \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} + \frac{1}{\sqrt{n}} \boldsymbol{u} \right\|_1 - \left\| \widehat{\boldsymbol{\beta}}_n^{\text{OLS}} \right\|_1 \right\} \\ &\to \lambda_0 \sum_{j=1}^{p} \left[u_j \, \text{sgn}(\beta_{0,j}) \mathbb{1}_{\{\beta_{0,j} \neq 0\}} + |u_j| \mathbb{1}_{\{\beta_{0,j} = 0\}} \right]. \end{aligned}$$

Then, by Lemma 3.3,

$$V_n(\boldsymbol{u}) \xrightarrow{\text{c.d.}} V(\boldsymbol{u}) \equiv -2\boldsymbol{u}'\boldsymbol{Z} + \boldsymbol{u}'C\boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} \left[u_j \, \text{sgn}(\beta_{0,j}) \mathbb{1}_{\{\beta_{0,j} \neq 0\}} + |u_j| \mathbb{1}_{\{\beta_{0,j} = 0\}} \right],$$

where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \sigma^2 C)$. Finally, conditional on data, $V_n(\boldsymbol{u})$ is convex, and $V(\boldsymbol{u})$ has unique minimum. Therefore, it follows from Geyer (1996) that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{w} - \boldsymbol{\beta}_0\right) \Big| \text{data} = \arg\min_{\boldsymbol{u}} \left\{ V_n(\boldsymbol{u}) \big| \text{data} \right\} \xrightarrow{\text{c.d.}} \arg\min_{\boldsymbol{u}} V(\boldsymbol{u}).$$

□

References

Chatterjee, A. and Lahiri, S. N. (2011), "Strong consistency of lasso estimators," *Sankhya: The Indian Journal of Statistics, Series A*, 73, 55–78.

Geyer, C. (1996), "On the asymptotics of convex stochastic optimization," Unpublished manuscript.

Knight, K. and Fu, W. (2000), "Asymptotics for lasso-type estimators," *The Annals of Statistics*, 28, 1356–1378.

Lai, T. L., Robbins, H., and Wei, C. Z. (1978), "Strong consistency of least squares estimates in multiple regression," *Proceedings of National Academy of Sciences*, 75, 3034 – 3036.

Mackey, L., Jordan, M., Chen, R., Farrell, B., and Tropp, J. (2014), "Matrix concentration inequalities via the method of exchangable pairs," *The Annals of Probability*, 42, 906–945.

Newton, M. A. (1991), "The weighted likelihood bootstrap and an algorithm for prepivoting," Ph.D. thesis, University of Washington, Seattle.

Newton, M. A. and Raftery, A. (1994), "Approximate bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 56, 3–48.

Pollard, D. (1991), "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 7, 186–199.

Strawderman, R. L. (1994), "A note on necessary and sufficient conditions for proving that a random symmetric matrix converges to a given limit," *Statistics & Probability Letters*, 21, 367–370.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58, 267–288.