Lecture 7 bis : Methods for Regression

Lasso estimator

K. Meziani

Model

Consider the n-sample $(\{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1}^n$ such that

$$y_i = x_i^\top \beta^* + \epsilon_i \quad , i = 1, \dots, n$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $x_i \in \mathbb{R}^p$ and $\beta^* \in \mathbb{R}^p$. Fix $p \geq 2$ and $n \geq 1$.

Matrix form:

$$Y = \sum_{j=1}^{n} \beta_j^* X_j + \epsilon = X\beta^* + \epsilon, \tag{1}$$

where $X_j \in \mathbb{R}^n$, $X = [X_1 | \cdots | X_p]$ and $\epsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbb{I}_n)$.

Remark

This lecture is based on the paper of Bickel, Ritov and Tsybakov,*Simultaneous analysis of Lasso and Dantzig selector* (2009, AOS).

Notations

Sparsity set. For all $\beta \in \mathbb{R}^p$, we denote by $J(\beta)$ the sparsity set, the subset of indices $\{1, \cdots, p\}$ where the vector β has non-zero coordinates

$$J(\beta) = \{j : \beta_j \neq 0\}$$

Sparsity of β. The sparsity of the vector β is characterized by the value $M(\beta)$, the cardinality of $J(\beta)$:

$$M(\beta) = \sum_{j=1}^{p} \mathbb{1}_{\beta_j \neq 0} = |J(\beta)|$$

Notations

Gram matrix: $\Psi_n = \frac{X^T X}{n} = \frac{1}{n}\sum_{i=1}^{n} {x_{i,j}}^2$

Some norms. For all $a \in R^n$, $b \in \mathbb{R}^p$ and $J_0 \subseteq \{1, ..., p\}$, we denote

$$||a||_n^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2 \quad , \quad ||b||_1 = \sum_{j=1}^{p} |b_j| \quad , \quad ||b||_{2,J_0} = \sqrt{\sum_{j \in J_0} {b_j}^2} \quad \text{and} \quad ||b||_{1,J_0} = \sum_{j \in J_0} |b_j|.$$

Assumptions $\mathcal{H}$

- The Gram matrix Ψ_n is such that its diagonal elements are equal to 1.
- The sparsity index is such that $M(\beta^*) \leq s$ for $1 \leq s \leq p$.

Definition **LASSO**

$$\widehat{\beta}^{\lambda,L} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|_n{}^2 + 2\lambda\|\beta\|_1 \right\}$$

where the regularization parameter $\lambda > 0$.

Remarks

We are typically interested in the case where $p > n$ and even $p >> n \Rightarrow$ Gram matrix Ψ_n is degenerate, *i.e*

$$\min_{\substack{\delta \in \mathbb{R}^p \\ \delta \neq 0}} \frac{(\delta^\top \psi_n \delta)^{1/2}}{\sqrt{n}||\delta||_2} \equiv \min_{\substack{\delta \in \mathbb{R}^p \\ \delta \neq 0}} \frac{||X\delta||_2}{\sqrt{n}||\delta||_2} = 0$$

OLSE does not work in this case, since it requires positive definiteness of ψ_n, *i.e*

$$\min_{\substack{\delta \in \mathbb{R}^p \\ \delta \neq 0}} \frac{||X\delta||_2}{\sqrt{n}||\delta||_2} > 0$$

The Lasso require much weaker assumptions. Replace

- the min by the minimum over a restricted set of vectors
- the norm $||\delta||_2$ by the ℓ_2 norm of only a part of δ.

For $J_0 = J(\beta^*)$, one of the properties of the Lasso is that the residuals $\delta = \widehat{\beta}^{\lambda,L} - \beta^*$ satisfy, with probability close to 1,

$$||\delta||_{1,J_0^c} \leq 3||\delta||_{1,J_0}.$$

RE(s, c_0) Assumption

Restricted Eigenvalues
sparsity
Constant (3 comme vu ci-dessus)

For an integer s such that $1 \leq s \leq p$ and a positive number c_0, we assume that the following condition is satisfied:

$$\kappa(s, c_0) := \min_{\substack{J_0 \subseteq \{1,\ldots,p\} \\ |J_0| \leq s}} \min_{\substack{\delta \neq 0 \\ \|\delta\|_{1,J_0{}^c} \leq c_0 \|\delta\|_{1,J_0}}} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta\|_{2,J_0}} > 0.$$

hypothèse faible (ce qu'on recherche en recherche)

plus petit vp de la matrice restreinte à l'espace J0.

Theorem~1

[a]Consider model (1). Under $\mathcal{H}$ and RE$(s,3)$ Assumptions, the estimator $\widehat{\beta}^{\lambda,L}$, for $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, with $A > 2\sqrt{2}$ is such that with probability greater than $1 - p^{1-\frac{A^2}{8}}$ ::

petit

Inégalités de concentration

$$\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_2^2 \leq \frac{16A^2}{\kappa(s,3)^2}\sigma^2 s\log(p). \tag{2}$$

$$\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq \frac{16A}{\kappa(s,3)^2}\sigma s\sqrt{\frac{\log(p)}{n}}, \tag{3}$$

$$M(\widehat{\beta}^{\lambda,L}) \leq \frac{64\phi_{\max}}{\kappa(s,3)^2}, \tag{4}$$

→ valeur propre maximale

where $\phi_{\max}$ denote the maximal eigenvalue of the Gram matrix ψ_n.

---

[a]Bickel, Ritov and Tsybakov,*Simultaneous analysis of Lasso and Dantzig selector* (2009, AOS)

To prove this theorem, we need previous results.

Lemma~1

Let $V_j := \frac{1}{n}\sum_{i=1}^{n} \epsilon_i x_{i,j} = \frac{1}{n} X_j^\top \epsilon$, we define the set

$$\Omega = \bigcap_{j=1}^{p} \left\{ 2|V_j| \leq \lambda \right\} = \bigcap_{j=1}^{p} \underbrace{\left\{ \left| \frac{1}{n} X_j^\top \epsilon \right| \leq \lambda/2 \right\}}_{\text{ces variables n'ont pas été prises}}$$

espace où le Lasso arrête de sélectionner des variables car trop corrélées au bruit.

Then for $A > 2\sqrt{2}$ and $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, we have

$$\mathbb{P}(\Omega) > 1 - p^{1-\frac{A^2}{8}},$$

Proof of Lemma~1

Denote by $\eta = \sqrt{n}V_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i,j}\epsilon_i$, then

$$
\begin{aligned}
\mathbb{E}(\eta) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i,j}\mathbb{E}(\epsilon_i) = 0 \\
\mathbb{V}\mathrm{ar}(\eta) &= \frac{1}{n} \sum_{i=1}^{n} x_{i,j}^2 \mathbb{V}\mathrm{ar}(\epsilon_i) = \frac{\sigma^2}{n} \sum_{i=1}^{n} x_{i,j}^2 = \frac{\sigma^2}{n} X_j^\top X_j = \sigma^2 \quad \text{by Assumption } \mathcal{H}
\end{aligned}
$$

↳ éléments diagonaux

Then $\frac{\eta}{\sigma} \sim \mathcal{N}(0, 1)$ and

union des complémentaires

$$
\begin{aligned}
\mathbb{P}\left(\Omega^c\right) &= \mathbb{P}\left(\bigcup_{j=1}^{p} \{2|V_j| > \lambda\}\right) = \mathbb{P}\left(\bigcup_{j=1}^{p} \{2|\eta| > \lambda\sqrt{n}\}\right) \leq \sum_{j=1}^{p} \mathbb{P}\left(2|\eta| > \lambda\sqrt{n}\right) \\
&\leq p\mathbb{P}\left(2|\eta| > \lambda\sqrt{n}\right) = p\mathbb{P}\left(\left|\frac{\eta}{\sigma}\right| > \frac{\lambda\sqrt{n}}{2\sigma}\right) \leq pe^{-\frac{\lambda^2 n}{8\sigma^2}} = pe^{-A^2\frac{\log(p)}{8}} = p^{1-A^2/8}.
\end{aligned}
$$

$Z \sim \mathcal{N}(0,1)$

$\mathbb{P}(|Z| > t) \leq e^{-\frac{t^2}{2}}$

Donc Ω est vrai avec grande probabilité.

Lemma~2

Consider model (1). Under $\mathcal{H}$ Assuptions, The estimator $\widehat{\beta}^{\lambda,L}$ for $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, with $A > 2\sqrt{2}$ is such that with probability greater than $1 - p^{1-\frac{A^2}{8}}$:

$$\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n^{\,2} + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq 4\lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0}, \tag{5}$$

where $J_0 = J(\beta^*)$.

The proof will be given last

$\delta = \hat{\beta}^L - \beta^*$

Lemma 2 = $\|X\delta\|_n^2 + \lambda\|\delta\|_1 \leq 4\lambda\|\delta\|_{1,J_0}$

Consequence of the Lemma~2

With probability$> 1 - p^{1-\frac{A^2}{8}}$, we have

$$
\begin{aligned}
& \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq \overbrace{\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n^2}^{>0} + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \\
\Rightarrow \quad & \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq 4\lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} \\
\Rightarrow \quad & \|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq 4\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} \\
\Rightarrow \quad & \underbrace{\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0}}_{\text{somme sur } J_0} + \underbrace{\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c}}_{\text{somme sur } J_0^c} \leq 4\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0}
\end{aligned}
$$

Lemma 2

$/\lambda$

Then

$$\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c} \quad \leq \quad 3\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0}. \tag{6}$$

$\|\delta\|_{1,J_0} \leq 3\|\delta\|_{1,J_0^c}$

KKT Conditions

Consider (1).Under $\mathcal{H}$ assumption, any Lasso solution $\widehat{\beta}^{L}$ satisfies the following necessary and sufficient condition (KKT)

$$\begin{cases} \frac{1}{n} X_j^\top (Y - X\widehat{\beta}^{\lambda,L}) = \lambda \,\mathrm{sign}(\widehat{\beta}_j^{\lambda,L}) & \text{if} \quad \widehat{\beta}_j^{\lambda,L} \neq 0 \\ \frac{1}{n} |X_j^\top (Y - X\widehat{\beta}^{\lambda,L})| \leq \lambda & \text{if} \quad \widehat{\beta}_j^{\lambda,L} = 0 \end{cases}$$

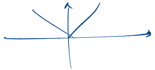
KKT Conditions - Proof

$$\widehat{\beta}^{\lambda,L} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \underbrace{\|Y - X\beta\|_n{}^2 + 2\lambda\|\beta\|_1}_{\text{on cherche la dérivée}} \right\} := \arg\min_{\beta\in\mathbb{R}^p} f(\beta),$$

The lasso solution is not differentiable at any point where β_j is equal to zero. ($\rightarrow$ use the concept of a *subdifferential*).

- The subdifferential of the ℓ_1 penalty at point β is the vector $\cancel{v \in \mathbb{R}^p}$ with components

$$\partial(|\cdot|)(\beta_j) = \begin{cases} \cancel{2}\,\text{sign}(\beta_j), & \text{if} \quad \beta_j \neq 0 \\ \cancel{2}\,\nu_j \in [-1, 1], & \text{if} \quad \beta_j = 0 \end{cases}$$



- We know that the gradient of the first terms at point β is

$$-\frac{2}{n} X^\top (Y - X\beta).$$

Any Lasso solution $\widehat{\beta}^L$ satisfies the following necessary and sufficient condition (KKT)

$$\overrightarrow{0} \in \partial(f)(\widehat{\beta}^L) \Leftrightarrow \text{ KKT Conditions}$$

Theorem~1

[a]Consider model (1). Under $\mathcal{H}$ and RE$(s,3)$ Assumptions, the estimator $\widehat{\beta}^{\lambda,L}$, for $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, with $A > 2\sqrt{2}$ is such that with probability greater than $1 - p^{1-\frac{A^2}{8}}$ ::

$$
\begin{aligned}
\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_2^2 &\leq \frac{16A^2}{\kappa(s,3)^2}\sigma^2 s\log(p). \\
\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 &\leq \frac{16A}{\kappa(s,3)^2}\sigma s\sqrt{\frac{\log(p)}{n}}, \\
M(\widehat{\beta}^{\lambda,L}) &\leq \frac{64\phi_{\max}}{\kappa(s,3)^2},
\end{aligned}
$$

where $\phi_{\max}$ denote the maximal eigenvalue of the Gram matrix ψ_n.

[a]Bickel, Ritov and Tsybakov,*Simultaneous analysis of Lasso and Dantzig selector* (2009, AOS)

Proof of Theorem~1 : the prediction bound

By Lemma~2, for $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, with $A > 2\sqrt{2}$ is s.t. w.h.p.

$$\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n^2 + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq 4\lambda\|\underbrace{\widehat{\beta}^{\lambda,L} - \beta^*}_{=\delta}\|_{1,J_0}$$

$\sum_{j\in J_0} |\delta_j| \leq \sqrt{\sum 1}\sqrt{\sum \delta_j^2}$

By Cauchy Schwarz, and for $M(\beta^*) \leq s$:

$$\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n^2 + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq 4\sqrt{s}\lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{2,J_0} \tag{7}$$

$\kappa \leq \frac{\|X\delta\|_2}{\sqrt{n}\|\delta\|_{2,J_0}}$

$(\Leftrightarrow)\ \kappa\|\delta\|_{2,J_0} \leq \frac{\|X\delta\|_2}{\sqrt{n}}$

By RE(s,3) assumption, for $\delta = \widehat{\beta}^{\lambda,L} - \beta^*$ and $\kappa := \kappa(s, 3)$

$$\kappa\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{2,J_0} \leq \frac{1}{\sqrt{n}}\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_2 = \|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n. \tag{8}$$

Then using (7) and (8)

$$\begin{aligned}
\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n^2 &\leq \|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n^2 + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq 4\sqrt{s}\lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{2,J_0} \\
&\leq \frac{4\sqrt{s}\lambda}{\kappa}\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n.
\end{aligned}$$

By simplifying by $\|X\widehat{\beta}^{\lambda,L} - \beta^*\|_n$ and squaring, we have

$$\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n^2 \leq \frac{16s\lambda^2}{\kappa^2}. \tag{9}$$

For $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, we get (2).

Theorem~1

[a]Consider model (1). Under $\mathcal{H}$ and RE$(s,3)$ Assumptions, the estimator $\widehat{\beta}^{\lambda,L}$, for $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, with $A > 2\sqrt{2}$ is such that with probability greater than $1 - p^{1-\frac{A^2}{8}}$ ::

$$
\begin{aligned}
\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_2^2 &\leq \frac{16A^2}{\kappa(s,3)^2}\sigma^2 s\log(p). \\
\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 &\leq \frac{16A}{\kappa(s,3)^2}\sigma s\sqrt{\frac{\log(p)}{n}}, \\
M(\widehat{\beta}^{\lambda,L}) &\leq \frac{64\phi_{\max}}{\kappa(s,3)^2},
\end{aligned}
$$

where $\phi_{\max}$ denote the maximal eigenvalue of the Gram matrix ψ_n.

[a]Bickel, Ritov and Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector* (2009, AOS)

Proof of Theorem~1: the estimation bound

Under RE(s,3), for $\delta = \widehat{\beta}^{\lambda,L} - \beta^*$ and $\kappa := \kappa(s,3)$,

$$\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{2,J_0} \leq \frac{1}{\kappa}\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n. \tag{10}$$

Using the previous prediction bound (9) $\|X\delta\|_n \leq \frac{4\sqrt{s}\lambda}{\kappa}$

$$\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{2,J_0} \leq \frac{1}{\kappa}\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n \leq \frac{4\sqrt{s}\lambda}{\kappa^2}. \tag{11}$$

As $\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c} \leq 3\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0}$, a consequence of Lemma~2

$$\begin{aligned}
\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 &= \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c} \leq 4\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} \\
&\leq 4\sqrt{s}\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{2,J_0} \quad \text{Cauchy Sch. inequ.} \\
&\leq \frac{16 s\lambda}{\kappa^2} \quad \text{by eq.(11)}
\end{aligned}$$

for $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, we get (3).

Theorem~1

[a]Consider model (1). Under $\mathcal{H}$ and RE$(s,3)$ Assumptions, the estimator $\widehat{\beta}^{\lambda,L}$, for $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, with $A > 2\sqrt{2}$ is such that with probability greater than $1 - p^{1-\frac{A^2}{8}}$ ::

$$
\begin{aligned}
\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_2^2 &\leq \frac{16A^2}{\kappa(s,3)^2}\sigma^2 s\log(p). \\
\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 &\leq \frac{16A}{\kappa(s,3)^2}\sigma s\sqrt{\frac{\log(p)}{n}}, \\
M(\widehat{\beta}^{\lambda,L}) &\leq \frac{64\phi_{\max}}{\kappa(s,3)^2},
\end{aligned}
$$

where $\phi_{\max}$ denote the maximal eigenvalue of the Gram matrix ψ_n.

[a]Bickel, Ritov and Tsybakov,*Simultaneous analysis of Lasso and Dantzig selector* (2009, AOS)

Proof of Theorem~1: the estimation support bound

On Ω, w.h.p with high proba

$$\left|\frac{1}{n}X_j^\top \epsilon\right| \leq \lambda/2, \quad \forall j = 1, \cdots, p \tag{12}$$

Moreover, the Lasso estimator $\widehat{\beta}^{\lambda,L}$ satisfies the KKT condition:

$$\begin{cases} \frac{1}{n}X_j^\top(Y - X\widehat{\beta}^{\lambda,L}) = \lambda \operatorname{sign}(\widehat{\beta}_j^{\lambda,L}) & \text{if} \quad \widehat{\beta}_j^{\lambda,L} \neq 0 \\ \frac{1}{n}|X_j^\top(Y - X\widehat{\beta}^{\lambda,L})| \leq \lambda & \text{if} \quad \widehat{\beta}_j^{\lambda,L} = 0 \end{cases}$$

Then, we have $$\left|\frac{1}{n}X_j^\top(Y - X\widehat{\beta}^{\lambda,L})\right| = \lambda \quad \text{if} \quad \widehat{\beta}_j^{\lambda,L} \neq 0 \tag{13}$$

Combining (12) and (13), and as $X\beta^* = Y - \epsilon$, we have

$$\begin{aligned} \left|\frac{1}{n}X_j^\top(X\beta^* - X\widehat{\beta}^{\lambda,L})\right| &= \left|\frac{1}{n}X_j^\top(Y - X\widehat{\beta}^{\lambda,L} - \epsilon)\right| \geq \left|\frac{1}{n}X_j^\top(Y - X\widehat{\beta}^{\lambda,L})\right| - \left|\frac{1}{n}X_j^\top \epsilon\right| \\ &\geq \lambda - \lambda/2 = \lambda/2 \quad \text{if} \quad \widehat{\beta}_j^{\lambda,L} \neq 0 \end{aligned}$$

Proof of Theorem~1: the estimation support bound

$$\left|\frac{1}{n}X_j^\top(X\beta^* - X\widehat{\beta}^{\lambda,L})\right| \quad \geq \quad \lambda/2 \quad \text{if} \quad \widehat{\beta}_j^{\lambda,L} \neq 0 \qquad (14)$$

Denote by $\widehat{J} = J(\widehat{\beta}^{\lambda,L})$ and $M(\widehat{\beta}^{\lambda,L})$ the cardinal of $\widehat{J}$, then

$$\begin{aligned}\sum_{j=1}^{p}\left(\frac{1}{n}X_j^\top(X\beta^* - X\widehat{\beta}^{\lambda,L})\right)^2 \quad &\geq \quad \sum_{j\in\widehat{J}}\left(\frac{1}{n}X_j^\top(X\beta^* - X\widehat{\beta}^{\lambda,L})\right)^2 \geq \sum_{j\in\widehat{J}}(\lambda/2)^2 \quad \text{by eq.(14)} \\ &= \quad M(\widehat{\beta}^{\lambda,L})\lambda^2/4 \qquad (15)\end{aligned}$$

minoré

First note that $\frac{X^\top X}{n}$ and $\frac{XX^\top}{n}$ have same maximal eigenvalue $\phi_{\max}$. Then,

$$\begin{aligned}\sum_{j=1}^{p}\left(\frac{1}{n}X_j^\top(X\beta^* - X\widehat{\beta}^{\lambda,L})\right)^2 \quad &= \quad \frac{1}{n}\left(X\beta^* - X\widehat{\beta}^{\lambda,L}\right)^\top \frac{XX^\top}{n}\left(X\beta^* - X\widehat{\beta}^{\lambda,L}\right) \\ &\leq \quad \frac{1}{n}\left(X\beta^* - X\widehat{\beta}^{\lambda,L}\right)^\top \phi_{\max}\left(X\beta^* - X\widehat{\beta}^{\lambda,L}\right) \\ &= \quad \phi_{\max}\frac{1}{n}\|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_2^2 \qquad (16)\end{aligned}$$

majoré

Proof of Theorem~1: the estimation support bound

Combining (15) and (16), it comes

$$M(\widehat{\beta}^{\lambda,L}) \leq \frac{4\phi_{\max}}{\lambda^2}\|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 \tag{17}$$

Using prediction bound (9),

$$M(\widehat{\beta}^{\lambda,L}) \leq \frac{4\phi_{\max}}{\lambda^2}\frac{16s\lambda^2}{\kappa^2} = \frac{64\phi_{\max}s}{\kappa^2}. \tag{18}$$

Lemma~2

Consider model (1). Under $\mathcal{H}$ Assuptions, The estimator $\widehat{\beta}^{\lambda,L}$ for $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$, with $A > 2\sqrt{2}$ is such that with probability greater than $1 - p^{1-\frac{A^2}{8}}$:

$$\|X(\widehat{\beta}^{\lambda,L} - \beta^*)\|_n^2 + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \leq 4\lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0}, \qquad (19)$$

where $J_0 = J(\beta^*)$.

By definition of $\widehat{\beta}^{\lambda,L}$, we have for all $\beta \in \mathbb{R}^p$

$$\|Y - X\widehat{\beta}^{\lambda,L}\|_n^2 + 2\lambda\|\widehat{\beta}^{\lambda,L}\|_1 \leq \|Y - X\beta\|_n^2 + 2\lambda\|\beta\|_1$$

As $Y = X\beta^* + \epsilon$, it comes

$$\|X\beta^* - X\widehat{\beta}^{\lambda,L} + \epsilon\|_n^2 + 2\lambda\|\widehat{\beta}^{\lambda,L}\|_1 \leq \|X\beta^* - X\beta + \epsilon\|_n^2 + 2\lambda\|\beta\|_1$$

$$\|\varepsilon\|_n^2 + \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 + \tfrac{2}{n}\langle X\beta^* - X\widehat{\beta}^{\lambda,L}, \epsilon\rangle + 2\lambda\|\widehat{\beta}^{\lambda,L}\|_1 \leq \|X\beta^* - X\beta\|_n^2 + \tfrac{2}{n}\langle X\beta^* - X\beta, \epsilon\rangle + 2\lambda\|\beta\|_1 + \|\varepsilon\|_n^2$$

In particular for $\beta = \beta^*$ and $V_j := \frac{1}{n}\sum_{i=1}^n \epsilon_i x_{i,j}$, we have

$$\begin{aligned}
& \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 + \frac{2}{n}\langle X\beta^* - X\widehat{\beta}^{\lambda,L}, \epsilon\rangle + 2\lambda\|\widehat{\beta}^{\lambda,L}\|_1 \leq 2\lambda\|\beta^*\|_1 \\
\Leftrightarrow \quad & \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 \leq 2\lambda\left(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1\right) + \frac{2}{n}\langle X\widehat{\beta}^{\lambda,L} - X\beta^*, \epsilon\rangle \\
\Leftrightarrow \quad & \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 \leq 2\lambda\left(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1\right) + \frac{2}{n}\sum_{j=1}^p \sum_{i=1}^n \epsilon_i x_{i,j}(\widehat{\beta}_j^{\lambda,L} - \beta_j^*) \\
\Leftrightarrow \quad & \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 \leq 2\lambda\left(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1\right) + 2\sum_{j=1}^p V_j(\widehat{\beta}_j^{\lambda,L} - \beta_j^*)
\end{aligned}$$

Proof of Lemme~2

Then w.h.p $> 1 - p^{1-\frac{A^2}{8}}$, on $\Omega = \bigcap_{j=1}^{p} \left\{2|V_j| \le \lambda\right\}$

$$
\begin{aligned}
& \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 \le 2\lambda\left(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1\right) + 2\sum_{j=1}^{p} V_j(\widehat{\beta}_j^{\lambda,L} - \beta_j^*) \\
\Leftrightarrow \quad & \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 \le 2\lambda\left(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1\right) + \lambda\sum_{j=1}^{p} (\widehat{\beta}_j^{\lambda,L} - \beta_j^*) \\
\Leftrightarrow \quad & \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 \le 2\lambda\left(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1\right) + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \\
\Leftrightarrow \quad & \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \le 2\lambda\left(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1\right) + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \\
\Leftrightarrow \quad & \|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \le 2\lambda(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1 + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_1) \qquad (20)
\end{aligned}
$$

Morover $p = J_0 \cup J_0^c$, where $J_0 = J(\beta^*) = \{j :\ \beta_j^* \neq 0\}$, we have

$$
\left\{
\begin{aligned}
\|\beta^*\|_1 &= \|\beta^*\|_{1,J_0} + \|\beta^*\|_{1,J_0^c} = \|\beta^*\|_{1,J_0} \\
\|\widehat{\beta}^{\lambda,L}\|_1 &= \|\widehat{\beta}^{\lambda,L}\|_{1,J_0} + \|\widehat{\beta}^{\lambda,L}\|_{1,J_0^c} \\
\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 &= \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c}
\end{aligned}
\right.
$$

Proof of Lemme~2

As $\|a\| - \|b\| \leq \|a-b\|$, we have

$$
\begin{aligned}
&\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1 + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 \\
&\quad = \|\beta^*\|_{1,J_0} - \|\widehat{\beta}^{\lambda,L}\|_{1,J_0} - \|\widehat{\beta}^{\lambda,L}\|_{1,J_0^c} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c} \\
&\quad = \left[\|\beta^*\|_{1,J_0} - \|\widehat{\beta}^{\lambda,L}\|_{1,J_0}\right] - \|\widehat{\beta}^{\lambda,L}\|_{1,J_0^c} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c} \\
&\quad \leq \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} - \|\widehat{\beta}^{\lambda,L}\|_{1,J_0^c} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c} \\
&\quad \leq 2\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0^c} - \|\widehat{\beta}^{\lambda,L}\|_{1,J_0^c} \\
&\quad \leq 2\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0} + \|\cancel{\widehat{\beta}^{\lambda,L} -}\, \beta^*\|_{1,J_0^c} \\
&\quad \leq 2\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0}
\end{aligned}
$$

Then w.h.p. $> 1 - p^{1-\frac{A^2}{8}}$, on Ω and eq. (20):

$$
\begin{aligned}
\|X\beta^* - X\widehat{\beta}^{\lambda,L}\|_n^2 + \lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_1 &\leq 2\lambda(\|\beta^*\|_1 - \|\widehat{\beta}^{\lambda,L}\|_1 + \|\widehat{\beta}^{\lambda,L} - \beta^*\|_1) \\
&\leq 4\lambda\|\widehat{\beta}^{\lambda,L} - \beta^*\|_{1,J_0}
\end{aligned}
$$