Statistics of LASSO estimator

Now we look at a setting in which d is large and

$$Y_i = \langle \vartheta^*, x_i \rangle + \varepsilon_i$$

where ϑ^* is a sparse vector, meaning that the number of non zero components is "small" compared to d. In this situation we define the LASSO estimator

$$\hat{\vartheta}_{\ell_1} \in \operatorname*{argmin}_{\vartheta \in \mathbb{R}^d} \frac{1}{2n} \|X\vartheta - Y\|^2 + \lambda \|\vartheta\|_1 .$$

The optimality condition yields

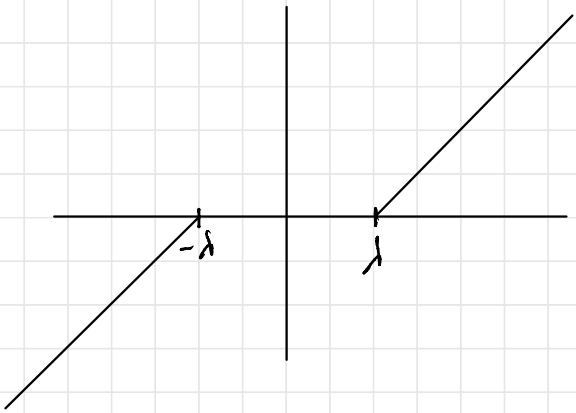
$$0 \in \frac{1}{n} X^T (X\vartheta - Y) + \lambda \partial \|\cdot\|_1 (\vartheta)$$

$$\Leftrightarrow \quad \vartheta - \frac{1}{n} X^T (X\vartheta - Y) \in \vartheta + \lambda \partial \|\cdot\|_1 (\vartheta)$$

$$= (I + \lambda \partial \|\cdot\|_1)(\vartheta)$$

$$\Leftrightarrow \quad \vartheta = \operatorname{prox}_{\lambda \|\cdot\|_1} \left(\vartheta - \frac{1}{n} X^T (X\vartheta - Y) \right)$$

$$\Leftrightarrow \quad \vartheta = \operatorname{soft}_\lambda \left(\vartheta - \frac{1}{n} X^T (X\vartheta - Y) \right)$$



when the component i of $\vartheta - \frac{1}{n} X^T (X\vartheta - y)$ has absolute value less than λ, $\vartheta_i = 0$.

So, the l_1 regularization term promotes sparsity.

Let's start with the statistical analysis.

$$\hat{\vartheta}_{\ell_1} \in \operatorname*{argmin}_{\vartheta \in \mathbb{R}^d} \frac{1}{2n} \|X\vartheta - Y\|^2 + \lambda \|\vartheta\|_1 .$$

$$Y = X\vartheta^* + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma^2 I_n).$$

LEMMA 1

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \leq 2n\lambda \left(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1}\|_1\right) + 2\|\vartheta^* - \hat{\vartheta}_{\ell_1}\|_1 \|X^T\varepsilon\|_\infty$$

Proof:

1. $\|X\hat{\vartheta}_{\ell_1} - Y\|^2 + 2n\lambda\|\hat{\vartheta}_{\ell_1}\|_1 \leq \|X\vartheta^* - Y\|^2 + 2n\lambda\|\vartheta^*\|_1$

2. $\|X\hat{\vartheta}_{\ell_1} - Y\|^2 = \|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 + \|X\vartheta^* - Y\|^2 + 2\langle X(\hat{\vartheta}_{\ell_1} - \vartheta^*), -\varepsilon\rangle$

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \leq \cancel{\|\varepsilon\|^2} + 2n\lambda\left(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1}\|_1\right) - \cancel{\|\varepsilon\|^2} + 2\langle \hat{\vartheta}_{\ell_1} - \vartheta^*, X^T\varepsilon\rangle$$

$$\leq 2n\lambda\left(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1}\|_1\right) + 2\|\hat{\vartheta}_{\ell_1} - \vartheta^*\|_1 \|X^T\varepsilon\|_\infty .$$

Now from LEMMA 1 we derive

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \leq 2n\lambda\left(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1}\|_1\right) + 2\left(\|\vartheta^*\|_1 + \|\hat{\vartheta}_{\ell_1}\|_1\right)\|X^T\varepsilon\|_1$$

$$= 2\|\vartheta^*\|_1\left(\|X^T\varepsilon\|_\infty + n\lambda\right) + 2\|\hat{\vartheta}_{\ell_1}\|_1\left(\|X^T\varepsilon\|_\infty - n\lambda\right)$$

The purpose now is to make

$$\|X^T\varepsilon\|_\infty \le n\lambda \quad \text{in high probability}$$

so to remove the term $2\|\hat{\vartheta}_{\mathcal{E}^c}\|_1(\|X^T\varepsilon\|_\infty - n\lambda)$ in the bound above.

Let's analyze the random vector $X^T\varepsilon$.

$$X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} \in \mathbb{R}^{n\times d} \qquad X^T = [X_1 \dots X_n] \in \mathbb{R}^{d\times n}.$$

$$X = (X_{ij})_{\substack{1\le i\le n \\ 1\le j\le d}}$$

$X_{\cdot,j}$ are the columns of X and $X_{\cdot,j}^T$ are the rows of X^T

$$\|X^T\varepsilon\|_\infty = \max_{1\le j\le d} |\langle X_{\cdot,j}, \varepsilon\rangle|$$

$$\langle X_{\cdot,j}, \varepsilon\rangle = \sum_{i=1}^n X_{ij}\varepsilon_i \quad \text{it is a normal r.v.}$$

with zero mean and variance

$$\mathrm{Var}(\langle X_{\cdot,j}, \varepsilon\rangle) = \sum_{i=1}^n X_{ij}^2 \mathrm{Var}(\varepsilon_i) = \|X_{\cdot,j}\|^2\sigma^2.$$

$$\underline{\langle X_{\cdot,j}, \varepsilon\rangle \sim N(0, \|X_{\cdot,j}\|^2\sigma^2)}$$

Reminder

If $Z \sim N(0, \sigma^2)$

$$\mathbb{P}(|Z| > t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

We look at $\mathbb{P}\left(\{\|X^T \varepsilon\|_\infty > \lambda n\}\right)$

$$\|X^T \varepsilon\|_\infty = \max_{1 \leq j \leq d} |\langle X_{\cdot,j}, \varepsilon \rangle|$$

$$\{\|X^T \varepsilon\|_\infty > \lambda n\} = \bigcup_{j=1}^{d} \{|\langle X_{\cdot,j}, \varepsilon \rangle| > \lambda n\}$$

$$\mathbb{P}\left(\{\|X^T \varepsilon\|_\infty > n\lambda\}\right) \leq \sum_{j=1}^{d} \mathbb{P}\left(|\langle X_{\cdot,j}, \varepsilon \rangle| > n\lambda\right)$$

$$\leq 2 \sum_{j=1}^{d} \exp\left(\frac{-(n\lambda)^2}{2 \|X_{\cdot,j}\|^2 \sigma^2}\right)$$

Hyp: $\|X_{\cdot,j}\| \leq M\sqrt{n}$ $(X_{\cdot,j} \in \mathbb{R}^n)$

$$\leq 2d \exp\left(-\frac{n\lambda^2}{2M^2\sigma^2}\right)$$

$$2d \exp\left(- \frac{n\lambda^2}{2\beta^2\sigma^2} \right) = \delta \iff \frac{n\lambda^2}{2M^2\sigma^2} = \log \frac{2d}{\delta}$$

$$\iff \boxed{\lambda = M\sigma \sqrt{\frac{2\log(2d/\delta)}{n}}}$$

With this choice we have

$$\mathbb{P}\left(\|X^T\varepsilon\|_\infty > \lambda n \right) \leq \delta$$

and hence

$$\mathbb{P}\left(\|X^T\varepsilon\|_\infty \leq \lambda n \right) \geq 1 - \mathbb{P}\left(\|X^T\varepsilon\|_\infty > \lambda n \right) \geq 1 - \delta.$$

3. Therfore with probability $\geq 1-\delta$ we have

$$\|X^T\varepsilon\|_\infty \leq \lambda n$$

and hence

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \leq \|\vartheta^*\|_1 \, 4n\lambda .$$

In the end we have

$$\boxed{MSE\left[X\hat{\vartheta}_{\ell_1}\right] \leq \|\vartheta^*\|_1 \, 4M\sigma \sqrt{\frac{2\log(2d/\delta)}{n}}}$$

with probability greater than $1-\delta$.

In the regime of $d \geq n$, say $n = d \cdot q$ with $q < 1$ we have

$$MSE[X\hat{\vartheta}_{\ell_1}] \leq \|\vartheta^*\|_1 \, 4M\sigma \sqrt{\frac{2\log 2d/\delta}{q \cdot d}}$$

Therefore if $\|\vartheta^*\|_1 = O(d^{\alpha} \cdot \log^{\gamma} d)$ with $\alpha < 1/2$ and $\gamma > 0$

($\vartheta^* \in \mathbb{R}^d$)

we have that

$$MSE[X\hat{\vartheta}_{\ell_1}] \to 0 \quad \text{as} \quad d \to +\infty \quad \text{in high probability.}$$

For instance if $\|\vartheta^*\|_\infty \leq B$, then $\|\vartheta^*\|_1 \leq B \underbrace{\|\vartheta^*\|_0}_{k}$

$k = \#\,\text{supp}\,\vartheta^*$

and hence we have

$$MSE[X\hat{\vartheta}_{\ell_1}] \leq 4MBk\sigma \sqrt{\frac{2\log(2d/\delta)}{n}}$$

Fast rates for Lasso estimator

Def X has incoherence $\kappa \in \mathbb{N}^*$ iff $\left\| \frac{X^T X}{n} - I_d \right\|_\infty \leq \frac{1}{32\kappa}$

Def $\theta \in \mathbb{R}^d$, $S \subset \{1, \dots, d\}$

$$\theta_S \in \mathbb{R}^d \quad (\theta_S)_j = \begin{cases} 0 & \text{if } j \notin S \\ \theta_j & \text{if } j \in S. \end{cases}$$

Lemma 2 Suppose that $\kappa \leq d$, $S \subset \{1, \dots, d\}$ $|S| \leq \kappa$. and that $\theta \in \mathbb{R}^d$ is s.t. $\|\theta_{S^c}\|_1 \leq 3 \|\theta_S\|_1$

Then $\|\theta\|_2^2 \leq \frac{2}{n} \|X\theta\|_2^2$

Proof: $\theta = \theta_S + \theta_{S^c}$

$$\frac{\|X\theta\|_2^2}{n} = \frac{\|X\theta_S\|^2}{n} + \frac{\|X\theta_{S^c}\|}{n} + \frac{2\langle X\theta_S, X\theta_{S^c} \rangle}{n}$$

$$\frac{\|X\theta_S\|^2}{n} = \left\langle \frac{X^T X}{n} \theta_S, \theta_S \right\rangle = \left\langle \left(\frac{X^T X}{n} - I_d \right) \theta_S, \theta_S \right\rangle + \|\theta_S\|^2$$

$$\left| \left\langle \left(\frac{X^T X}{n} - I_d \right) \theta_S, \theta_S \right\rangle \right| \leq \left\| \frac{X^T X}{n} - I_d \right\|_\infty \|\theta_S\|_1^2$$

Therefore 1. $\frac{\|X\theta_S\|^2}{n} \geq \|\theta_S\|^2 - \frac{\|\theta_S\|_1^2}{32\kappa}$

Similarly

2. $$\frac{\|X\theta_{S^c}\|^2}{n} \geq \|\theta_{S^c}\|^2 - \frac{\|\theta_{S^c}\|_1^2}{32k} \geq \|\theta_{S^c}\|^2 - \frac{9\|\theta_S\|_1^2}{32}$$

3. $$2\left|\left\langle \frac{X^TX}{n}\theta_S, \theta_{S^c}\right\rangle\right| = 2\left\langle \left(\frac{X^TX}{n} - I_d\right)\theta_S, \theta_{S^c}\right\rangle + 2\underbrace{\langle \theta_S, \theta_{S^c}\rangle}_{=0}$$

$$\leq \frac{2}{32k}\|\theta_S\|_1\|\theta_{S^c}\|_1$$

$$\leq \frac{6}{32k}\|\theta_S\|_1^2$$

In the end we have

$$\frac{\|X\theta\|^2}{n} \geq \|\theta_S\|^2 + \|\theta_{S^c}\|^2 - \frac{\|\theta_S\|_1^2}{32k} - \frac{9\|\theta_S\|_1^2}{32k} - \frac{6\|\theta_S\|_1^2}{32k}$$

$$= \|\theta_S\|^2 + \|\theta_{S^c}\|^2 - \frac{16}{32k}\|\theta_S\|_1^2$$

$$\|\theta_S\|_1 \leq \sqrt{|S|}\,\|\theta_S\|_2 \leq \sqrt{k}\,\|\theta_S\|_2$$

$$\frac{\|X\theta\|^2}{n} \geq \|\theta_S\|^2 + \|\theta_{S^c}\|^2 - \frac{16k}{32k}\|\theta_S\|_2^2$$

$$= \|\theta_S\|^2 + \|\theta_{S^c}\|^2 - \frac{\|\theta_S\|_2^2}{2}$$

$$\geq \frac{\|\theta_S\|^2 + \|\theta_{S^c}\|^2}{2} = \frac{\|\theta\|^2}{2}. \qquad \square$$

Theorem

Suppose that X has incoherence $\kappa \in \mathbb{N}^*$ and that $|\operatorname{supp} \vartheta^*| \leq \kappa$. Set $\lambda = 4\sigma \sqrt{\frac{\log(2d/\delta)}{n}}$

Then with probability $\geq 1-\delta$ we have

$$\boxed{\text{MSE}[X\hat{\vartheta}_{\ell_1}] \leq 9 \cdot 2^5 \kappa \sigma^2 \frac{\log(2d/\delta)}{n}}$$

Proof: It follows from LEMMA 1 that

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \leq 2n\lambda(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1}\|_1) + 2\|\hat{\vartheta}_{\ell_1} - \vartheta^*\|_1 \|X^T\varepsilon\|_\infty$$

Now we make $\|X^T\varepsilon\|_\infty \leq \frac{n\lambda}{2}$ with high probability

$$\mathbb{P}(\|X^T\varepsilon\|_\infty \geq t) \leq \sum_{i=1}^{d} \mathbb{P}(|\langle X_{\cdot,i}, \varepsilon\rangle| \geq t)$$

$$\leq 2\sum_{i=1}^{d} \exp\left(-\frac{t^2}{2\|X_{\cdot,i}\|^2\sigma^2}\right)$$

From the incoherence we have

$$\left(\frac{X^TX}{n} - I_d\right)_{ij} = \frac{\langle X_{\cdot,i}, X_{\cdot,j}\rangle}{n} - \delta_{ij}$$

and hence $\left| \frac{\|X_{\cdot,j}\|^2}{n} - 1 \right| \leq \frac{1}{32k}$

which implies that $\frac{\|X_{\cdot,j}\|^2}{n} \leq 1 + \frac{1}{32k} \leq 2.$

and hence $\|X_{\cdot,j}\|^2 \leq 2n \quad (\|X_{\cdot,j}\| \leq \sqrt{2n})$

Therefore $\mathbb{P}(\|X^T\varepsilon\|_\infty > t) \leq 2d \exp\left(-\frac{t^2}{4n\sigma^2}\right)$

if we pick $t = \frac{\lambda n}{2}$ we have

$$\mathbb{P}\left(\|X^T\varepsilon\|_\infty > \frac{\lambda n}{2}\right) \leq 2d \exp\left(-\frac{\lambda^2 n}{16\sigma^2}\right)$$

$$2d \exp\left(-\frac{\lambda^2 n}{16\sigma^2}\right) = \delta \iff \frac{\lambda^2 n}{16\sigma^2} = \log\frac{2d}{\delta}$$

$$\iff \boxed{\lambda = 4\sigma\sqrt{\frac{\log(2d/\delta)}{n}}}$$

Thus, with that choice we have

$$\mathbb{P}\left(\|X^T\varepsilon\|_\infty > \frac{\lambda n}{2}\right) \leq \delta \quad \text{and hence}$$

$$\mathbb{P}\left(\|X^T\varepsilon\|_\infty \le \frac{d\eta}{2}\right) \ge 1-\delta.$$

Therefore with p. $\ge 1-\delta$ we have

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \le 2n\lambda\left(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1}\|_1\right) + n\lambda\|\hat{\vartheta}_{\ell_1} - \vartheta^*\|_1$$

Now set $S = \operatorname{supp}\vartheta^*$. Then we have

$$(\hat{\vartheta}_{\ell_1} - \vartheta^*)_{S^c} = \hat{\vartheta}_{\ell_1, S^c}, \text{ therefore}$$

$$\begin{aligned}\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 &\le -n\lambda\|\hat{\vartheta}_{\ell_1,S^c}\|_1 + 2n\lambda\left(\|\vartheta^*\|_1 - \|\hat{\vartheta}_{\ell_1,S}\|_1\right) \\ &\quad + n\lambda\|\hat{\vartheta}_{\ell_1,S} - \vartheta^*\|_1 \\ &\le -n\lambda\|\hat{\vartheta}_{\ell_1,S^c}\|_1 + 3n\lambda\|\hat{\vartheta}_{\ell_1,S} - \vartheta^*\|_1.\end{aligned}$$

If we define $\vartheta = \hat{\vartheta}_{\ell_1} - \vartheta^*$ we have.

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \le -n\lambda\|\vartheta_{S^c}\|_1 + 3n\lambda\|\vartheta_S\|_1.$$

and in particular: $\|\vartheta_{S^c}\|_1 \le 3\|\vartheta_S\|_1$

Therefore, by Lemma 2,

$$\|\vartheta_S\|_1 \le \sqrt{|S|}\,\|\vartheta_S\|_2 \le \sqrt{K}\,\|\vartheta\|_2 \le \sqrt{\frac{2n}{n}}\,\|X\vartheta\|_2.$$

In the end. we have with $p. \geq 1-\delta$

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|_2^2 \leq 3n\lambda\|\vartheta_S\|_1 \leq 3n\lambda\sqrt{\frac{2k}{n}}\|X(\hat{\vartheta}_{\ell_1} - \vartheta^*)\|$$

and hence

$$\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \leq 18\, n\lambda^2 k$$

that is

$$\frac{1}{n}\|X\hat{\vartheta}_{\ell_1} - X\vartheta^*\|^2 \leq 18 \cdot k \cdot 16\sigma^2 \frac{\log(2d/\delta)}{n}. \qquad \square$$