High-dimensional Statistics — 6th Lecture

A. DALALYAN

MASTER MVA

0. RECALL

We observe $(Y_i)_{i=1,\dots,n}$ such that

(1) $$y = f^* + \xi \quad ; \quad \xi \sim \mathcal{N}(0, \sigma^2 I_n)$$

where $f^* = (f_1^*, \dots, f_n^*) \in \mathbb{R}^n$ is an unknown vector (called signal)

$\xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$ is a random noise

Dictionary: $f^* = \Phi \cdot \beta^*$ where $\Phi \in \mathbb{R}^{n\times p}$ is a known design matrix

Sparsity: the dimension p of β^* is large, possibly larger than n, but many coordinates of β^* vanish:

$$s = \|\beta^*\|_0 = \sum_{j=1}^p \mathbb{1}(|\beta_j^*| \neq 0) \ll p.$$

Lasso: $$\boxed{\hat\beta^{\text{Lasso}} \in \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|y - \Phi\beta\|_2^2 + C\|\beta\|_1 \right\}}$$

THEOREM: If Φ is orthogonal, i.e., $\frac{1}{n}\Phi^T\Phi = I_p$, then $\hat\beta^{\text{Lasso}}$ coincides with the soft thresholding procedure.

More precisely, for $\lambda = \frac{C}{2n}$, we have

$$\hat\beta_j^{\text{Lasso}} = \left(\left|\tfrac{1}{n}(\Phi^T y)_j\right| - \lambda\right)_+ \cdot \operatorname{sign}\left((\Phi^T y)_j\right), \quad j = 1, \dots, p.$$

1. Remarks

1) From now on, we will parameterize the Lasso by λ instead of C:

(2) $$\hat\beta_\lambda^{\text{Lasso}} \in \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|y - \Phi\beta\|_2^2 + \lambda\|\beta\|_1 \right\}.$$

2) The condition $\frac{1}{n}\Phi^T\Phi = I_p$ essentially means that the dictionary

3) Even if Φ is not ortho-normal (ON), the Lasso is efficiently computable even for very large values of p.

<u>Lemma</u>: $\hat{\beta}$ is a solution to (2) if and only if there exist $(\hat{u}, \hat{v}, \hat{t}) \in \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}$ such that $(\hat{\beta}, \hat{u}, \hat{v}, \hat{t})$ is a solution of the second-order cone program (SOCP):

$$\left.\begin{array}{l} \min \left\{ \frac{1}{2n} \times t + \lambda \sum_{j=1}^{p} u_j \right\} \\ \text{subject to :} \quad -u_j \leqslant \beta_j \leqslant u_j \quad \forall\, j = 1, \dots, p \\ \qquad\qquad\qquad y - \Phi\beta = v \\ \qquad\qquad\qquad \|(v; t; \tfrac{1}{2})\|_2 \leqslant t + \tfrac{1}{2} \end{array}\right\} \text{(SOCP1)}$$

<u>Proof.</u>

1) We prove first that if $(\hat{\beta}, \hat{u}, \hat{v}, \hat{t})$ is a solution of (SOCP 1) then $\hat{\beta}$ is a solution of (2). Indeed, let β be any vector of $\mathbb{R}^p$. Define $u_j = |\beta_j|$, $j = 1, \dots, p$; $v = y - \Phi\beta$ and $t = \|y - \phi\beta\|_2^2$. One easily checks that all the constraints of (SOCP1) are fulfilled for (β, u, v, t) defined is such a way. Therefore, (β, u, v, t) is a feasible solution and

$$\text{(3)} \qquad \frac{1}{2n} \times \hat{t} + \lambda \sum_{j=1}^{p} \hat{u}_j \leqslant \frac{1}{2n} \times t + \lambda \sum_{j=1}^{p} u_j = \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 .$$

On the other hand, since $(\hat{\beta}, \hat{u}, \hat{v}, \hat{t})$ is a solution of (SOCP1) we have [notice that the third constraint is equivalent to $\|v\|_2^2 \leqslant t$] : $\|y - \phi\hat{\beta}\|_2^2 \leqslant t^2$ and $\|\hat{\beta}\|_1 \leqslant \|\hat{u}\|_1$.

Combined with (3), this implies that

$$\frac{1}{2n} \|y - \phi\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \leqslant \frac{1}{2n} \|y - \phi(\beta)\|_2^2 + \lambda \|\beta\|_1 .$$

Since this holds true for every $\beta \in \mathbb{R}^p$, $\hat{\beta}$ is a solution of (2).

2) Let $\hat\beta$ be a solution of (2) and define

$$\hat u_j = |\hat\beta_j| \ , j = 1, \dots, p \ ; \quad \hat v = y - X\hat\beta \quad \text{and} \quad \hat t = \|\hat v\|_2^2 .$$

One easily checks that $(\hat\beta, \hat u, \hat v, \hat t)$ verifies the constraints of (SOCP 1). Furthermore, if (β, u, v, t) is another vector satisfying the constraints of (SOCP 1), we have

$$(4) \qquad \frac{1}{2n}\|y - \Phi\hat\beta\|_2^2 + \lambda\|\hat\beta\|_1 \leqslant \frac{1}{2n}\|y - \Phi\beta\|_2^2 + \lambda\|\beta\|_1$$

The left-hand side of (4) is equal to

$$\frac{1}{2n}\hat t + \lambda\sum_{j=1}^{p}\hat u_j$$

whereas the right-hand side of (4) is bounded as follows:

$$\frac{1}{2n}\|y - \Phi\beta\|_2^2 + \lambda\|\beta\|_1 \leqslant \frac{1}{2n}\|v\|_2^2 + \lambda\sum_{j=1}^{p} u_j \leqslant \frac{1}{2n} t + \lambda\sum_{j=1}^{p} u_j .$$

This implies that

$$\frac{1}{2n}\hat t + \lambda\sum_{j=1}^{p}\hat u_j \leqslant \frac{1}{2n} t + \lambda\sum_{j=1}^{p} u_j$$

for every (β, u, v, t) satisfying the constraints. $\blacksquare$

3. Prediction loss of the Lasso: the "slow" rates.

$$y = \Phi\beta^* + \xi \qquad ; \qquad \hat\beta^{\text{Lasso}} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n}\|y - \Phi\beta\|_2^2 + \lambda\|\beta\|_1 \right\}$$

We wish to evaluate the prediction loss

$$\ell_n(\hat f, f^*) = \frac{1}{n}\|\hat f - f^*\|_2^2 = \frac{1}{n}\|\Phi(\hat\beta - \beta^*)\|_2^2 \triangleq \ell_n(\hat\beta, \beta^*) .$$

Note: we divide by n since $\|\hat f - f^*\|_2^2$ is a sum containing n terms.

Condition C1: For $\forall j \in \{1, \dots, p\}$, we have $\|\Phi^j\|_2^2 = \sum_{i=1}^{n} \varphi_{ij}^2 \leqslant n$.

Lemma For every $\delta \in (0,1)$, the event $\mathcal{B} = \{\|\Phi^T \xi\|_\infty \leq \sigma \sqrt{2n \ln(p/\delta)}\}$ satisfies $\mathbb{P}(\mathcal{B}) \geq 1-\delta$.

Proof. We will prove that $\mathbb{P}(\mathcal{B}^c) \leq \delta$. Set $x = \sqrt{2 \ln(p/\delta)}$ and $\xi_j = (\Phi^j)^T \xi / (\sigma \cdot \|\Phi^j\|_2)$; $j = 1, \ldots, p$. Since Gaussian distribution is stable by affine transformations, we have $\xi_j \sim \mathcal{N}(0,1)$ $\forall j$.

Therefore,

$$
\begin{aligned}
\mathbb{P}(\mathcal{B}^c) &= \mathbb{P}\Big(\max_j |(\Phi^j)^T \xi| > \sigma \sqrt{2n \ln(p/\delta)} \Big) \\
&= \mathbb{P}\Big(\bigcup_{j=1}^{p} \{ |(\Phi^j)^T \xi| > \sigma \sqrt{n} \cdot x \} \Big) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\Big(|(\Phi^j)^T \xi| > \sigma \sqrt{n}\, x \Big) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\Big(|(\Phi^j)^T \xi| > \sigma \|\Phi^j\|_2 \cdot x \Big) \\
&= \sum_{j=1}^{p} \mathbb{P}\big(|\xi_j| > x \big) = 2p\, \mathbb{P}(\xi_1 > x)
\end{aligned}
$$

where for the last equality we used the symmetry of the Gaussian distribution. To complete the proof we will use the Gaussian tail bound : $\mathbb{P}(\xi_1 > x) \leq \frac{1}{2} \exp(-x^2/2)$, $\forall x > 0$.

This leads to

$$\mathbb{P}(\mathcal{B}^c) \leq 2p \times \frac{1}{p} \times \exp\Big(-\frac{1}{2}\big(\sqrt{2 \ln(p/\delta)}\big)^2\Big) = \delta . \quad ▨$$

For completeness, we provide the proof of the Gaussian tail bound.

$$\mathbb{P}(\xi_1 > x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du .$$

Set $G(x) = \mathbb{P}(\xi_1 > x) - \frac{1}{2} e^{-x^2/2}$. One easily checks that

$$G'(x) = \frac{-1}{\sqrt{2\pi}} e^{-x^2/2} + \frac{x e^{-x^2/2}}{2} = \frac{e^{-x^2/2}}{2}\left(x - \sqrt{\frac{2}{\pi}}\right).$$

This implies that G is decreasing on $]0, \sqrt{2/\pi}[$ and increasing on $]\sqrt{2/\pi}, +\infty[$. Therefore, $G(x) \leq \max(G(0); G(+\infty)) = 0$ ▨

From now on, we will always work on the event $\mathcal{B}$.

<u>Theorem</u>: If Φ satisfies condition C1, then on $\mathcal{B}$, it holds that

(5) $$\ell_n(\hat\beta^L, \beta^*) \leqslant \inf_{\beta \in \mathbb{R}^p} \left\{ \ell_n(\beta, \beta^*) + 4\lambda \|\beta\|_1 \right\}$$

provided that λ is chosen $\geqslant \sigma \sqrt{\frac{2}{n} \ln(p/\delta)}$.

<u>Proof</u>. Let $\beta \in \mathbb{R}^p$ be any vector. Then

$$\frac{1}{2n} \|y - \Phi\hat\beta\|_2^2 + \lambda \|\hat\beta\|_1 \leqslant \frac{1}{2n} \|y - \Phi\beta\|_2^2 + \lambda \|\beta\|_1 .$$

$$\frac{1}{2n} \|\Phi(\beta^* - \hat\beta) + \xi\|_2^2 + \lambda \|\hat\beta\|_1 \leqslant \frac{1}{2n} \|\Phi(\beta^* - \beta) + \xi\|_2^2 + \lambda \|\beta\|_1$$

$$\frac{1}{n} \|\Phi(\beta^* - \hat\beta)\|_2^2 + 2\lambda \|\hat\beta\|_1 \leqslant \frac{2}{n} \xi^T \Phi(\hat\beta - \beta) + 2\lambda \|\beta\|_1 + \frac{1}{n} \|\Phi(\beta^* - \beta)\|_2^2$$

$$\ell_n(\hat\beta, \beta^*) \leqslant \ell_n(\beta, \beta^*) + 4\lambda \|\beta\|_1 + \underbrace{\frac{2}{n} \xi^T \Phi(\hat\beta - \beta) - 2\lambda\left(\|\beta\|_1 + \|\hat\beta\|_1\right)}_{\leqslant 0 \text{ on } \mathcal{B}}$$

En effet

$$\frac{2}{n} \xi^T \Phi(\hat\beta - \beta) \leqslant \frac{2}{n} (\Phi^T \xi)^T (\hat\beta - \beta) \leqslant \frac{2}{n} \|\Phi^T \xi\|_\infty \cdot \|\hat\beta - \beta\|_1$$

$$\leqslant \frac{2}{n} \sigma \sqrt{2n \ln(p/\delta)} \times \|\hat\beta - \beta\|_1$$

$$\leqslant 2\lambda \|\hat\beta - \beta\|_1 \leqslant 2\lambda \left(\|\hat\beta\|_1 + \|\beta\|_1\right)$$ ▨

<u>Commentaires</u>

① On dit que (5) est une Oracle Inequality with "slow" rate. If an oracle tells us what is the best sparse approximation $\bar\beta$ of β^*, then can use it as estimator of β^* and get the loss $\ell_n(\bar\beta, \beta^*)$. Since we do not know $\bar\beta$, we use $\hat\beta^{\text{Lasso}}$ and get the loss bounded by $\ell_n(\bar\beta, \beta^*) + \text{Const} \times \underbrace{\sqrt{\frac{\ln(p)}{n}}}_{\text{this is called "slow" rate.}} \times \|\bar\beta\|_1$