Statistical Estimates for LASSO

2022年2月8日 9:14

Recall the Setup:

- $y = X\theta^* + w$, $X \in \mathbb{R}^{n\times d}$, $n \ll d$; $\underbrace{\mathcal{M}}_{\text{model}} \subseteq \bar{\mathcal{M}} \subseteq \mathbb{R}^d$ $\quad \bar{\mathcal{M}}^{\perp}$: perturbation

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \{ \mathcal{L}(\theta) + \lambda R(\theta) \}$$

Here, $\mathcal{L}$ is a smooth convex loss, R is a decomposable norm wrt $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$, $\lambda \geq 2R^*(\nabla\mathcal{L}(\theta^*))$.

- Localization of error vector

$$\hat{\Delta} = \hat{\theta} - \theta^* \in \mathcal{C} = \{ \Delta \in \mathbb{R}^d : R(\Delta_{\bar{\mathcal{M}}^{\perp}}) \leq 3R(\Delta_{\bar{\mathcal{M}}}) + 4R(\theta^*_{\mathcal{M}^{\perp}}) \}$$

- RSC of $\mathcal{L}$ over $\mathcal{C}$

$$\mathcal{L}(\theta^* + \Delta) \geq \mathcal{L}(\theta^*) + \nabla\mathcal{L}(\theta^*)^T\Delta + \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) \quad \forall \Delta \in \mathcal{C}$$

Theorem: Under the above setup,

$$\underbrace{\|\hat{\theta} - \theta^*\|_2^2}_{\hat{\Delta}} \leq \underbrace{\frac{9\lambda^2}{4\kappa^2}\psi^2(\bar{\mathcal{M}})}_{\text{tightness measure}} + \underbrace{\frac{2}{\kappa}\left[\tau^2(\theta^*) + 2\lambda R(\theta^*_{\mathcal{M}^{\perp}})\right]}_{\text{model specification error}},$$

where

$$\psi(\mathcal{M}) = \sup_{u \in \mathcal{M}\setminus\{0\}} \frac{R(u)}{\|u\|_2}$$

Example: LASSO with exactly sparse model

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \{ \underbrace{\tfrac{1}{2n}\|y - X\theta\|_2^2}_{\mathcal{L}(\theta)} + \lambda \underbrace{\|\theta\|_1}_{R(\theta)} \} \quad \text{— (L)}$$

By taking S to be the support of θ^* with $|S| = s$, $\mathcal{C}$ takes the form

$S = \operatorname{supp}(\theta^*) = \{ i : \theta^*_i \neq 0 \}$

$\mathcal{M}(S) = \{ \theta : \theta_j = 0 \ \forall j \notin S \} = \bar{\mathcal{M}}(S)$

$$\mathcal{C} = \{ \Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 \}$$

Want:

① $\mathcal{L}$ is RSC over $\mathcal{C}$:

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa \|\Delta\|_2^2 \quad \forall \Delta \in \mathcal{C}. \qquad \text{--- } (*)$$

② Estimate $R^*(\nabla \mathcal{L}(\theta^*))$: $\quad R^*(\theta) = \|\theta\|_\infty$

$$\therefore \quad R^*(\nabla \mathcal{L}(\theta^*)) = \frac{1}{n}\|X^T w\|_\infty$$

$\hookrightarrow w = y - X\theta^*$

Proposition: Suppose that

(i) X satisfies $(*)$; [RSC]

(ii) $\frac{\|X_j\|_2}{\sqrt{n}} \leq 1 \quad \forall j$, where X_j is the j^{th} col of X, [scaling]

(iii) w is sub-Gaussian with parameter $\sigma > 0$: w has zero mean and for any fixed $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$,

$$\Pr\left[\,|v^T w| \geq t\,\right] \leq 2\exp\left(-t^2/2\sigma^2\right) \quad \forall t > 0 \quad \left[\begin{array}{c}\text{statistical}\\ \text{model of}\\ \text{noise}\end{array}\right]$$

(iv) θ^* is s-sparse. [sparsity]

Then, by taking $\lambda = 4\sigma\sqrt{\frac{\log d}{n}}$ in (L), then

$$\|\hat{\theta} - \theta^*\|_2^2 \leq 36\,\frac{\sigma^2}{\kappa^2}\cdot\frac{s\log d}{n}$$

with probability $\geq 1 - 2/d$.

Proof: In the theorem, $\tau(\cdot) = 0$, $R(\theta^*_{\mathcal{M}^\perp}) = 0$. It remains to estimate $\Psi(\mathcal{M})$ and λ

1°: $$\Psi(\mathcal{M}) = \sup_{u \in \mathcal{M}\setminus\{0\}} \frac{\|u\|_1}{\|u\|_2} = \sqrt{s} \qquad \left[\forall u \in \mathbb{R}^d: \|u\|_1 \leq \sqrt{d}\cdot\|u\|_2\right]$$

2°: From our calculation, we need to estimate $\frac{1}{n}\|X^T w\|_\infty$

Observe: The j^{th} entry of $X^T w$ is $X_j^T w$. Now,

$$\Pr\left[\,\left|\frac{X_j^T w}{n}\right| \geq t\,\right] = \Pr\left[\,\left|\left(\frac{X_j^T}{\sqrt{n}}\right) w\right| \geq \sqrt{n}\, t\,\right]$$

$$\leq 2\exp\left(-\frac{nt^2}{2\sigma^2}\right) \qquad \forall t > 0,$$

$$\le \quad 2\exp\left(-\frac{nt^2}{2\sigma^2}\right) \qquad \forall t>0,\ j=1,\cdots,d$$

By the Union bound,

$$\Pr\left[\frac{1}{n}\|X^T w\|_\infty \ge t\right] \le 2d\exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

Set $t^2 = \dfrac{4\sigma^2 \log d}{n}$, Then, the RHS $= \dfrac{2}{d}$. Hence, with probability $1-\dfrac{2}{d}$, we have $\lambda = 4\sigma\sqrt{\dfrac{\log d}{n}} \ge \dfrac{2}{n}\|X^T w\|_\infty$

Establishing the RSC of $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$ over

$\mathcal{C} = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1\}$; i.e.,

$$\frac{\|X\Delta\|_2^2}{2n} \ge \kappa\|\Delta\|_2^2 \qquad \forall \Delta \in \mathcal{C}. \qquad \text{— } (*)$$

<u>Note</u>: one would not expect $(*)$ to hold for arbitrary X:

e.g. fix $\Delta \in \mathcal{C}$ and construct X s.t. $\Delta \in \text{null}(X)$.

$\rightarrow$ need, e.g., a statistical model of X in order to proceed.

<u>Theorem</u>: Suppose that the rows of X are iid $\mathcal{N}(0, I_d)$.

Then, with probability $\ge 1 - c_1\exp(-c_2 n)$,

$$\begin{array}{cc} & d \\ n & \begin{bmatrix} - \; x_j^T \; - \end{bmatrix} \\ & X \end{array}$$

where $c_1, c_2 > 0$ are constants,

$$(**) \qquad \frac{\|X\Delta\|_2}{\sqrt{n}} \ge \frac{1}{4}\|\Delta\|_2 - 9\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \qquad \forall \Delta \in \mathbb{R}^d$$

Implication: Note that for $\Delta \in \mathcal{C}$,

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \underset{\substack{\uparrow \\ \Delta\in\mathcal{C}}}{\le} 4\|\Delta_S\|_1 \le 4\sqrt{s}\|\Delta_S\|_2 \le 4\sqrt{s}\|\Delta\|_2$$

Hence, $(**)$ implies

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \left(\frac{1}{4} - 36\sqrt{\frac{s\log d}{n}}\right)\|\Delta\|_2 \quad \forall \Delta \in \mathcal{C}$$

Thus, as long as

$$\frac{1}{4} - 36\sqrt{\frac{s\log d}{n}} > 0 \iff n > 144^2 \, s \log d,$$

the bound is meaningful.

The proof of the theorem has 3 main steps:

Step 1: Observe that $(**)$ is trivial if $\Delta = 0$. So we may assume $\|\Delta\|_2 = 1$ by homogeneity. For any $r > 0$, define

$$V(r) = \{ \Delta \in \mathbb{R}^d : \|\Delta\|_2 = 1, \ \|\Delta\|_1 \leq r \}$$

Proposition 1:

$$\mathbb{E}\left[\inf_{\Delta \in V(r)} \frac{\|X\Delta\|_2}{\sqrt{n}}\right] \geq 3\left[\frac{1}{4} - \sqrt{\frac{\log d}{n}}\, r\right]$$

whenever $V(r) \neq \emptyset$.