ROBUST ESTIMATION IN LINEAR MODEL

A. DALALYAN (joint with Ph. Thompson)

28/03/2019 Les Houches

Plan

1) Problem formulation
2) Main results
3) Sketch of the proof

1 PROBLEM

We observe n input-output pairs $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$

$$y_i = x_i^T \cdot \beta^* + \underbrace{\sqrt{n}\,\theta_i^*}_{\text{Contamination}} + \underbrace{\xi_i}_{\overset{iid}{\sim} \mathcal{N}(0,\sigma^2) \text{ noise}} \qquad i = 1, \dots, n$$

Our goal is to reconstruct the vector β^*.

Matrix notation : $\boxed{Y = X\beta^* + \sqrt{n}\,\theta^* + \xi}$

- $\beta^* \in \mathbb{R}^p$ is s-sparse with $s \ll p \wedge n$
- $\theta^* \in \mathbb{R}^n$ is k-sparse with $k \ll n$
- $\xi \sim \mathcal{N}_n(0, \sigma^2 I_n)$
- $x_i \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$ with $\Sigma_{ii} = 1 \ \forall i \in [p]$.

Remark The vector θ^* may be arbitrary (adversarial contamination). It is random and depends on X and ξ.

Lasso or Huber-ℓ_1

A natural estimator is the ℓ_1-penalized LSE.

$$\boxed{\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2n} \|Y - X\beta - \sqrt{n}\,\theta\|_2^2 + \lambda\left(\|\beta\|_1 + \|\theta\|_1\right) \right\}}$$

Candes & Randall (2008) ; Nguyen & Tran (2012)

[$s = p << n$]

stylized communication

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \leq \frac{C \cdot \sigma^2}{\mu_{\min}(\Sigma)} \left(\frac{s \log p}{n} + \underbrace{\frac{k \log n}{n}}_{\text{suboptimal}} \right) \quad \text{w.h.p.}$$

Question: Can this rate be improved?

Minimax risk: (Chen et al. 2016) $\exists\, c_1, c_2, c_3 > 0$ s.t. $\forall\ k \leq c_1 \cdot n$

$$\inf_{\bar{\beta}} \sup_{\beta^*, \theta^*} \mathbb{E}\left[\|\Sigma^{1/2}(\bar{\beta} - \beta^*)\|_2^2 \right] \begin{array}{l} \leq c_2 \sigma^2 \left(\boxed{\frac{s \log(2p/s)}{n} + \frac{k^2}{n^2}} \right) \\ \geq c_3 \sigma^2 \left(\quad \text{— " —} \quad \right) \end{array}$$

- sup over β^* such that $\|\beta^*\|_0 \leq s$ & θ^* s.t. $\|\theta^*\|_0 \leq k$
- inf over all possible estimators of β^*.
- estimator in (Chen et al 2016) is not computable in poly-time.

② Main Result

Our main result reads as follows.

THEOREM Let $\hat{\Sigma}$ satisfy RE condition with $\varkappa$.

For some universal constants $c_1, c_2, c_3 > 0$ the following claim holds true. If

(1) $$c_1(s+k)\log(pn) \leqslant n \quad \& \quad c_2 \log(1/\delta) \leqslant n$$

then the choice $\lambda \geqslant 3\sigma\sqrt{\frac{\log(np/\delta)}{n}}$

ensures that

(2) $$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2 \leqslant c_3 \underbrace{\frac{\lambda^2}{n}\left(\frac{s^2}{\varkappa}\right.}_{T_1} + k^2\Big)\log\left(\frac{pn}{\delta}\right) + 9\underbrace{\frac{\lambda^2 s}{\varkappa}}_{T_3}$$

with prob. $\geqslant 1-\delta$.

Remarks 1) T_1 is always of smaller order than T_3. because of condition (1)

2) If we choose $\lambda = 3\sigma\sqrt{\frac{\log(np/\delta)}{n}}$ we get the optimal rate

$$\sigma^2\left(\frac{s}{n\varkappa} + \frac{k^2}{n^2}\right)$$

3) For the Lasso, the presence of $\varkappa$ is unavoidable.

4) If we do explicitly the minimization w.r.t. θ in the lasso problem, we get

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Phi\left(\frac{y_i - x_i^T \beta}{\lambda_k \sqrt{n}} \right) + \frac{\lambda_s}{(\lambda_k \sqrt{n})^2} \|\beta\|_1 \right\}$$

where $\Phi(u) = \min\left(\frac{u^2}{2}, |u| - \frac{1}{2} \right)$ is Huber's function. (Here, we used two different λ's for the penalty term $\|\beta\|_1$ and $\|\theta\|_1$)

3 Proof

For simplicity, we assume $\Sigma = I_p$. Then, of course, we have RE with $\varkappa = 1$.

We will use only two properties of X: w.p. $\geq 1-\delta$

(P1) $$|\theta^T X \beta| \leq \underbrace{\left(3 + \sqrt{2\log(1/\delta)}\right)}_{\times \|\theta\|_2 \cdot \|\beta\|_2} + 3\sqrt{\log n}\, \|\theta\|_1 \cdot \|\beta\|_2 + 3\sqrt{\log p}\, \|\theta\|_2 \cdot \|\beta\|_1$$

(P2) $$\|X\beta\|_2 \geq \left(\sqrt{n-1} - \sqrt{2\log(1/\delta)}\right) \|\beta\|_2 - 3\sqrt{\log p}\, \|\beta\|_1.$$

Under (P1) and (P2), the matrix $[X \ \sqrt{n} I_n]$ satisfies the RE condition with $\varkappa = 1/2$.

Therefore (by standard arguments (BRT 2009))

$$\|\hat{\beta} - \beta^*\|_2^2 + \|\hat{\theta} - \theta^*\|_2^2 \leq C \cdot \lambda^2 \cdot (s+k)$$

$$\|\hat{\beta} - \beta^*\|_1 + \|\hat{\theta} - \theta^*\|_1 \leq C \cdot \lambda (s+k)$$

We cannot improve the rate for estimating θ^* but the rate of estimation of β^* can be improved using (P1).

We know that

$$\hat{\beta} \in \arg\min \left\{ \frac{1}{2n} \underbrace{\|Y - X\beta - \hat{\theta}\|_2^2}_{\|X(\beta^* - \beta) + (\theta^* - \hat{\theta}) + \xi\|_2^2} + \lambda \|\beta\|_1 \right\}$$

From KKT conditions:

$$\frac{1}{n}\|X(\hat{\beta} - \beta^*)\|_2^2 \leq \frac{1}{\sqrt{n}} \underbrace{(\hat{\beta} - \beta^*)}_{u}{}^T X^T \underbrace{(\hat{\theta} - \theta^*)}_{v} + \frac{1}{\sqrt{n}} (\hat{\beta} - \beta^*)^T X^T \xi + \lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\overset{(P1)}{\leq} \frac{C}{\sqrt{n}}\left(\|u\|_1 \cdot \|v\|_2 + \|u\|_2 \cdot \|v\|_1\right) + \frac{\lambda}{2}\|u\|_1 + \lambda\|u_J\|_1 - \lambda\|u\|_1$$

with $u = \hat{\beta} - \beta^*$ & $v = \hat{\theta} - \theta^*$.

Using this & (P2) we get

$$\|u\|_2^2 \leq \underbrace{\frac{C\|u\|_1^2}{n}}_{\text{OK}} + \frac{C}{\sqrt{n}}\left(\|u\|_1 \cdot \|v\|_2 + \underbrace{\|u\|_2 \cdot \|v\|_1}_{\leq \frac{\sqrt{n}}{4C}\|u\|_2^2 + \frac{4C}{\sqrt{n}}\|v\|_1^2}\right) + \underbrace{\frac{\lambda}{2}\|u_J\|_1}_{\text{OK}} - \lambda\|u\|_1$$

λ is chosen so that this term ≤ 0

$\boxed{\text{THANK YOU!}}$