Proof of Corr. 6.1 in PB&vdG

Basic Inequality (Lemma 6.1 in PB&vdG)

$$\|X(\hat\beta-\beta^0)\|_2^2/n + \lambda\|\hat\beta\|_1 \leq 2\varepsilon^T X(\hat\beta-\beta^0)/n + \lambda\|\beta^0\|_1$$

Proof:

$$\underbrace{\|\underset{\overset{\|}{X\beta^0+\varepsilon}}{Y}-X\hat\beta\|_2^2/n}_{} + \lambda\|\hat\beta\|_1 \leq \underbrace{\|Y-X\beta^0\|_2^2/n}_{\cancel{\|\varepsilon\|_2^2/n}} + \lambda\|\beta^0\|_1$$

$$\|X(\hat\beta-\beta^0)\|_2^2/n - 2\varepsilon^T X(\hat\beta-\beta^0)/n + \cancel{\|\varepsilon\|_2^2/n}$$

∎

From Basic Inequality: term

$$2|\varepsilon^T X(\hat{\beta}-\beta^0)/n| = 2\Big|\sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{p} X_i^{(j)}(\hat{\beta}_j-\beta_j^0)/n\Big|$$

$$\underset{\text{Hölder}}{\leq} 2 \max_{j=1,\dots,p} \Big|\sum_{i=1}^{n} \varepsilon_i X_i^{(j)}/n\Big| \, \|\hat{\beta}-\beta^0\|_1$$

$$= 2 \max_{j=1,\dots,p} |\varepsilon^T X^{(j)}/n| \cdot \|\hat{\beta}-\beta^0\|_1$$

trick: consider the set

"noise level"

$$\mathcal{T} = \Big\{ \max_{j=1,\dots,p} 2|\varepsilon^T X^{(j)}/n| \leq \lambda_0 \Big\}, \quad \lambda \overset{\text{"="}}{\geq} 2\lambda_0$$

- we will show later: $\mathbb{P}[\mathcal{T}]$ is large

- on $\mathcal{T}$:

$$\|X(\hat\beta-\beta^0)\|_2^2/n + \lambda\|\hat\beta\|_1 \overset{\text{Basic Ineq.}}{\leq} \lambda_0\|\hat\beta-\beta^0\|_1 + \lambda\|\beta^0\|_1$$

$$\underset{\Delta\text{-ineq.}}{\leq} \lambda_0\|\hat\beta\|_1 + \lambda_0\|\beta^0\|_1 + \lambda\|\beta^0\|_1 \qquad \lambda \geq 2\lambda_0$$

$$\underset{\lambda\geq 2\lambda_0}{\Longrightarrow} \|X(\hat\beta-\beta^0)\|_2^2/n \leq \frac{3}{2}\lambda\|\beta^0\|_1$$

on $\mathcal{T}$: Theorem and Cor. 6.1 is proved

remains to show that $\mathbb{P}[\mathcal{T}] \to 1 \ (n \to \infty)$

<u>Lemma 6.2 in PB&vdG (2011):</u>

assume $\hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i^{(j)})^2 \equiv 1 \quad \forall j$

Then for $\lambda_0 = 2\sigma \sqrt{\frac{t^2 + 2\log(p)}{n}}$

$$\mathbb{P}[\mathcal{T}] \geq 1 - 2\exp(-t^2/2)$$

$\rightsquigarrow$ use $t^2 = t_n^2 = \log(p_n) \rightsquigarrow \mathbb{P}[\mathcal{T}] \geq \ldots$

$$\geq 1 - \mathcal{O}(p_n^{-1}) \to 1$$

$$(n, p_n \to \infty)$$

Proof of lemma 6.2:

Consider $V_j = \varepsilon^T X^{(j)} / \sqrt{n\sigma^2}$

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

but dependent across $j = 1, \ldots, p$

$$\mathbb{P}\left[\max_{j=1,\ldots,p} |V_j| > \sqrt{t^2 + 2\log(p)}\right]$$

$$\underset{\substack{\text{union bound}\\ \text{Bonferroni bound}}}{\leq} \sum_{j=1}^{p} \mathbb{P}\left[|V_j| > \sqrt{t^2 + 2\log(p)}\right]$$

$$\underset{\text{tail bound for } \mathcal{N}(0,1)}{\leq} p \cdot 2\exp\left(-\frac{t^2 + 2\log(p)}{2}\right) = 2\exp(-t^2/2)$$

$$\mathbb{P}[\mathcal{T}] = \mathbb{P}\Big[\max_{j=1,\dots,p} 2\,|\varepsilon^T X^{(j)}/n| \le \lambda_0\Big]$$

$$\dots = \mathbb{P}\Big[\max_{j=1,\dots,p} |V_j| \le \sqrt{t^2 + 2\log(p)}\Big]$$

$$\ge 1 - 2\exp(-t^2/2)$$

$\blacksquare$

the issue with $\underbrace{C \geq \hat{\sigma} \geq \sigma}_{=: A_n}$

(e.g. $\hat{\sigma}^2 = \|Y\|_2^2/n$)

$\mathbb{P}[A_n] \longrightarrow 1$: assumption

we have analy τ with σ

new τ " $\hat{\sigma}$?

$$\mathbb{P}[\tau^c] = \mathbb{P}[\tau^c \cap A_n] + \mathbb{P}[\tau^c \cap A_n^c]$$

$$\leq \mathbb{P}[\tau^c] + \mathbb{P}[A_n^c] \leq 2\exp\left(-t^2/2\right) + o(1)$$

II.5 Estimation of β^0 and oracle inequality

if we want to ensure that $\hat{\beta} \approx \beta^0$ we necessarily need conditions on X

(e.g. $X^{(1)} = -X^{(2)}$ $\rightsquigarrow$ cannot distinguish between β_1^0 and β_2^0)

Conditions on X for identifiability

suppose $X\theta = X\beta^0$

sufficient condition on X to imply $\theta = \beta^0$?

$$0 = \| X(\theta - \beta^0) \|_2^2 / n = (\theta - \beta^0)^T \underbrace{\hat{\Sigma}}_{X^T X / n} (\theta - \beta^0)$$

$$\geq \underbrace{\boxed{\Lambda_{min}^2 (\hat{\Sigma})}}_{\text{minimal eigenvalue of } \hat{\Sigma}} \| \theta - \beta^0 \|_2^2$$

if $\Lambda_{min}^2 (\hat{\Sigma}) \neq 0 \implies \| \theta - \beta^0 \|_2^2 = 0 \implies \theta = \beta^0$

but for $p > n$: $\Lambda_{min}^2 (\hat{\Sigma}) = 0$

now suppose that we restrict θ and β^0 to be <u>sparse</u>

$$\mathrm{supp}(\theta) = \{ j ; \theta_j \neq 0 \} = S_\theta \; ; \quad |S_\theta| =: s_\theta \text{ "small"}$$

$$\mathrm{supp}(\beta^0) = \quad \cdot \text{———} \quad = S_0 \qquad |S_0| =: s_0 \text{ "small"}$$

$$\rightsquigarrow \mathrm{supp}(\theta - \beta^0) \subseteq S_\theta \cup S_0$$

$$\|\theta - \beta^0\|_0 := |\mathrm{supp}(\theta - \beta^0)| \leq s_\theta + s_0$$

we can thus consider sub-matrices $\underbrace{\hat{\Sigma}_S}_{\text{corresponds to columns and row of } S \subseteq \{1 \ldots p\}}$

$$\phi^2_{min}(m) = \min_{S \subseteq \{1, \dots, p\}} \left\{ \lambda^2_{min}(\hat{\Sigma}_S) \, ; \, |S| \leq m \right\}$$

$$\Longleftrightarrow \cancel{\phi} \; \phi^2_{min}(m) = \min_{\substack{\beta \neq 0 \\ \|\beta\|_0 \leq m}} \frac{\beta^T \hat{\Sigma} \beta}{\|\beta\|_2^2}$$

sparse eigenvalues (Meinshausen & Yu, 2009)

now it's reasonable that $\phi^2_{min}(m) > 0$ if $m < n$