36-710: Advanced Statistical Theory **Spring 2019**

Lecture 12: March 19

Lecturer: Alessandro Rinaldo *Scribes: Tom Yan*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Recall:

$$Y = X\beta^* + \epsilon$$

where X is the fixed design matrix, $\epsilon \in \mathrm{SG}_n(\sigma^2)$.

We have:

$$\beta^* = (X^TX)^{-1}X^TY$$

as the OLS solution (which can be one of infinitely many solutions).

Our target for inference is $X\beta^*$.

Theorem 12.1 *There exists universal constants $c > 0$ s.t:*

$$\frac{1}{n}||X(\hat{\beta} - \beta^*)||^2 \leq C\sigma^2(\frac{r + \log(1/d)}{n})$$

where $r = rank(X^TX)$.

Proof:

Step 1: as per last time, use basic inequality:

$$\begin{aligned} ||X(\hat{\beta} - \beta^*)||^2 &\leq 2\epsilon^T(\hat{\beta} - \beta^*) \\ &= 2\epsilon^T\frac{X(\hat{\beta} - \beta^*)}{||X(\hat{\beta} - \beta^*)||} \end{aligned}$$

And so, $||X(\hat{\beta} - \beta^*)|| \leq 2\epsilon^T\frac{X(\hat{\beta}-\beta^*)}{||X(\hat{\beta}-\beta^*)||}$ which is an unit vector in $\mathbb{R}^n$.

The wrong step here would be to bound the RHs using $\epsilon^T v \in \mathrm{SG}(\sigma^2)$, which is true for each fixed v, but not true for $\frac{X(\hat{\beta}-\beta^*)}{||X(\hat{\beta}-\beta^*)||}$ which is random and depends on ϵ.

This relates to the principle not to use data to both identify the parameter of interest and estimate the parameter.

Instead the right thing to do is to use a crude bound via discretization:

$$||X(\hat{\beta} - \beta^*)|| \leq 2\text{sup}_{v \in B_n} \epsilon^T v$$

A slightly more refined approach use that X has rank r:

Let Φ be a $n \times n$ matrix with orthonormal columns, which span the column range of X, i.e $X(\hat{\beta} - \beta^*) = \Phi z$ for some vector z.

Then:

$$\epsilon^T \frac{X(\hat{\beta} - \beta^*)}{||X(\hat{\beta} - \beta^*)||} = \epsilon^T \frac{\Phi z}{||\Phi z||} = \frac{\tilde{\epsilon}^T z}{||z||}$$

where $\tilde{\epsilon} = \Phi^T \epsilon \in \mathbb{R}^r$ and making use of the fact that $||\Phi z|| = ||z|| \Rightarrow ||X(\hat{\beta} - \beta^*)|| \leq 2\text{sup}_{z \in B_r} \tilde{\epsilon}^T z$.

We have that $\tilde{\epsilon} \sim \text{SG}_r(\sigma^2)$ (since $v^T \tilde{\epsilon} = (v^T \Phi)\epsilon \in SG(\sigma^2)$).

By continuity:

$$||X(\hat{\beta} - \beta^*)||^2 \leq 4\text{sup}_{z \in B_r} (\tilde{\epsilon}^T z)^2$$

$$\mathbb{E}[\text{sup}_{z \in B_r} (\tilde{\epsilon}^T z)^2] = \mathbb{E}[\sum_{j=1}^{r} \tilde{\epsilon}_j^2] \leq r\sigma^2$$

And so:

$$\frac{1}{n}\mathbb{E}[||X(\hat{\beta} - \beta^*)||^2] \leq 4\sigma^2 \frac{r}{n}$$

To obtain a bound about the probability, we use:

- $\text{sup}_{z \in B_r} (\tilde{\epsilon}^T z)^2 = (\text{sup}_{z \in B_r} \tilde{\epsilon}^T z)^2$
- $\text{sup}_{z \in B_r} \tilde{\epsilon}^T z \leq 2\text{max}_{w \in N_{1/2}} \tilde{\epsilon}^T w$

And so:

$$\begin{aligned} P(\text{sup}_{z \in B_r} \tilde{\epsilon}^T z \geq t) &\leq P(2\text{max}_{w \in N_{1/2}} \tilde{\epsilon}^T w \geq \sqrt{t}) \\ &\leq |N_{1/2}| \exp(-t/(8\sigma^2)) \end{aligned}$$

by Hoeffding for sub-Gaussian and union bound.

$\blacksquare$

Reflection: basic inequality, sup out, maximal inequality are common techniques.

Extensions: Let $\lambda_{min}(\frac{X^T X}{n})$ be the smallest eigenvalue of $\frac{X^T X}{n}$, assume it's positive.

Let A be PSD, then using the fact that $||X|^2 \leq \frac{X^TAX}{\lambda_{min}(A)}$ we get that:

$$||\hat{\beta}-\beta^*||^2 \leq \frac{1/n||X(\hat{\beta}-\beta^*)||^2}{\lambda_{min}(\frac{X^TX}{n})}$$

.

Penalized Regression/Lasso

Penalized regression:

$$\hat{\beta} \in \min_{\beta\in\mathbb{R}^d}||Y-X\beta||^2+\lambda_n f(\beta)$$

which includes a penalty term for the complexity of β.

A classic penalty term is $f(\beta)=||\beta||^2$ (ridge regression):

$$\beta_{ridge}=(X^X+\lambda_n I)^{-1}X^TY$$

which is always unique even if $n > d$.

The interpretation is, consider the SVD decomposition of X: $X=U\Lambda U^T$. Plugging this in:

$$\begin{aligned} X\hat{\beta}_{ridge}=X(X^TX+\lambda I)^{-1}X^TY &= U\Lambda U^TU(\Lambda^2+\lambda I)^{-1}U^TU\Lambda U^TY \\ &= UHU^TY \end{aligned}$$

where H is a diagonal matrix with $H_{jj}=\frac{\sigma_j^2}{\sigma_j^2+\lambda}$.

And so, $X\hat{\beta}_{ridge}=\sum_{j=1}^r u_j\frac{\sigma_j^2}{\sigma_j^2+\lambda}\langle\ u_j,Y\rangle$.

We can see that ridge gives higher weight to directions u_j with large σ_j^2 and may be considered a smarter projection, whereas for OLS, all basis u_j is weighted the same amount.