10-755: Advanced Statistical Theory I **Fall 2017**

Lecture 10: October 4

Lecturer: Alessandro Rinaldo *Scribes: Ciaran Evans*

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

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

10.1 Linear Models

Recall that we assume the following linear model:

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

where $(\varepsilon_1, ..., \varepsilon_n) \sim$ iid $SG(\sigma^2)$, and X is treated as fixed. In the case that $n >> d$ (many more observations that covariates), we have

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N_d(0, Q)$$

where $Q = \lim_{n\to\infty} \left(\frac{X^T X}{n}\right)^{-1}$ and $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Now, if $d > n$ or X is not full column rank, we will run into problems, because we cannot invert $X^T X$. However, it is still possible to compute an OLS solution for $\hat{\beta}$ by solving

$$(X^T X)\beta = X^T Y$$

One solution, of minimal norm, is $\hat{\beta} = (X^T X)^+ X^T Y$, where $(X^T X)^+$ is the Moore-Penrose generalized inverse of $X^T X$.

A fact about the generalized inverse is that $X^+ = (X^T X)^+ X^T$, so $\hat{\beta} = X^+ Y$. Thus, $X\hat{\beta} = XX^+ Y$. Since XX^+ is the projection onto the column space of X, then $X\hat{\beta}$ is unique, even though there may be infinitely many solutions for $\hat{\beta}$!

In particular, if $\Delta \in kernel(X)$, then $\hat{\beta} + \Delta$ is also a least squares solution. But,

$$X\hat{\beta} = X(\hat{\beta} + \Delta)$$

so the estimate of $E[Y] = X\beta^*$, $X\hat{\beta}$, is unique.

10.2 Aside: General Regression Framework

So far we've been discussing the framework in which we make very strong assumptions (e.g., linearity) about the true model. This may be dishonest, since it seems unlikely that the true model is really linear. A better,

more honest approach is the following general regression framework.

$(X, Y) \in \mathbb{R} \times \mathbb{R}^d$, having distribution P. We observed n pairs $(X_1, Y_1), ..., (X_n, Y_n) \sim$ iid P, and we want to do inference on

$$E[Y|X = x] = \mu(x).$$

Let

$$\beta^* = \text{argmin} E[(Y - \beta^T X)^2]$$

This gives us the *best linear approximation* to the regression function.

If $Cov(X) = \Sigma$ is non-singular, then

$$\beta^* = \Sigma^{-1}\alpha$$

where $\alpha = E[Y \cdot X] \in \mathbb{R}^d$. Then, $(\beta^*)^T X$ is the projection (in the L_2 sense) of Y into the set of linear functions of X, i.e. the best linear approximation of Y.

In the general regression framework, all inference delas with the best linear approximation, **not** with the true regression function (which is unknown).

10.3 Back to Linear Models with Strong Assumptions

Theorem 10.1 *There exists $c > 0$ such that*

$$MSE(\hat{\beta}) = \frac{1}{n}||X(\hat{\beta} - \beta^*)||^2 \leq c\sigma^2 \frac{r + \log(1/\delta)}{n}$$

with probability at least $1 - \delta$, for all $\delta \in (0, 1)$, where $r = rank(X)$. Further,

$$E[MSE(\hat{\beta})] \leq c\sigma^2 \frac{r}{n}$$

which is the minimax optimal rate.

Proof: By optimality of the least square solution $\hat{\beta}$,

$$\frac{1}{n}||Y - X\hat{\beta}||^2 \leq \frac{1}{n}||Y - X\beta^*||^2 = \frac{||\varepsilon||^2}{n}$$

Hence, $||Y - X\hat{\beta}||^2 - ||\varepsilon||^2 \leq 0$.

Developing squares, we have

$$\begin{aligned} ||Y - X\hat{\beta}||^2 &= ||X\beta^* + \varepsilon - X\hat{\beta}||^2 \\ &= ||X(\hat{\beta} - \beta^*)||^2 + ||\varepsilon||^2 - 2\varepsilon^T X(\hat{\beta} - \beta^*) \end{aligned}$$

Combining, we get our **basic inequality**:

$$||X(\hat{\beta} - \beta^*)||^2 \leq 2\varepsilon^T X(\hat{\beta} - \beta^*)$$

Notice that the RHS of the basic inequality is composed of two parts:

- **Random** vector: ε^T
- Vector with **structure**: $X(\hat{\beta} - \beta^*)$

Now we want to "sup out" over all possible choices of $X(\hat{\beta} - \beta^*)$. Let Φ be an $n \times r$ matrix whose columns for an orthonormal basis for the column space of X. Then,

$$X(\hat{\beta} - \beta^*) = \Phi v$$

for some unique $v \in \mathbb{R}^r$. Simplifying the basic inequality, we have

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

where $\tilde{\varepsilon} = \Phi^T \varepsilon$ and we use the fact that $||\Phi v|| = ||v||$ (since Φ has orthonormal columns). Hence,

$$||X(\hat{\beta} - \beta^*)||^2 \leq 4 \sup_{v \in \mathcal{S}^{r-1}} (\tilde{\varepsilon}^T v)^2$$

The bound in expectation follows easily. Since $\tilde{\varepsilon} \in SG_r(\sigma^2)$, then

$$\begin{aligned} 4E[\sup_{v \in \mathcal{S}^{r-1}} (\tilde{\varepsilon}^T v)^2] &= 4E\left(\sum_{i=1}^{r} (\tilde{\varepsilon}_i)^2\right) \\ &\leq 4r\sigma^2 \end{aligned}$$

because $\tilde{\varepsilon}_i \in SG(\sigma^2)$ and so $Var(\tilde{\varepsilon}_i) \leq \sigma^2$ (recall that the ε_i have mean 0).

Now to get the bound in probability, we can proceed as follows. We have

$$\sup_{v \in \mathcal{S}^{r-1}} (\tilde{\varepsilon}^T v)^2 \leq 2\tilde{\varepsilon}^T z \qquad x \in \mathcal{N}_{1/2}$$

and furthermore,

$$\sup_{v \in \mathcal{S}^{r-1}} (\tilde{\varepsilon}^T v)^2 = (\sup_{v \in \mathcal{S}^{r-1}} \tilde{\varepsilon}^T v)^2.$$

Thus,

$$\begin{aligned} P(\sup_{v \in \mathcal{S}^{r-1}} (\tilde{\varepsilon}^T v)^2 \geq t) &\leq P(2 \max_{z \in \mathcal{N}_{1/2}} \tilde{\varepsilon}^T z \geq \sqrt{t}) \\ &\leq |\mathcal{N}_{1/2}| \exp\left(-\frac{t}{8\sigma^2}\right) \end{aligned}$$

by Hoeffding. Set this to be $\leq \delta$ to get the result with the actual constants. ■

Now let's review the steps of the proof.

1. Basic inequality (upper bound, often in terms of inner product of random noise with vector with structure)
2. sup-out (over vector with structure)
3. Discretization + union bound and concentration

10.3.1 Remark: Estimating the Betas

If $d < n$ and X has rank d, then

$$||\hat{\beta} - \beta^*||^2 \leq \frac{MSE(\hat{\beta})}{\lambda_{\min}(\frac{X^T X}{n})} \leq \frac{1}{\lambda_{\min}} \sigma^2 d \frac{\log(1/d)}{n}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of the matrix. This makes use of the Courant-Fisher theorem, which for PSD matrices A gives

$$||x||^2 \lambda_{\min}(A) \leq x^T A x.$$

10.4 Penalized Linear Regression

We still have the model

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

In the case when $d >> n$ and we assume that Y only depends on a (small) subset of the parameters, it might be better to use penalized regression than ordinary least squares. Penalized regression solves

$$\hat{\theta} \in \text{ argmin } \left\{ \frac{1}{2n} ||Y - X\theta||^2 + \lambda_n f(\theta) \right\}$$

where $\lambda_n > 0$ and $f(\theta)$ is a measure of how "complex" the candidate solution is.

Some choices for $f(\theta)$:

- $f(\theta) = ||\theta||_0$ (best subset selection)
- $f(\theta) = ||\theta||_2^2$ (ridge regression)
- $f(\theta) = ||\theta||_1$ (LASSO)