36-710: Advanced Statistical Theory **Fall 2018**

Lecture 21: April 4, 2019

Lecturer: Alessandro Rinaldo *Scribes: Max Rubinstein*

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

21.1 Oracle Inequalities

We first return to the proof of the OLS oracle inequality.

Theorem 21.1 *For $\delta \in (0,1)$, with $p \geq 1-\delta$, $MSE(\hat{f}_{OLS}) \leq \inf_{\theta \in \mathbb{R}^m} MSE(f_\theta) + C\sigma^2(\frac{M+\log(\frac{1}{\delta})}{n})$*

Proof: Let $Y = f^* + \epsilon$. Then

$$\frac{1}{n}||Y - \hat{f}_{ols}||^2 \leq \frac{1}{n}||Y - f_{oracle}||^2 = \frac{1}{n}(||f^* - \hat{f}_{ols}||^2 - ||f^* - f_{oracle}||^2) \leq \frac{2}{n}\epsilon'(\hat{f}_{ols} - f_{oracle})$$

We then divide each side by $||\hat{f}_{ols} - f_{oracle}||$ to get

$$\frac{(||f^* - \hat{f}_{ols}||^2 - ||f_{oracle} - \hat{f}^*||^2)}{||\hat{f}_{ols} - f_{oracle}||} \leq \frac{2\epsilon'(\hat{f}_{ols} - f_{oracle})}{||\hat{f}_{ols} - f_{oracle}||} \leq 2 \sup_{v \in \mathbb{S}^{n-1}} \epsilon'\nu$$

Because both $\hat{f}$ and f_{oracle} lie in $\text{span}\{f_1, ..., f_n\}$, we can invoke the Pythagorean Theorem conclude

$$||f^* - \hat{f}_{ols}||^2 - ||f_{oracle} - \hat{f}^*||^2 = ||\hat{f} - f_{oracle}||^2$$

Substituting this in the LHS and cancelling the square from the denominator, squaring both the LHS and RHS expressions and dividing by n, we conclude that

$$\frac{1}{n}||\hat{f} - f_{oracle}||^2 \leq \frac{1}{n}(2 \sup_{v \in \mathbb{S}^{n-1}} \epsilon'\nu)^2 \leq \frac{\sigma^2 m}{n}$$

where we get the final inequality from the previous proof of the OLS bound. Again using the Pythagorean Theorem we substitute back to get that

$$MSE(\hat{f}) \leq MSE(f_{oracle}) + \frac{\sigma^2 m}{n}$$

■

21.2 Principal Components Analysis

Let $X \in \mathbb{R}^d$, $\mathbb{E}(X) = 0$, and $Cov(X) = \Sigma$, which is a d by d positive semi-definite matrix. Then let i index the order of the eigenvalues of Σ and U be a d by d orthonormal matrix and Λ is a d by d diagonal matrix where $\Lambda = diag(\lambda_i)_{i \in 1,...,d}$. We then have that

- $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d \geq 0$
- $\Sigma = U \Lambda U'$

21.2.1 What is PCA?

We now outline three view of principal components analysis (PCA).

21.2.1.1 View 1: Maximal Variance

Find a direction $\nu \in \mathbb{S}^{d-1}$ along which X has maximal variance $\nu^* = \arg\max_{\nu \in \mathbb{S}^{d-1}} Var(\nu' X) = \nu' \Sigma \nu$. Then ν^* is the largest (leading) eigenvector of Σ (1st column of U in the SVD) and is the direction of maximal variance.

More generally, we can find the r directions of largest variance:

$$V^*_{dxr} = \arg \max_{\nu \in \mathbb{V}_{d,r}} \mathbb{E}||V' X||^2$$

where $\mathbb{V}_{dxr} = \{V_{dxr}$ w/ orthonormal columns$\}$. Then $\mathbb{E}||X'V||^2 = \sum_{j=1}^r \mathbb{E}[(X'V_j)^2]$ where V_J is the jth column of V. The solution V^* consists of the first r leading eigenvectors of Σ (or the first r columns of U) and $\mathbb{E}||V^{*\prime} X||^2 = \sum_{j=1}^r \lambda_r$. In other words, the solution is an orthogonal projection of X onto the linear subspace spanned by $\nu \in \mathbb{S}^{d-1}$.

21.2.1.2 View 2: Optimal Projection

Suppose we want to find the optimal linear subspace S of $\mathbb{R}^d$ of dimension $1 \leq r \leq d$ such that

$$\mathbb{E}(||X - \pi_s X||^2)$$

is minimal and $\pi_s X$ is a projection of X onto S. Then $\pi_s = U_r U_r'$ where U_r are the first r columns of U (note that π_s is d by d, U_r is d by r). Moreover, we have that

$$\mathbb{E}||X - \pi_s X||^2 = \sum_{j=r+1}^{d} \lambda_j$$

21.2.1.3 View 3: Matrix Approximation

We can also think about the matrix Z^* that minimizes $||Sigma - Z||_F^2$ where Z is a matrix such that $rank(Z) \leq r$. Then we have that

$$Z^* = \sum_{j=1}^{r} \lambda_j U_j U_j'$$

and

$$||\Sigma - Z^*||_F^2 = \sum_{j=r+1}^{d} \lambda_j^2$$

ie Z is a low-rank approximation to Σ.

21.2.2 Estimating Eigenvalues and Eigenvectors

Our goal is to estimate $\lambda_1, ..., \lambda_r$ and $u_1 \geq ... \geq u_r$. In general we can think of

$$\hat{\Sigma}_n = \Sigma + E$$

where E is not necessarily positive definite. Then the problem becomes

$$\lambda_{max}(\hat{\Sigma}) = \lambda_{max}(\Sigma + E) \leq \lambda_{max}(\Sigma) + ||E||_{op}$$

In fact, $|\lambda_{max}(\Sigma) - \lambda_{max}(\hat{\Sigma})| \leq ||E||_{op}$ and Weyl's Inequality (which has a non-obvious proof) further provides us with a uniform bound on the target eigenvalues and the observed perturbed eigenvalues

$$\max_j |\lambda_j(\Sigma) - \lambda_j(\hat{\Sigma})| \leq ||E||_{op}$$

Notice that $||E||_{op} = ||\hat{\Sigma}_n - \Sigma||_{op}$, a quantity we've already studied. In brief, estimating λ_i is relatively easy; by contrast, estimating U_i is difficult.

We illustrate the difficulty of estimating eigenvalues through the following example:

$$A = I_2 + \begin{pmatrix} \epsilon & 0 \\ 0 & \epsilon \end{pmatrix}$$

$$B = I_2 + \begin{pmatrix} 0 & \epsilon \\ \epsilon & 0 \end{pmatrix}$$

We see that $\lim_{\epsilon \to 0} A = I_2$ and $\lim_{\epsilon \to 0} B = I_2$. Moreover, in both cases $||E||_{op} = \epsilon$.

However, for A the eigenvalues are (1, 0) and (0, 1) (our target subspace), while for B the eigenvalues are $1 + \epsilon$ and $1 - \epsilon$. In general, PCA might fail when the eigenvectors are too close to each other. Next lecture

we will discuss how to measure the distance between subspaces so that we can consider the problem of estimating them.