36-755: Advanced Statistical Theory 1	**Fall 2016**

Lecture 11: October 10

Lecturer: Alessandro Rinaldo — *Scribes: Pengtao Xie*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

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

11.1 Persistence

Setup: $Z_1, \cdots, Z_n$ are i.i.d samples drawn from distribution P, where $Z_i = (Y_i, X_i)$ with $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^d$, and $Y_i = f(X_i) + \epsilon$. f can be any function. We want to predict Y using vector X. We are only using linear predictors. Formally, for any $\beta \in \mathbb{R}^d$,let

$$R_P(\beta) = \mathbb{E}_P[(Y - X^\top \beta)^T] \tag{11.1}$$

We assume $\text{cov}[X] = \Sigma$ is non-singular, then the problem

$$\min_{\beta \in \mathbb{R}^d} R_P(\beta) \tag{11.2}$$

has unique solution $\beta^* = \Sigma^{-1}\alpha$ where $\alpha = \mathbb{E}[YX]$.

Suppose we have a sequence $\{P_n\}$ of probability distribution for $Z = (Y, X) \in \mathbb{R}^{d+1}$ where $d = d(n)$. We also have a sequence of sets $\{B_n\}$ where $B_n \subset \mathbb{R}^{d(n)}$. For each n, let the optimal constrained parameters be

$$\beta_n^* \in \text{argmin}_{\beta \in B_n} R_{P_n}(\beta) \tag{11.3}$$

Example of B_n: (1) $B_n = \{\theta \in \mathbb{R}^{d(n)}, \|\beta\|_1 \le b_n\}$ where $b_n > 0$; (2) $B_n = \{\theta \in \mathbb{R}^{d(n)}, \|\beta\|_0 \le k_n\}$

Definition of persistence: given a sequence $\{(P_n, \beta_n^*)\}$, a sequence of estimators $\{\hat{\beta}_n\}$ is persistent if $R_{P_n}(\hat{\beta}_n)$ converges to $R_{P_n}(\beta_n^*)$ in probability.

We will be looking at

$$\hat{\beta}_n = \text{argmin}_{\beta \in B_n} \hat{R}(\beta) \tag{11.4}$$

where

$$\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta) \tag{11.5}$$

Let $\widetilde{\Sigma} = \text{cov}[Z]$ and $\hat{Z}$ be the empirical covariance. Assume that $\|\widetilde{\Sigma} - \hat{\Sigma}\|_\infty = \max_{ij}|\widetilde{\Sigma}_{ij} - \hat{Sigma}_{ij}| \le \Delta_n(\delta)$ with probability $1 - \delta$ for all n and P_n. For $\beta \in \mathbb{R}^{d+1}$, let $\tilde{\beta} = (1, -\beta) \in \mathbb{R}^{d+1}$, then $y - x^\top \beta = z^\top \tilde{\beta}$. Let $\tilde{B}_n = \{(-1, \beta) \in \mathbb{R}^{d+1}, \beta \in B_n\}$, then $R_P(\beta) = R_p(\tilde{\beta})$.

Theorem: Assume $d = n^\alpha$ where $\alpha > 0$. Then

$$R_{P_n}(\hat{\beta}) \le R_{P_n}(\tilde{\beta}^*) + 2\Delta_n(b_n + 1)^2 \tag{11.6}$$

Proof: $R_{P_n}(\tilde\beta) = \tilde\beta^\top \tilde\Sigma \tilde\beta$ and $\hat R_{\tilde\beta} = \tilde\beta^\top \hat\Sigma \tilde\beta$. Then $\forall \tilde\beta \in \mathbb{R}^{d+1}$ and P_n, we have

$$\begin{aligned} &|R_{P_n}(\tilde\beta) - \hat R_{\tilde\beta}| = |\tilde\beta^\top(\tilde\Sigma - \hat\Sigma)\tilde\beta| \\ &\le \|\tilde\Sigma - \hat\Sigma\|_\infty \|\tilde\beta\|_1 (\text{Holder Inequality}) \\ &\le \Delta_n(\delta)(b_n+1)^2 \end{aligned} \tag{11.7}$$

Then

$$\begin{aligned} &R_{P_n}(\hat{\tilde\beta}_n) \le \hat R(\hat\beta_n) + \Delta_n(\delta)(b_n+1)^2 \\ &\le \hat R(\hat\beta_n^*) + \Delta_n(\delta)(b_n+1)^2 \\ &\le \hat R_{P_n}(\hat\beta_n^*) + \Delta_n(\delta)(b_n+1)^2 \end{aligned} \tag{11.8}$$

Remark: If $\Delta_n(\delta) \preceq \sqrt{\frac{\log d}{n} + \frac{\log(1/\delta)}{n}} \preceq \sqrt{\frac{\log n}{n}}$

If $d = n^\alpha$, $\delta = \frac{1}{n}$ and $\tilde\beta_n = \{\tilde\beta \in \mathbb{R}^{d+1} | \|\tilde\beta\|_1 \le b_n + 1\}$. Then $\hat{\tilde\beta}_n$ is persistent if $b_n = o((\frac{n}{\log n})^{\frac{1}{4}})$

11.2 PCA

Let $X \in \mathbb{R}^d$ be a random vector with $cov[X] = \Sigma$. Let $\lambda_i(\Sigma)$ be the eigenvalue of Σ and u_i be the eigenvector associated with $\lambda_i(\Sigma)$. Assume $\lambda_{\max} = \lambda_1(\Sigma) \ge \cdots \ge \lambda_d(\Sigma) \ge 0$.

PCA has several interpretations.

Optimal Linear Subspace: what is the direction $v \in \mathbb{S}^{d-1}$ such that $\text{var}[v^\top X]$ is maximal?

$v^* = \text{argmax}_{v \in \mathbb{S}^{(d-1)}} \text{var}[v^\top X]$ is the eigenvector associated to $\lambda_{\max}(\Sigma)$.

More generally, let $V_{d\times r} = \{V_{d\times r} \text{with orthogonal columns}\}$. The optimal solution of

$$\text{argmax}_{V \in V_{d\times r}} \mathbb{E}[\|V^\top X\|^2] \tag{11.9}$$

is the first r eigenvectors.

Low-rank Approximation: We want to find matrix Z^* such that

$$\begin{aligned} &Z^* \in \text{argmin} \|\Sigma - Z\|_F^2 \\ &s.t. \quad \text{rank}(Z) = r \end{aligned} \tag{11.10}$$

Then $Z^* = \sum_{i=1}^r \lambda_i \mu_i \mu_i^\top$ and $\|Z^* - \Sigma\|_F^2 = \sum_{j=i+1}^d \lambda_j^2$

Subspace: suppose we want to find subspace S of R^d of dimension $r \le d$.

$$\mathbb{E}\|X - \Pi_S X\|^2 \tag{11.11}$$

where Π_S is the orthonormal projection of X onto S. Then $\Pi_S = V_r V_r^\top$ where the columns of V_r are r largest eigenvectors.

The challenges are that we need to estimate eigenvalues and eigenvectors well.