36-709: Advanced Statistical Theory I **Spring 2019**

Lecture 24: April 18

Lecturer: Alessandro Rinaldo *Scribes: Arnav Choudhry*

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.

24.1 Sparse PCA

In the last class we had looked at sparse PCA.

Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} (0, \Sigma)$, where $X_i \in \mathcal{SG}_d(||\Sigma||_{op})$ for all $i \in \{1, \dots, n\}$. We assumed a spiked covariance model $\Sigma = \theta\nu\nu^T + I_d$, where $\theta > 0, \nu \in \mathbb{S}^{d-1}$ and $||\nu||_0 \leq k \leq d/2$ serves as the structural assumption of sparsity. Now let

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$$

One of the solution approaches to find $\hat{\nu}$ is,

$$\hat{\nu} \in \underset{\substack{\theta \in \mathbb{S}^{d-1} \\ ||\theta||_0 \leq k' \leq d/2}}{\arg\max} \; \theta^T \hat{\Sigma}_n \theta \tag{24.1}$$

such that $k' \leq k$. We reasoned that it is intractable in practice due to high computational costs, and established a bound as shown in the following theorem.

Theorem 24.1 *Given the spiked covariance model described above along with the conditions on θ and ν, and let $\hat{\nu}$ be defined as in 24.1. Then with probability at least $1 - \delta$, $\delta \in (0, 1)$:*

$$\min_{\epsilon \in \{-1,1\}} ||\epsilon\hat{\nu} - \nu|| \leq C\frac{1+\theta}{\theta}\max\{\sqrt{\eta_n}, \eta_n\}$$

where C is a constant, θ is the eigengap and

$$\eta_n = \frac{(k+k')\log\frac{de}{k+k'} + \log\left(\frac{1}{\delta}\right)}{n}$$

Proof: Continuing from last class, we saw that

$$\theta\sin^2(\angle(\nu, \hat{\nu})) \leq \langle \hat{\Sigma}_s - \Sigma_s, \hat{\nu}_s\hat{\nu}_s^T - \nu_s\nu_s^T \rangle$$

Using the Hölder inequality with the p-Schatten norm,

$$\theta\sin^2(\angle(\nu, \hat{\nu})) \leq ||\hat{\Sigma}_s - \Sigma_s||_\infty ||\hat{\nu}_s\hat{\nu}_s^T - \nu_s\nu_s^T||_1 \tag{24.2}$$

Where $||\cdot||_\infty$ is the $L_\infty-$Schatten norm, equivalent to the operator norm. $||\cdot||_1$ is the 1-Schatten norm. Focusing on the 1-Schatten norm,

$$\begin{aligned}||\hat{\nu}_s\hat{\nu}_s^T - \nu_s\nu_s^T||_1 &\leq \sqrt{2}||\hat{\nu}_s\hat{\nu}_s^T - \nu_s\nu_s^T||_2 \quad \text{[using Cauchy-Schwarz inequality]} \\ &= \sqrt{2}||\hat{\nu}_s\hat{\nu}_s^T - \nu_s\nu_s^T||_F\end{aligned}$$

Note that from hereon, we can remove the restriction to s, as we work with the eigenvector ν. As an exercise, the readers are asked to convince themselves of the fact that $||\hat{\nu}_s\hat{\nu}_s^T - \nu_s\nu_s^T||_F = \sqrt{1-(\nu^T\hat{\nu})^2}$ (It can be done with the fact that $||A||_F^2 = \langle A, A\rangle$). Using this fact, we have now,

$$\begin{aligned}||\hat{\nu}_s\hat{\nu}_s^T - \nu_s\nu_s^T||_1 &\leq \sqrt{2\left(1-(\nu^T\hat{\nu})^2\right)} \\ &= \sqrt{2\sin^2\left(\angle(\nu,\hat{\nu})\right)} \qquad (24.3)\end{aligned}$$

Substituting Equation 24.3 in Equation 24.2, we get

$$\begin{aligned}\theta\sin^2(\angle(\nu,\hat{\nu})) &\leq ||\hat{\Sigma}_s - \Sigma_s||_{op}\sqrt{2\sin^2\left(\angle(\nu,\hat{\nu})\right)} \\ \theta\sin(\angle(\nu,\hat{\nu})) &\leq \sqrt{2}||\hat{\Sigma}_s - \Sigma_s||_{op} \qquad (24.4)\end{aligned}$$

Next, as we saw before,

$$\min_{\epsilon\in\{-1,1\}}||\epsilon\hat{\nu}-\nu||^2 \leq 2\sin^2\left(\angle(\nu,\hat{\nu})\right)$$

We can use this to bound Equation 24.4. Therefore, we get

$$\min_{\epsilon\in\{-1,1\}}||\epsilon\hat{\nu}-\nu|| \leq \frac{\sqrt{8}}{\theta}||\hat{\Sigma}_s - \Sigma_s||_{op} \quad \text{a.s.} \qquad (24.5)$$

It should be noted here that, it is possible to reach this point in the proof using the Davis-Kahan theorem. It is not possible to use a concentration inequality for $||\hat{\Sigma}_s - \Sigma_s||_{op}$ as Σ_s is not a fixed matrix, it is random. Hence, to solve it we use the sup-out argument to write,

$$\min_{\epsilon\in\{-1,1\}}||\epsilon\hat{\nu}-\nu|| \leq \frac{\sqrt{8}}{\theta}\max_{\substack{T\subseteq\{1,\ldots,d\}\\ \text{s.t. } |T|\leq k+k'}}||\hat{\Sigma}_T - \Sigma_T||_{op}$$

At this point, if we knew something about the way s is selected, we could have a better bound. However, here we get a union bound.

$$\mathbb{P}\left(\max_{\substack{T\subseteq\{1,\ldots,d\}\\ \text{s.t. } |T|\leq k+k'}}||\hat{\Sigma}_T - \Sigma_T||_{op} \geq t||\Sigma||_{op}\right) \leq \binom{d}{k+k'}9^{k+k'}\exp\left\{-\frac{n}{2}\max\left\{\left(\frac{t}{32}\right)^2, \frac{t}{32}\right\}\right\} \qquad (24.6)$$

Here we get the $\binom{d}{k+k'}$ term by counting all the possible choices of T, and the rest of the bound by fixing T and and constructing an ϵ-net for the operator norm. Now we use the fact that

$$\binom{d}{k+k'} \leq \left(\frac{de}{k+k'}\right)^{k+k'}$$

to bound Equation 24.6.

$$\mathbb{P}\left(\max_{\substack{T\subseteq\{1,\ldots,d\}\\ \text{s.t. } |T|\leq k+k'}}||\hat{\Sigma}_T - \Sigma_T||_{op} \geq t||\Sigma||_{op}\right) \leq \exp\left\{-\frac{n}{2}\max\left\{\left(\frac{t}{32}\right)^2, \frac{t}{32}\right\} + (k+k')\log 9 + (k+k')\log\frac{de}{k+k'}\right\} \qquad (24.7)$$

It should be noted that in Equation 24.7 the bound does not depend on d. The proof is completed by setting the RHS in Equation 24.7 to δ and solving for t. ■

24.2 Uniform Law of Large Numbers (ULLN)

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mu$, then Law of Large Numbers says that

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

This is often times not enough, as we might want to estimate the Cumulative Distribution Function (CDF) well at all points $x \in \mathbb{R}$, as illustrated in the following example.

Example. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p$ with CDF F where

$$F(x) = \mathbb{P}(X_1 \leq x) \tag{24.8}$$

Fixing $x \in \mathbb{R}$, to estimate $F(x)$ we can use the empirical CDF,

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq x\} \tag{24.9}$$

Note that $\sum_{i=1}^{n} 1\{X_i \leq x\} \sim \text{Binomial}(F(x), n)$, which has mean $nF(x)$. Hence, we can say that $\hat{F}_n(x) \xrightarrow{p} F(x)$. In fact we can get finite sample bounds, but it is not enough if we want to estimate $F(\cdot)$ well, at all points $x \in \mathbb{R}$.

Theorem 24.2 *The Glivenko-Cantelli Theorem*

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \xrightarrow{a.e.} 0$$

where the empirical CDF, $\hat{F}_n(x)$, is defined as in Equation 24.9 and $F(x)$ is the true CDF as in Equation 24.8.

This is transitioning from the law of large numbers to the Uniform Law of Large Numbers (ULLN) which we will see in the next few sections.

24.3 More abstract version

We are going to be referencing a lot of empirical process theory here. For more details the reader is refered to [VW96].

Let p be a probability distribution over some space $\mathcal{X}$. Let $\mathcal{F}$ be a class of functions of the form $f : \mathcal{X} \to \mathbb{R}$. And finally, let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$. The empirical measure associated to $X_1, \ldots, X_n$ is the random probability measure of the form

$$A \subseteq \mathcal{X} \to P_n(A)$$

where

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \in A\}$$

For $f \in \mathcal{F}$, let

$$P(f) = \mathbb{E}[f(X)]$$

and

$$P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

where $X \sim p$. So, our target is to evaluate the quantity

$$||P_n - P||_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |P_n(f) - P(f)|$$

which is the suprmum of empirical processes. Back to the example of estimating the CDF F, we have

$$\mathcal{F} = \{1(-\infty, x], x \in \mathbb{R}\}$$

When $f(\cdot) = 1(-\infty, x](\cdot)$, then we can write

$$P(f) = \mathbb{E}[1(-\infty, x](X)] = \mathbb{P}(X \leq x) = F(x)$$

It can be similarly argued that

$$P_n(f) = \hat{F}_n(x)$$

and finally from Glivenko-Cantelli theorem,

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{f\in\mathcal{F}} |P_n(f) - P(f)| = ||P_n - P||_{\mathcal{F}} \tag{24.10}$$

To see more, the reader is directed to Section 4.2.1 of [W19].

24.4 ULLN using Rademacher complexity

We can interpret the Rademacher random variable (ϵ) as essentially being a random sign (it takes values $+1$ and -1 with probability 0.5). We will use this fact in this section to quantify the magnitude of class $\mathcal{F}$.

Let us fix $\mathcal{F}$ and the n-tuple $x_1^n = (x_1, \ldots, x_n)$. Let

$$\mathcal{F}(x_1^n) = \{(f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n\}$$

for $f \in \mathcal{F}$ and $f \subseteq \mathbb{R}^n$. The *empirical* Rademacher complexity of $\mathcal{F}$ at x_1^n is

$$R_n(\mathcal{F}(x_1^n)) = \mathbb{E}_{\epsilon=(\epsilon_1,\ldots,\epsilon_n)}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} f(x_i)\epsilon_i\right|\right]$$

and the Rademacher complexity of $\mathcal{F}$ is

$$\begin{aligned} R_n(\mathcal{F}) &= \mathbb{E}\left[R_n(\mathcal{F}(X_1^n))\right] \\ &= \mathbb{E}_{X_1^n,\epsilon}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} f(X_i)\epsilon_i\right|\right] \end{aligned}$$

Now if we can show that if $R_n(\mathcal{F}) \to 0$ then we get ULLN. Note the following theorem which we will explore in the next class, towards this end.

Theorem 24.3 *Let $\mathcal{F}$ be a class of real valued functions on $\mathcal{X}$ such that $||f||_\infty = \sup_{x\in\mathcal{X}} |f(x)| \leq b < \infty$ for $f \in \mathcal{F}$. Then the*

$$\mathbb{P}\left(||P_n - P||_{\mathcal{F}} \leq 2R_n(\mathcal{F}) + t\right) \geq 1 - \exp\left\{-\frac{nt^2}{2b^2}\right\}$$

Here we can see that if $R_n(\mathcal{F}) \to 0$, then $||P_n - P||_{\mathcal{F}} \xrightarrow{p} 0$

References

[VW96] Aad W. van der Vaart and Jon A. Wellner, "Weak Convergence and Empirical Processes With Applications to Statistics" *Springer*, 1996.

[W19] Martin J. Wainwright, "High-dimensional statistics: A non-asymptotic viewpoint" *Vol. 48. Cambridge University Press*, 2019.