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

Lecture 2: January 17

Lecturer: Alessandro Rinaldo *Scribes: Charvi Rastogi*

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

In the last lecture, we started discussing high dimensional statistics. In this lecture we look at how the statistical models change as the dimension of the problem grows.

2.1 Examples of high dimensional statistical models

2.1.1 Covariance Estimation

In the problem setting we obtain vector samples $X_1, X_2, \cdots X_n \overset{\text{iid}}{\sim} (0, \Sigma)$ in $\mathbb{R}^d$ where Σ is a $d \times d$ matrix. We want to estimate Σ using the empirical covariance matrix, given by $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$. Note that the empirical covariance matrix is an unbiased estimator of the covariance matrix, ie $\mathbb{E}[\widehat{\Sigma}_n] = \Sigma$.

We are interested in finding $\|\widehat{\Sigma}_n - \Sigma\|_\infty$, to quantify the goodness of the estimator. If this is fairly small, we could possibly say we have a good estimator. But we can't be sure if the estimate is Positive definite or not. How do we measure this?

Note: For $d \times d$ matrix A, $\|A\|_\infty = \max_{i,j} |A_{i,j}|$

There are two cases that we need to consider. Case 1, wherein d is fixed and Case 2 wherein the dimension of the problem d grows with n.

2.1.1.1 Fixed d

For a given pair (i, j) in $\langle 1, \cdots, d\rangle$, let $\widehat{\Sigma_{n(i,j)}} = \frac{1}{n}\sum_{k=1}^n Z_k^{(i,j)}$ where $Z_k^{(i,j)} = X_{k,i}X_{k,j}$. This implies that every entry is an average of product of two things. In particular, $Z_1^{(i,j)}, \cdots, Z_n^{(i,j)}$ are iid with $\mathbb{E}[\widehat{\Sigma_{n(i,j)}}] \to \Sigma_{(i,j)}$. By WLLN (weak law of large numbers),

$$\widehat{\Sigma_{n(i,j)}} \xrightarrow{P} \Sigma_{(i,j)} \quad \forall (i,j)$$

Following this, we see that

$$\|\widehat{\Sigma_n} - \Sigma\|_\infty \leq \sum_{i,j} |\widehat{\Sigma_{n(i,j)}} - \Sigma_{(i,j)}| \tag{2.1}$$

Since $|\widehat{\Sigma_{n(i,j)}} - \Sigma_{(i,j)}| \xrightarrow{P} 0 \ \forall (i,j)$, each term can be expressed as $op(1)$.

Aside: Last time, we defined $o(n)$. In particular, if $x_n = o(1)$, this is equivalent to saying that, $x_n \to 0$ as $n \to \infty$. Here, x_n represents a deterministic sequence. What if we had random sequences?

If $\{X_n\}_{n=1,2,\cdots}$ is a sequence of random vectors and $\{y_n\}_{n=1,2,\cdots}$ is a sequence of positive numbers, then

$$X_n = op(1) \iff X_n \xrightarrow{P} 0.$$

This tells us that Eq. (2.1) can be expressed as

$$\|\widehat{\Sigma_n} - \Sigma\|_\infty \le \sum_{i,j} op(1) = \frac{d(d+1)}{2} op(1) \tag{2.2}$$

If d is fixed as n goes to infinity, $\|\widehat{\Sigma_n} - \Sigma\|_\infty \le op(1)$ since the rest can be written of as a constant. Furthermore, if $Z_k^{(i,j)}$ has a second moment(that is entries of the random vector have a fourth moment) then, by CLT

$$\|\widehat{\Sigma_n} - \Sigma\|_\infty + Op\Big(\frac{1}{\sqrt{n}}\Big) \tag{2.3}$$

This provides a rate of convergence for the estimator chosen.

Aside: The Big-O notation may be familiar, and is defined for deterministic sequences, say $\{x_n\}, \{y_n\}$. If $x_n = O(y_n)$, $\exists\, c > 0$ and $n_0 = n_0(c)$ such that $\forall\ n > n_0$, $\frac{|x_n|}{|y_n|} < c$. Similarly, for a sequence of random vectors $\{X_n\}$ and a sequence of positive numbers y_n where $X_n = Op(y_n)$, $\forall\ \epsilon > 0$, $\exists c = c(\epsilon)$ such that $\forall\ n > n_0 : P\big(\frac{\|x_n\|}{y_n} > c\big) < \epsilon$. This implies that the sequence of random vectors is bounded in probability.

Continuing with our covariance estimation problem, let $X_1, \cdots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$. Then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu \tag{2.4}$$

$$\overline{X}_n = \mu + op(1) \tag{2.5}$$

By central limit theorem,

$$\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0,1) \tag{2.6}$$

$$\overline{X}_n = \mu + Op(\frac{1}{\sqrt{n}}). \tag{2.7}$$

We are ignoring σ here because it is a constant. The statement obtained through CLT implies the first statement and also gives us a rate.

2.1.1.2 d increases with n

If d is a function of n, we need different tools/language. In HW1 you'll show that with probability at least $1 - \frac{1}{n}$,

$$\|\widehat{\Sigma}_n - \Sigma\|_\infty \le C\Big(\frac{\log d_n + \log n}{n}\Big)^{\frac{1}{2}}$$

$$\|\widehat{\Sigma}_n - \Sigma\|_\infty = Op\Big(\frac{\log d_n}{n}\Big)^{\frac{1}{2}}$$

The increased rate of convergence shows the price you pay for the growing dimension. This may be a misleading result because it seems to imply you can do well for $d >> n$ but you should recall that the metric under study isn't a good one to begin with.

2.2 High Dimensional Probability Distributions

Commonly known probability distributions do not look similar in a high dimensional space, imagining how they behave isn't necessarily intuitive. However, the good part is that they tend to concentrate [keithball].

For example, consider the Euclidean unit ball. Take $r > 0$, and $\|x\| = \sqrt{\sum_i x_i^2}$ is the Euclidean norm, then the Euclidean ball is given by

$$B_d(0, r) = \{x \in \mathbb{R}^d \ : \ \|x\| \leq r\}.$$

The infinity norm is defined as $\|x\|_\infty = \max_i |x_i|$. Let the cube be defined as

$$C_d(0, r) = \{x \in \mathbb{R}^d \ : \ \|x\|_\infty \leq r\}$$

In two dimensions the Euclidean unit ball, $B - 2(0, 1)$ is a circle with radius 1 and the unit cube $C_2(0, 1)$ is a square symmetric about the origin with each side = 2.

Let's look at the volume of the sets considered above. Volume of the Euclidean norm ball $B_d(0, r) = r^d v_d$, where $v_d = \mathrm{Vol}(B_d(0, 1))$.

$$v_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(d/2 + 1)} \overset{\text{large}d}{\sim} \left(\frac{2\pi e}{d}\right)^{\frac{d}{2}}.$$

The gamma function is given by $\Gamma(x) = \int_0^\infty \exp(-z) z^{x-1} dz$. Note that the volume of the Euclidean unit ball goes to zero really fast in high dimensions. Although, this doesn't hold for $C_d(0, 1)$ which is equal to 2^d even in higher dimensions.

Assume X is uniformly distributed over $B_d(0, 1)$, $\mathbb{E}[\|x\|] = \frac{d}{d+1}$. Now, pick $\epsilon \in (0, 1)$

$$P(1 - \epsilon \leq \|x\|) = \frac{v_d - (1 - \epsilon)^d v_d}{v_d} = 1 - (1 - \epsilon)^d \geq 1 - \exp(-\epsilon d).$$

The probability that $\|x\|$ is close to 1 goes to 1 exponentially fast in d. Similarly, for the normal distribution, if $X \sim \mathcal{N}_d(0, I_d)$, then with high probability $\|x\| \sim \sqrt{d}$. This implies that if you distribute points according to the normal distribution, the whole space never gets filled in.

Let's go back to the unit cube, $C_d(0, 1) = \{x \in \mathbb{R}^d \ : \ \|x\|_\infty \leq 1\}$. It turns out that

$$\lim_{d \to \infty} \mathbb{P}\Big(\frac{\sqrt{d}}{3}(1 - \epsilon) \leq \|X\| \leq \frac{\sqrt{d}}{3}(1 + \epsilon)\Big) = 1 \ \ \forall \, \epsilon \in (0, 1).$$

Ref [mledoux] for more detailed explanations and discussions. The main idea is that if $X_1, \cdots, X_n$ are independent random variables and $f : \mathbb{R}^n \to R$ such that it doesn't depend too much on any of its coordinates, then $f(X_1, \cdots, X_n)$ is very close to $\mathbb{E}[f(X_1, \cdots, X_n)]$.

2.2.1 Basic tail concentration bounds

Let $X_1, \cdots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$. By central limit theorem, $\overline{X}_n = \frac{1}{n} \sum_i X_i = \mu + Op(\frac{1}{\sqrt{n}})$. Note that this is a purely asymptotic statement and doesn't tell us about the behaviour for intermediate values of n, say $n = 30$. We would like to know

$$\mathbb{P}(|\overline{X}_n - \mu| \geq t) \ \text{ for some } t > 0$$

We know that

$$\lim_{n \to \infty} \mathbb{P}\big(\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) > t\big) = \mathbb{P}(z \geq t)$$

where $Z \sim \mathcal{N}(0,1)$. Let $\phi(t) = \mathbb{P}(Z \leq t)$, then we have

$$\Big(\frac{1}{t} - \frac{1}{t^3}\Big)\phi(t) \leq 1 - \phi(t) \leq \frac{1}{t}\phi(t) \leq \frac{1}{2}\exp(\frac{-t^2}{2})$$

Following this, we may be tempted to conclude that

$$\mathbb{P}(|\overline{X}_n - \mu| \geq t) \lesssim \exp\left(\frac{-nt^2}{2\theta^2}\right)$$

Although this is good for intuition, this isn't exactly correct. We now look at the finite version of CLT, also known as **Berry Esseen Bound**

Berry Esseen Bound : Let $X_1, \cdots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ then

$$\sup_{x \in \mathbb{R}} \Big|\mathbb{P}\Big(\frac{\sum_i (X_i - \mu)}{\sqrt{n}\sigma} \leq x\Big) - \mathbb{P}(Z \leq x)\Big| \leq C\frac{\gamma}{n}; \;\; \gamma = \frac{\mathbb{E}[|X_i - \mu|^3]}{\sigma^3}, \;\; C \leq \frac{1}{2}$$

Note that you need three moments for this bound.