36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 8: February 13

Lecturer: Siva Balakrishnan *Scribe: Zeyu Tang*

In this lecture, we consider estimation of covariance matrices.

Recall that in the previous lecture, we deal with Wigner type of matrices (whose entries are i.i.d. and with zero mean. Here for covariance matrices (as an example of Wishart type of matrices), their entries are correlated, which makes the estimation potentially harder. In this lecture, we will first consider estimating covariance matrix of Gaussian data; then see what would happen if we drop the Gaussianity, namely, estimating covariance matrix in the sub-Gaussian scenario.

8.1 Covariance matrix estimation: Gaussian data

In this section, we consider the estimation of covariance matrix for Gaussian case. The fact that the data is Gaussian enables us use a trick to reason it just as we did for Wigner type of matrices. (Notice that this trick only works for Gaussian case.)

8.1.1 Problem setup

Consider $x_1, x_2, \ldots, x_n \sim \mathcal{N}(0, \Sigma)$ where each of them $x_i \in \mathbb{R}^d$. A natural estimator for the covariance matrix Σ is the sample covariance (the i-th row of $X \in \mathbb{R}^{n\times d}$ as x_i):

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = \frac{X^T X}{n} \tag{8.1}$$

From Eq. 8.1, the squared largest singular value of $\frac{X}{\sqrt{n}}$ equals to the largest eigenvalue of $\widehat{\Sigma}$:

$$\gamma^2_{\max}(\frac{X}{\sqrt{n}}) = \lambda_{\max}(\widehat{\Sigma}) \tag{8.2}$$

This is always true, but notice that we have not taken advantage of Gaussianity of data. The fact that data is Gaussian allows us to further factor X as the product of a Wigner type matrix W and the square root of Σ:

$$X = W\sqrt{\Sigma} \tag{8.3}$$

where $X, W \in \mathbb{R}^{n\times d}$.

Therefore, for Gaussian data, instead of reasoning about the max/min eigenvalue of the covariance matrix Σ, we can reason about the max/min singular value of $\frac{W\sqrt{\Sigma}}{\sqrt{n}}$.

8.1.2 Bounds for max/min singular values

We want to show that, for $\delta > 0$, we have two facts:

$$P\left(\frac{\gamma_{\max}(X)}{\sqrt{n}} \geq (1+\delta)\left\|\sqrt{\Sigma}\right\|_{\text{op}} + \sqrt{\frac{\text{tr}(\Sigma)}{n}}\right) \leq \exp(-\frac{n\delta^2}{2}) \tag{8.4}$$

$$P\left(\frac{\gamma_{\min}(X)}{\sqrt{n}} \leq (1-\delta)\left\|\sqrt{\Sigma}\right\|_{\text{op}} - \sqrt{\frac{\text{tr}(\Sigma)}{n}}\right) \leq \exp(-\frac{n\delta^2}{2}) \tag{8.5}$$

We will prove the first claim today. The proof of second requires some additional tools which we will see in the future lectures. But before we give out the proof, let's take a look at what the claim Eq. 8.4 is telling us.

Consider the case when Σ is identity matrix, $x_1, x_2, \ldots, x_n \sim \mathcal{N}(0, I_d)$. From the previous two claims, for any $\delta > 0$, with probability at least $1 - 2\exp(-\frac{n\delta^2}{2})$, we have:

$$\frac{\gamma_{\max}(X)}{\sqrt{n}} \leq 1 + \delta + \sqrt{\frac{d}{n}} \tag{8.6}$$

$$\frac{\gamma_{\min}(X)}{\sqrt{n}} \leq 1 - \delta - \sqrt{\frac{d}{n}} \tag{8.7}$$

Therefore, if denote $\epsilon = \delta + \sqrt{\frac{d}{n}}$, recall Eq. 8.2 we have:

$$\lambda_{\max}(\widehat{\Sigma}) \leq (1+\epsilon)^2 \tag{8.8}$$

$$\lambda_{\min}(\widehat{\Sigma}) \geq (1-\epsilon)^2 \tag{8.9}$$

which, if write in terms of operator norm, is a valid bound for the estimation of covariance matrix:

$$\left\|\widehat{\Sigma} - I\right\|_{\text{op}} \leq \epsilon^2 + 2\epsilon \tag{8.10}$$

When Σ is not an identity matrix, $x_1, x_2, \ldots, x_n \sim \mathcal{N}(0, \Sigma)$, follow the same way of reasoning:

$$\begin{aligned}\left\|\frac{X^TX}{n} - \Sigma\right\|_{\text{op}} &= \left\|\sqrt{\Sigma}\left(\frac{W^TW}{n}\right)\sqrt{\Sigma} - \Sigma\right\|_{\text{op}} \\ &= \left\|\sqrt{\Sigma}\left(\frac{W^TW}{n} - I\right)\sqrt{\Sigma}\right\|_{\text{op}} \\ &= \left\|\sqrt{\Sigma}\right\|_{\text{op}}\left\|\frac{W^TW}{n} - I\right\|_{\text{op}}\end{aligned} \tag{8.11}$$

where W has covariance matrix I_d, which is the case we have just considered.

Let's go back to the proof of the upper bound (Eq. 8.4). There are two steps in the proof, one being establish concentration, the other being bounding the expectation. We will ignore for now the bound of expectation (which is a Gaussian width of some set), and see how to establish concentration inequality.

Concentation We want $\frac{\gamma_{\max}(X)}{\sqrt{n}}$ to have some nice concentration property:

$$P\left(\frac{\gamma_{\max}(X)}{\sqrt{n}} \geq \mathrm{E}\left[\frac{\gamma_{\max}(X)}{\sqrt{n}}\right] + t\right) \leq \exp(-\frac{t^2}{2C}) \tag{8.12}$$

The key insight here is that a Lipschitz function of Gaussian concentrates. Considering Eq. 8.3, let $f(W) = \frac{\gamma_{\max}(W\sqrt{\Sigma})}{\sqrt{n}}$ and we want to show $f(\cdot)$ is Lipschitz, i.e., we want to find an L that satisfies:

$$|f(W_1) - f(W_2)| \leq L\left\|W_1 - W_2\right\|_{\text{F}} \tag{8.13}$$

By Weyl's inequality, we can find L as:

$$\begin{aligned}|f(W_1) - f(W_2)| = \left|\frac{\gamma_{\max}(W_1\sqrt{\Sigma})}{\sqrt{n}} - \frac{\gamma_{\max}(W_2\sqrt{\Sigma})}{\sqrt{n}}\right| &\leq \frac{\left\|(W_1 - W_2)\sqrt{\Sigma}\right\|_{\text{op}}}{\sqrt{n}} \\ &\leq \frac{\left\|\Sigma\right\|_{\text{op}}}{\sqrt{n}}\left\|W_1 - W_2\right\|_{\text{F}}\end{aligned} \tag{8.14}$$

Therefore we have the concentration inquality:

$$P\left(\frac{\gamma_{\max}(X)}{\sqrt{n}} \geq \mathrm{E}\left[\frac{\gamma_{\max}(X)}{\sqrt{n}}\right] + t\right) \leq \exp(-\frac{nt^2}{2\left\|\sqrt{\Sigma}\right\|_{\text{op}}^2}) \tag{8.15}$$

8.1.3 Take-away

There are several things worth noticing:

1. Since the data is Gaussian, by showing the function mapping is Lipschitz, we can get concentration "for free". However, this is not the case for non-Gaussian data.
2. What we actually showed is that:

$$P\left(\frac{\left\|\widehat{\Sigma}-\Sigma\right\|_{\text{op}}}{\|\Sigma\|_{\text{op}}} \geq 2\big(\sqrt{\frac{d}{n}}+\delta\big)+\big(\sqrt{\frac{d}{n}}+\delta\big)^2\right) \leq 2\exp(-\frac{n\delta^2}{2}), \quad \forall \delta > 0 \tag{8.16}$$

   We can see that only when $\sqrt{\frac{d}{n}} \to 0$ can we consistently estimate Σ.

8.2 Covariance matrix estimation: sub-Gaussian data

In this section, we consider the estimation of covariance matrix for sub-Gaussian case.

8.2.1 Problem setup & sketch of proof

Consider $x_1, x_2, \ldots, x_n \sim P_X(\Sigma)$ where each of them $x_i \in \mathbb{R}^d$ satisfing σ sub-Gaussianity:

$$\mathrm{E}\left[\exp(tu^T x_i)\right] \leq \exp(\frac{t^2\sigma^2}{2}), \quad \forall \|u\|_2 = 1, u \in \mathbb{R}^d \tag{8.17}$$

We want to show that

$$P\left(\frac{\left\|\widehat{\Sigma}-\Sigma\right\|_{\text{op}}}{\sigma^2} \geq C\big(\sqrt{\frac{d}{n}}+\frac{d}{n}\big)+\delta\right) \leq 2\exp(-c\, n \min\{\delta^2, \delta\}), \quad \forall \delta > 0 \tag{8.18}$$

where C, c, and σ^2 are constants to be determined.

Again, we can make use the result in Wigner case and go through a two-step process to further correct the result:

1. Discretization: for fixed unit vector u, make use of the result in Wigner case (we need to correct the result since the entries are not i.i.d.)

$$P\left(\left\|\widehat{\Sigma}-\Sigma\right\|_{\text{op}} \geq t\right) \leq |5|^d P\left(u^T(\widehat{\Sigma}-\Sigma)u \geq \frac{t}{4}\right) \tag{8.19}$$

2. Bound the moment generating function: we want to establish an upper bound for

$$\mathrm{E}\left[\exp\{tu^T(\widehat{\Sigma}-\Sigma)u\}\right] \tag{8.20}$$

Consider the $u^T\widehat{\Sigma}u$ term in the expression of the mgf function Eq. 8.20:

$$u^T\widehat{\Sigma}u = \frac{u^TX^TXu}{n}, \text{ where } Xu = \begin{bmatrix} u^Tx_1 \\ u^Tx_2 \\ \vdots \\ u^Tx_n \end{bmatrix} \tag{8.21}$$

Therefore the expression of mgf function is ($Z_i = u^Tx_i$ is σ sub-Gaussian):

$$\begin{aligned} \mathrm{E}\left[\exp\{tu^T(\widehat{\Sigma}-\Sigma)u\}\right] &= \mathrm{E}\left[\exp\left\{\frac{t}{n}\Big(\sum_{i=1}^n(u^Tx_i)^2 - \mathrm{E}\big[(u^Tx)^2\big]\Big)\right\}\right] \\ &= \mathrm{E}\left[\exp\left\{\frac{t}{n}\Big(\sum_{i=1}^n Z_i^2 - \mathrm{E}\big[Z^2\big]\Big)\right\}\right] \end{aligned} \tag{8.22}$$

In order to bound this mgf function, we need to make use of the (σ^4, σ^2) sub-Exponential tail bound of Z_i^2.

8.2.2 sub-Gaussian & sub-Exponential

We first give the definition of sub-Exponential variable as well as their tail bounds, and then we look at the example of bounded variables.

We say a random variable Z is (σ^2, b) sub-Exponential if

$$\mathrm{E}[\exp\{tZ\}] \le \exp\{\frac{t^2\sigma^2}{2}\}, \text{ for } |t| \le \frac{1}{b} \tag{8.23}$$

which is saying that the variable has a sub-Gaussian type of bound for small t.

One useful fact would be: the square of σ sub-Gaussian variable is $c\cdot(\sigma^4, \sigma^2)$ sub-Exponential variable (c is some constant).

One canonical example is bounded variable (recall Bernstein's inequality). If Z has mean μ and bounded support $[0, b]$, with variance σ^2, we have

$$P(|Z-\mu| \ge t) \le 2\exp\Big\{-\frac{t^2}{2(\sigma^2+bt)}\Big\} \tag{8.24}$$

Therefore

$$P(|Z-\mu| \ge t) \le \begin{cases} 2\exp\big\{-\frac{t^2}{2\sigma^2}\big\} & \text{for } 0 \le t \le \frac{\sigma^2}{b} \\ 2\exp\big\{-\frac{t}{2b}\big\} & \text{for } t \ge \frac{\sigma^2}{b} \end{cases} \tag{8.25}$$