MATH888: High-dimensional probability and statistics	Fall 2021

Lecture 20 — October 22, 2021

Sebastien Roch, UW-Madison *Scribe: Shuqi Yu*

1 Overview

In the last lecture, we showed the bound on approximate isometries. In this lecture, we focus on estimating the covariance matrix from finite sample, which has a well-known application in the principal components analysis (PCA). And the material follows from section 4.7, 5.2, 9.2 in [3].

2 Covariance estimation

Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with mean zero, the covariance matrix is $\mathbf{\Sigma} := \mathbb{E}\mathbf{X}\mathbf{X}^\top$, *i.e* $\mathbf{\Sigma}_{ij} = \mathbb{E}\mathbf{X}_i\mathbf{X}_j$. We first show the following property of covariance matrix. For $\mathbf{u} \in \mathbb{S}^{d-1}$, $\mathbb{E}\langle \mathbf{X}, \mathbf{u}\rangle = 0$ as $\mathbf{X}$ is mean zero, then

$$Var\langle \mathbf{X}, \mathbf{u}\rangle = \mathbb{E}\left(\langle \mathbf{X}, \mathbf{u}\rangle^2\right) = \mathbb{E}\mathbf{u}^\top\mathbf{X}\mathbf{X}^\top\mathbf{u} = \mathbf{u}^\top\mathbb{E}(\mathbf{X}\mathbf{X}^\top)\mathbf{u} = \mathbf{u}^\top\mathbf{\Sigma}\mathbf{u} \tag{1}$$

Suppose there are n *i.i.d* samples $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(n)} \sim \mathbf{X}$, the sample covariance is defined as

$$\hat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}^{(i)}\mathbf{X}^{(i)\top}$$

Put differently, let $\mathbf{A} \in \mathbb{R}^{n\times d}$ have row i is $\mathbf{X}^{(i)}$, then $\hat{\mathbf{\Sigma}}_n = \frac{\mathbf{A}^\top\mathbf{A}}{n}$.

2.1 Corvariance estimation for high dimensional distributions

Recall the following theorem we proved last time for sub-gaussian isotropic random vectors.

Theorem 1 (Theorem 4.6.1 in [3])**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ have rows that are independent, mean zero, sub-gaussian isotropic random vectors with $K = \max_{i=1,\cdots n}\|\mathbf{A}_{i\cdot}\|_{\psi_2}$, then with probability $1 - 2e^{-u}$,*

$$\left\|\frac{1}{n}\mathbf{A}^\top\mathbf{A} - \mathbf{I}_d\right\|_2 \le CK^2\left(\sqrt{\frac{d+u}{n}} + \frac{d+u}{n}\right)$$

Remark 2. *Since rows of $\mathbf{A}$ are isotropic whose covariance are $\mathbf{I}_d$, $\left\|\frac{1}{n}\mathbf{A}^\top\mathbf{A} - \mathbf{I}_d\right\|_2 = \left\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\|_2$.*

Next, we extend to the general sub-gaussian setting.

Theorem 3 (Theorem 4.7.1 in [3])**.** *Let* $\mathbf{X}$ *be a sub-gaussian random vector in* $\mathbb{R}^d$ *with invertible covariance matrix* $\mathbf{\Sigma}$ s.t

$$\|\langle \mathbf{X}, \mathbf{u}\rangle\|_{\psi_2} \leq K\sqrt{\mathbb{E}(\langle \mathbf{X}, \mathbf{u}\rangle^2)} \qquad \textit{for all } \mathbf{u} \in \mathbb{S}^{d-1}$$

with probability $1 - 2e^{-u}$*, then*

$$\left\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\|_2 \leq CK^2 \|\mathbf{\Sigma}\|_2 \left(\sqrt{\frac{d+u}{n}} + \frac{d+u}{n}\right).$$

Remark 4. $\mathbb{E}(\langle \mathbf{X}, \mathbf{u}\rangle^2) = \langle \mathbf{u}, \mathbf{\Sigma u}\rangle$ *by* (1)*.*

Proof. We first bring the random vectors $\mathbf{X}_1, \cdots, \mathbf{X}_n$ to the isotropic position. By slides(`https://people.math.wisc.edu/~roch/hdps/roch-hdps-slides12.pdf`) or Exercise 3.2.2 in [3], there exist isotropic mean zero vectors $\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(n)}$ s.t

$$\mathbf{X}^{(i)} = \mathbf{\Sigma}^{1/2}\mathbf{Z}^{(i)}, \text{ for all } i = 1, \cdots, n.$$

We proved previously, $\left\|\mathbf{Z}^{(i)}\right\|_{\psi_2} \leq K$. Let $\hat{\mathbf{R}}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{Z}^{(i)}\mathbf{Z}^{(i)\top} - \mathbf{I}_d$, then

$$\left\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\|_2 = \left\|\mathbf{\Sigma}^{1/2}\hat{\mathbf{R}}_n\mathbf{\Sigma}^{1/2}\right\|_2 \leq \left\|\hat{\mathbf{R}}_n\right\|_2 \|\mathbf{\Sigma}\|_2$$

where we apply the property $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$ and $\left\|\mathbf{A}^{1/2}\right\|_2^2 = \left(\sqrt{\sigma_{max}(A)}\right)^2 = \|\mathbf{A}\|_2$.

Consider the $n \times d$ random matrix $\mathbf{A}$ whose rows are $\mathbf{Z}^{(i)}$, then

$$\frac{1}{n}\mathbf{A}^\top\mathbf{A} - \mathbf{I}_d = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}^{(i)}\mathbf{Z}^{(i)\top} - \mathbf{I}_d = \hat{\mathbf{R}}_n$$

Then we conclude by applying Theorem 1 for $\mathbf{A}$. □

2.2 Covariance estimation for lower-dimensional distributions

We found that the covariance matrix $\mathbf{\Sigma}$ of an n-dimensional distribution can be estimated from $m = O(n)$ sample points for sub-gaussian distributions. For approximately lower-dimensional distributions, smaller sample can be sufficient for covariance estimation, which means that the distribution tends to concentrate near a small subspace.

Theorem 5 (Theorem 9.2.4 in [3])**.** *Let* $\mathbf{X}$ *be a sub-gaussian random vector in* $\mathbb{R}^d$ *with invertible covariance matrix* $\mathbf{\Sigma}$ s.t

$$\|\langle \mathbf{X}, \mathbf{u}\rangle\|_{\psi_2} \leq K\sqrt{\mathbb{E}(\langle \mathbf{X}, \mathbf{u}\rangle^2)} \qquad \textit{for all } \mathbf{u} \in \mathbb{S}^{d-1}$$

with probability $1 - 2e^{-u}$*, then*

$$\left\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\|_2 \leq CK^4 \|\mathbf{\Sigma}\|_2 \left(\sqrt{\frac{r+u}{n}} + \frac{r+u}{n}\right).$$

where $r = tr(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_2$.

2.3 General covariance estimation

Next, we state a more general version of covariance estimation by applying matrix Berstein inequality, see the slides(`https://people.math.wisc.edu/~roch/hdps/roch-hdps-slides20.pdf`) for more details.

Theorem 6 (Theorem 5.6.1 in [3])**.** *Let* $\mathbf{X}$ *be a random vector in* $\mathbb{R}^d$*,* $d \geq 2$*. Assume that for some* $K \geq 1$*,*

$$\|\mathbf{X}\|_2 \leq K \left(\mathbb{E}\,\|\mathbf{X}\|_2^2\right)^{1/2} \qquad \textit{almost surely.}$$

Then, with probability $1 - 2e^{-u}$*, we have*

$$\left\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\|_2 \leq C\,\|\mathbf{\Sigma}\|_2 \left(\sqrt{\frac{K^2 r(\log d + u)}{n}} + \frac{K^2 r(\log d + u)}{n}\right)$$

where $r = tr(\mathbf{\Sigma}) / \|\mathbf{\Sigma}\|_2 \leq n$*.*

References

[1] Rick Durrett, *Probability—theory and examples (fifth edition)*, Cambridge University Press, 2019.

[2] Ramon van Handel, *APC 550: Probability in High Dimension*, Lecture Notes, 2016. `https://web.math.princeton.edu/~rvan/APC550.pdf`

[3] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, Cambridge University Press, 2018.