MATH888: High-dimensional probability and statistics Fall 2021

Lecture 21 — October 25, 2021

Sebastien Roch, UW-Madison *Scribe: Max B, Sebastien Roch*

1 Overview

In the last lecture, we discussed estimation of the covariance matrix of both subgaussian and more general random vectors. In this lecture, we discuss an application to principal component analysis (PCA). This lecture is based on Section 8.2 from Wainwright [1].

2 Application to Principal Component Analysis

2.1 General Setting and Goal

Recall the setting of covariance estimation from the previous lecture: For a mean-zero random vector $\mathbf{X} \in \mathbb{R}^d$, the covariance matrix is $\mathbf{\Sigma} := \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]$, so that $\mathbf{\Sigma}_{ij} = \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_j\right]$.

PCA is a dimension-reduction technique. $\mathbf{X}$ may be a high-dimensional, but its action might be mostly concentrated in lower dimensional space. Hence we are interested in estimating the direction $\mathbf{u} \in \mathbb{S}^{d-1}$ that maximizes the variance of the scalar projection $\langle \mathbf{X}, \mathbf{u} \rangle$. Define

$$\mathbf{u}_1 \in \mathrm{argmax}_{\mathbf{u}\in\mathbb{S}^{d-1}} \mathrm{Var}\left[\langle \mathbf{X}, \mathbf{u} \rangle\right].$$

Since $\mathbf{X}$ is mean-zero, it follows that $\langle \mathbf{X}, \mathbf{u} \rangle$ is also mean-zero. Therefore $\mathrm{Var}\left[\langle \mathbf{X}, \mathbf{u} \rangle\right] = \mathbb{E}\left[\langle \mathbf{X}, \mathbf{u} \rangle^2\right]$. Moreover, at the beginning of the previous lecture we showed that $\mathbb{E}\left[\langle \mathbf{X}, \mathbf{u} \rangle^2\right] = \langle \mathbf{u}, \mathbf{\Sigma}\mathbf{u} \rangle$. Therefore

$$\mathbf{u}_1 \in \mathrm{argmax}_{\mathbf{u}\in\mathbb{S}^{d-1}} \langle \mathbf{u}, \mathbf{\Sigma}\mathbf{u} \rangle.$$

Therefore by variational calculus (See, e.g. the "Review of eigenvalues" section in Lecture 16), the solution $\mathbf{u}_1$ is the maximal eigenvector of $\mathbf{\Sigma}$ (i.e. corresponding to the largest eigenvalue of $\mathbf{\Sigma}$,)

Remark: Having obtained the first principal component $\mathbf{u}_1$, we could then go on to obtain $\mathbf{u}_2 \in \mathrm{argmax}_{\mathbf{u}\in\mathbb{S}^{d-1}\cap\mathrm{span}\{\mathbf{u}_1\}^\perp} \mathrm{Var}(\langle \mathbf{X}, \mathbf{u} \rangle)$ and so forth... but we won't do that today.)

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

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

Our goal is to estimate $\mathbf{u}_1$ with these samples. To this end, define

$$\hat{\mathbf{u}} \in \mathrm{argmax}_{\mathbf{u}\in\mathbb{S}^{d-1}} \left\langle \mathbf{u}, \hat{\mathbf{\Sigma}}\mathbf{u} \right\rangle$$

We wish to know how close are $\mathbf{u}_1$ and $\hat{\mathbf{u}}$. In particular, we wish to estimate $\|\mathbf{u}_1 - \hat{\mathbf{u}}_1\|_2$, the Euclidean distance between $\mathbf{u}_1$ and $\hat{\mathbf{u}}$. Equipped with the theorems from the previous lecture, we have estimates of $\left\|\hat{\mathbf{\Sigma}}_{\mathbf{n}} - \mathbf{\Sigma}\right\|$; moreover, $\hat{\mathbf{\Sigma}}$ can be seen as a perturbation $\mathbf{\Sigma}$, so we shall utilize perturbation theorems as well.

We consider the following toy model.

2.2 Spiked Covariance Model

Definition 1 (Spiked covariance model)**.** *Let $\boldsymbol{W} \in \mathbb{R}^d$ be an isotropic, subgaussian random vector with mean zero, and let ϵ be an independent real-valued subgaussian random variable with mean zero and variance 1.*[1] *The* ***spiked covariance model*** *is given by the random vector $\mathbf{X}$ with distribution*

$$\mathbf{X} \sim \mathbf{W} + \sqrt{\nu}\epsilon\theta^*$$

where $\nu > 0$, $\theta^* \in \mathbb{S}^{d-1}$ are fixed.

The idea here is that θ^* is a "secret direction", so that we are adding variation in the θ^* direction. The constant ν can be thought of as the "strength" of the hidden signal.

Since $\mathbf{X}$ is mean-zero and isotropic, its covariance is easily computed as

$$\mathbf{\Sigma} = I_d + \nu\theta^*\theta^{*\top}$$

where I_d is the $d \times d$ identity matrix. The θ^* direction maximizes variance. Indeed, $\theta^* = \mathbf{u}_1$ has eigenvalue $1+\nu$ since $\mathbf{\Sigma}\theta^* = (1+\nu)\theta^*$ whereas if $\langle \mathbf{z}, \mathbf{u}_1 \rangle = 0$ then z is an eigenvector with eigenvalue 1, since $\mathbf{\Sigma z} = \mathbf{z}$. Hence the eigengap is $(1+\nu) - 1 = \nu$.

We make the further assumption that the subgaussian norms are all 1. Call this assumption $(*)$.

Theorem 2 (Cor 8.7 in [1])**.** *Assume $n > d$. Given n iid samples from the spiked covariance model with $(*)$, and assuming that $\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}} \le C_0$, it holds that if $\hat{\theta}$ is the maximal eigenvector of $\hat{\mathbf{\Sigma}}_n$, then $\left\|\hat{\theta} - \theta^*\right\|_2 \le C_1\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}}$ with probability $1 - C_2\exp\{-C_3 d\}$.*

Remarks:

1. The assumption that $\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}} \le C_0$ means that $\frac{d}{n}$ cannot be too large.
2. The constants C_1, C_2, and C_3 are universal, i.e. they do not depend on n or d.

2.3 A Perturbation Bound

Consider the **perturbation matrix** $\mathbf{P} := \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}$. Consider the real $d \times d$ orthonormal matrix U whose columns are the eigenvectors of $\mathbf{\Sigma}$. Then $U = (\theta^* | U_2)$ where U_2 is a $d \times (d-1)$ matrix whose columns are an orthonormal basis of span $\{\theta^*\}^\perp$. Define the **transformed perturbation matrix**

$$\tilde{\mathbf{P}} := U^\top \mathbf{P} U = \begin{pmatrix} \tilde{p}_{11} & \tilde{\mathbf{p}}^\top \\ \tilde{\mathbf{p}} & \tilde{\mathbf{P}}_{22} \end{pmatrix}$$

[1] Note [1] uses ξ rather than ϵ

where $\tilde{p}_{11} \in \mathbb{R}$, $\tilde{\mathbf{p}} \in \mathbb{R}^{d-1}$, and $\tilde{\mathbf{P}}_{22} \in \mathbb{R}^{(d-1)\times(d-1)}$.

We will use the following bound version of Davis-Kahan (see Lecture 16) to prove Theorem 2.

Theorem 3 (Theorem 8.5 in [1]). *If* $\|\mathbf{P}\|_2 < \frac{\nu}{2}$ *then*

$$\left\|\hat{\theta} - \theta^*\right\|_2 \leq \frac{2\|\tilde{\mathbf{p}}\|_2}{\nu - 2\|\mathbf{P}\|_2}$$

Remark: This theorem is similar to the more basic Davis-Kahan result discussed in Lecture 16 (Theorem 4.5.5 in Vershynin [2]). To see the similarity, note that since $\|\theta^*\|_2 = 1$ and $\left\|U_2^\top\right\|_2 = 1$,

$$\|\tilde{\mathbf{p}}\|_2 = \left\|U_2^\top \mathbf{P}\theta^*\right\|_2 \leq \left\|U_2^\top\right\|_2 \|\mathbf{P}\|_2 \|\theta^*\|_2 = \|\mathbf{P}\|_2 = \left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_2$$

so that by also using the inequality $\|\mathbf{P}\|_2 < \nu/2$, the right hand side of Theorm 3 can be bounded above by $C\left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_2$.

2.4 Preview of Proof of Theorem 2

In order to gainfully employ Theorem 3 in order to prove Theorem 2, it is necessary to bound both $\|\mathbf{P}\|_2$ and $\|\tilde{\mathbf{p}}\|_2$. Since $X^{(i)} = \mathbf{w}_i + \sqrt{\nu}\epsilon_i\theta^*$ for $i = 1, \ldots, n$, therefore by (1)

$$\hat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n \left[\mathbf{w}_i + \sqrt{\nu}\epsilon_i\theta^*\right]\left[\mathbf{w}_i + \sqrt{\nu}\epsilon_i\theta^*\right]^\top.$$

Using this, the perturbation matrix may be decomposed in the following way:

$$\begin{aligned}\mathbf{P} &= \hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma} \\ &= \nu\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - 1\right)\theta^*\theta^{*\top} + \sqrt{\nu}\left(\bar{\mathbf{w}}\theta^{*\top} + \theta^*\bar{\mathbf{w}}^\top\right) + \left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i^\top - I_d\right)\end{aligned}$$

where $\bar{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^n \epsilon_i\mathbf{w}_i$. We then deal with each of the terms individually, which is the subject of the next lecture.

References

[1] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* CUP.

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