IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: P Rigollet
Scribe: Maggie Makar, Ziao Lin

Lecture 11
Mar. 12, 2020

Goals: In this lecture we will focus on *Principal component analysis* (PCA), where the task is to project a high dimensional vector X onto a low dimensional space. At the crux of PCA is studying Σ, the covariance matrix of X.

We assume that the true Σ follows a spiked covariance model. We consider the empirical estimator $\hat{\Sigma}$, and quantify how close it is to the true Σ in terms of Σ's eigenspace and dimension as well as number of samples.Our analysis will rely on the Davis-Kahan theorem from the previous 2 lectures.

1. SPIKED COVARIANCE MODEL

Consider the following problem. Suppose we observe some data $X_i, \ldots, X_n \sim \mathcal{N}_d(0, \Sigma)$. We want to consider some model that allows us to uncover a low dimensional space in which X lies (e.g., for visualization purposes). Specifically, we will consider a linear structure where we take a vector $v \in R^d$. The expectation of the observed matrix $X = [X_1, X_2, ..., X_n]^\top \in R^{n\times d}$ would be represented as $E[X] = Yv$, where $Y = [Y_1, Y_2, ..., Y_n]^\top \in R^{n\times 1}$ and $y_i \in R$.

Realistically, we would not observe perfectly aligned points. Instead, data is typically corrupted by some noise in the full d dimension. We denote the noise by Z and assume that $Z_1, \cdots, Z_n \sim \mathcal{N}(0, I_d)$, with $Z \perp\!\!\!\perp Y$. So we can represent the observed $X_i = Y_i v + Z_i$. Because Y_i, and Z_i might not be on the same scale, we introduce a tuning parameter $\sqrt{\theta}$ for some $\theta > 0$, and we say that $X_i = \sqrt{\theta} Y_i v + Z_i$. We also assume that v has been normalized, i.e. $|v|_2 = 1$. Since $Z \perp\!\!\!\perp Y$, we have that $X \sim \mathcal{N}(0, \Sigma)$ based on a linear transformation of a multivariate random vector also has a multivariate normal distribution, with

$$
\begin{aligned}
\Sigma &= \mathbb{E}[X_i X_i^\top] \\
&= \mathbb{E}[(\sqrt{\theta} Y_i v + Z_i)(\sqrt{\theta} Y_i v + Z_i)^\top] \\
&= \theta \mathbb{E}[Y_i^2] v v^\top + \mathbb{E}[Z_i Z_i^\top] \\
&= \theta v v^\top + I_d
\end{aligned}
$$

where the last equality follows from the fact that $\mathbb{E}[Y_i^2] = 1$, and $\mathbb{E}[Z_i Z_i^\top] = I_d$. When $|v|_2$ is fixed to be $= 1$, this model is referred to as the *spiked covariance model.* Under the spiked covariance model, we can claim the following:

Claim: v is an eigenvector of Σ.

This is because $\Sigma v = \theta(v^\top v)v + I_d v = (1 + \theta)v$. We also have that:

$$
\begin{aligned}
&\max_{|u|_2=1} u^\top \Sigma u \\
&= \theta(u^\top v)^2 + 1 \\
&= v^\top v,
\end{aligned}
$$

where the last equality follows from the fact that this quantity is maximized when u, and v are aligned. Knowing that $\forall u \perp v$, $u^\top \Sigma u = 1 < 1 + \theta$. This identifies all our eignvalues:

$$\lambda_1 = (1+\theta) \geq \lambda_2 = 1 \geq \lambda_3 = 1 \ldots \lambda_d = 1$$

2. ESTIMATING Σ

We will take the empirical covariance estimate,

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

to be an estimator for Σ. By LLN, we have that this is a consistent estimator. We know that the largest eigenvector is v and the associated eigenvalue is λ_1. So if we want to identify what v is, we can apply Davis-Kahan:

$$|\sin(\angle(\hat{v}, v))| \leq \frac{2||\hat{\Sigma} - \Sigma||_{op}}{\lambda_1 - \lambda_2} = \frac{2||\hat{\Sigma} - \Sigma||_{op}}{\theta}$$

where $\hat{v}$ is the leading eigenvector of $\hat{\Sigma}$. This tells us that the norm we need to control in order to do PCA is the operator norm. Note that even if $\hat{\Sigma}$ and Σ is positive semidefinite since they are real symmetric matrices, the difference $E = \hat{\Sigma} - \Sigma$ in general is not guaranteed to be positive semidefinite. Thus we cannot directly apply the leading eigenvector u_1 into $u_1^T E u_1$ to get operator norm. We will instead move on to control this operator norm using ε-Nets.

Let $E := ||\hat{\Sigma} - \Sigma||_{op}$. We have that:

$$E_{jk} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(j)} X_i^{(k)} - \mathbb{E}[X_i^{(j)} X_i^{(k)}]$$

Using the definition of the operator norm (see lecture 8, expression 3.30) and a previous result (see lecture 9, proof of Lemma), we have that:

$$||E||_{op} \leq 2 \max_{x \in \mathcal{N}_d, y \in \mathcal{N}_d} x^\top (\hat{\Sigma} - \Sigma) y,$$

where $\mathcal{N}_d$ is the $\frac{1}{4}$-net of $B_2(\mathbb{R}^d)$, and we can control $|\mathcal{N}_d|$ to get $|\mathcal{N}_d| \leq 9^d$. We have that:

$$x^\top (\hat{\Sigma} - \Sigma) y = \frac{1}{n}\sum_{i=1}^{n} (x^\top X_i)(y^\top X_i) - \mathbb{E}[(x^\top X_i)(y^\top X_i)]. \tag{2.1}$$

It turns out that the distribution of this variable is subexponential. To see that note that:

$$x^\top X_i \sim \mathcal{N}(0, x^\top \Sigma x).$$

If we take $|x|_2 \leq 1$, we have that $x^\top \Sigma x \leq ||\Sigma||_{op}$, we have that

$$x^\top X_i \sim subG(||\Sigma||_{op}).$$

Since the term 2.1 includes a product of 2 subGaussian variables, it is subExponential, which means that we will likely use Brenstien's inequality. To use Brenstien's inequality:

$$\begin{aligned}
&||(x^\top X_i)(y^\top X_i) - \mathbb{E}[(x^\top X_i)(y^\top X_i)]||_{\varphi_1} \\
&\leq ||(x^\top X_i)(y^\top X_i)||_{\varphi_1} + ||\mathbb{E}[(x^\top X_i)(y^\top X_i)]||_{\varphi_1} \\
&\leq ||(x^\top X_i)||_{\varphi_2}||(y^\top X_i)||_{\varphi_2} + ||(x^\top X_i)||_{\varphi_2}||(y^\top X_i)||_{\varphi_2} \\
&\leq 2\sqrt{||\Sigma||_{op}}\sqrt{||\Sigma||_{op}} \\
&\leq 2||\Sigma||_{op},
\end{aligned}$$

where the first inequality follows from triangle inequality, and the second inequality is an application of Jensen's inequality due to the convexity of φ_1-norm plus the inequality s.t. $||xy||_{\varphi_1} \leq ||x||_{\varphi_2}||y||_{\varphi_2}$. The third inequality follows from the property of subGaussian variables. We can now apply Bernstein:

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^{n}(x^\top X_i)(y^\top X_i) - \mathbb{E}[(x^\top X_i)(y^\top X_i)] > t\bigg) \\
&\leq \sum_{x,y} \exp\bigg(-Cn\Big(\frac{t^2}{||\Sigma||_{op}^2} \wedge \frac{t}{||\Sigma||_{op}}\Big)\bigg) \\
&\leq 9^{2d} \exp\bigg(-Cn\Big(\frac{t^2}{||\Sigma||_{op}^2} \wedge \frac{t}{||\Sigma||_{op}}\Big)\bigg),
\end{aligned}$$

for some constant C. And the second inequality follows from the fact that the terms in the sum do not depend on x, y.
Now let's denote the desired threshold to be δ, then resolve the above inequality we will get: $t \leq C||\Sigma||_{op}[\sqrt{\frac{d+lg(1/\delta)}{n}} + \frac{d+lg(1/\delta)}{n}]$ for some constant C. Then we can hopefully control $||E||_{op} \leq C||\Sigma||_{op}\sqrt{\frac{d}{n}}$ for some constant C. Plug in the results back to Davis-Kahan, we eventually get a bound on the difference in angle between the two leading eigenvenctors $\hat{v}$ and v:

$$|\sin(\angle(\hat{v}, v))| \leq C\frac{1+\theta}{\theta}\sqrt{\frac{d}{n}}$$

for some constant C.
The result can be generalized to multiple spiked model with some scaling factor proportional to the square root of number of spikes.

3. SPARSE PCA

A slightly different model that could have generated Σ is known as the *sparse spiked model.* In this model v is assumed to be sparse. Consider the example where $v \in \mathbb{R}^2$. The spiked covariance model assumes that v_1, v_2 are a linear combination of possibly all the dimensions in the original space. Instead, the sparse spiked covariance matrix assumes that v_1, v_2 are a linear combination of a small subset of cardinality s contribute to the principle directions v_1, v_2. In that case, we would want to include a sparsity constraint when estimating $\hat{v}_1, \hat{v}_2$. The estimator becomes:

$$\hat{v} = \max_{|u|_2=1, u \in B_0(s)} u^\top \hat{\Sigma} u.$$

Because we're considering B_0 in the constraint, this problem is computationally very expensive. Significant research has been done to find efficient ways to solve this problem (e.g., convex relaxations, ScoTLASS)

Summary: By applying Davis-Kahan theorem, we derive a upper bound on the difference in angle between the two leading eigenvectors in sample covariance estimator $\hat{\Sigma}$ and the truth covariance matrix Σ in *Principal component analysis* (PCA). The results and methods used here are generalizable to multiple spiked model.