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

Lecture 23: April 16

Lecturer: Alessandro Rinaldo *Scribes: Lorenzo Tomaselli*

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

23.1 Applications of Davis–Kahan Theorem

In this lecture we look at some applications of the Davis–Kahan theorem. We first provide a bound on the distance between the estimated and the actual leading eigenvectors of the covariance matrix in the spiked covariance model. Then, we bound the number of misclustered nodes in the task of community recovery in the settings of the Stochastic Block Models.

23.1.1 Spiked Covariance Model

Let $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} (0, \Sigma)$ in $\mathbb{R}^d$ and let the covariance matrix, Σ, be a spiked version of I_d:

$$\Sigma = \theta \nu \nu^\top + I_d$$

for $\theta > 0$ and $\nu \in \mathbb{S}^{d-1}$. From the above definition of Σ, it follows that its leading eigenvalue, λ_1, is equal to $1 + \theta$, with corresponding leading eigenvector ν, while the other eigenvalue is 1 with algebraic multiplicity $d - 1$. The main question that we address in this section is: are we able to estimate the leading eigenvector of Σ, i.e. ν, in this simple theoretical model?

For this purpose, let $\hat{\nu}$ be the leading eigenvector of the sample covariance matrix $\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$. We want to estimate how close is $\hat{\nu}$ to ν and we proceed as follows:

$$\min_{\varepsilon \in \{-1, 1\}} \|\varepsilon \hat{\nu} - \nu\|^2 = 2 - 2 \left|\hat{\nu}^\top \nu\right| \overset{\text{(I)}}{\leq} 2 - 2 \left|\hat{\nu}^\top \nu\right|^2 \overset{\text{(II)}}{=} 2 \sin^2 (\angle(\hat{\nu}, \nu))$$

where for inequality (I) we used Cauchy-Schwartz to bound $\left|\hat{\nu}^\top \nu\right| \leq 1$ and for equality (II) we used the fact that $\sin^2 (\angle(\hat{\nu}, \nu)) + \cos^2 (\angle(\hat{\nu}, \nu)) = 1$ and $\angle(\hat{\nu}, \nu) = \cos^{-1} \left(\left|\hat{\nu}^\top \nu\right|\right)$. We, hence, use Davis–Kahan theorem as in [YWS14] to bound $\sin^2 (\angle(\hat{\nu}, \nu))$ and we get:

$$\min_{\varepsilon \in \{-1, 1\}} \|\varepsilon \hat{\nu} - \nu\|^2 \leq 8 \frac{\|\hat{\Sigma}_n - \Sigma\|_{op}^2}{\theta^2} \tag{23.1}$$

where θ is the eigengap.

By making some additional assumptions on the X_i's, we can provide a bound for $\|\hat{\Sigma}_n - \Sigma\|_{op}$ in (23.1). For example, let $X_i \in SG_d (\|\Sigma\|_{op})$ for all $i \in \{1, 2, \ldots, n\}$ where, in our settings, $\|\Sigma\|_{op} = 1 + \theta$. Then, with

probability at least $1-\delta$, with $\delta \in (0,1)$, we have:

$$\min_{\varepsilon \in \{-1,1\}} \|\varepsilon\,\hat{\nu} - \nu\| \leq C\,\frac{1+\theta}{\theta}\max\left\{\sqrt{\frac{d+\log\left(\frac{1}{\delta}\right)}{n}},\ \frac{d+\log\left(\frac{1}{\delta}\right)}{n}\right\} \tag{23.2}$$

In Eq.(23.2) we note that as θ, i.e. the eigengap, gets smaller then it is hard to distinguish the leading eigenvector from the other ones and the bound explodes. Therefore, in order to provide a bound on the distance between the estimated and the leading eigenvector of the spiked covariance matrix, we need to estimate $\|\hat{\Sigma}_n - \Sigma\|_{op}$ and the eigengap needs to be bounded away from zero.

From Eq.(23.2) we see that, when θ does not change with n, in order for the bound to go to zero for large n, we need that $d = o(n)$. In high dimensions, instead, where we allow d to grow and θ to decay with n, we note that, in order for the bound to go to zero, $\sqrt{\frac{d+\log\left(\frac{1}{\delta}\right)}{n}}$ should go faster to zero than $\frac{1}{\theta}$ to infinity.

23.1.2 Community Recovery in Stochastic Block Model

As a reference for the following application see *sec. 4.5* of [RV18]. Suppose we are dealing with a Stochastic Block Model with k communities and n nodes, where $k = 2$. The actual community assignment is unknown to us and we want to recover it from the adjacency matrix A, whose entries are defined as follows. For all $i \neq j$, where $i, j \in \{1, 2, \ldots, n\}$, and $0 < q < p < 1$:

$$A_{ij} = \begin{cases} \text{Bernoulli}\,(p) & i,\ j \text{ are in the same community} \\ \text{Bernoulli}\,(q) & \text{otherwise} \end{cases}$$

and $A_{ii} = 0$ for all i. Moreover, once we condition on a particular random partition of the nodes in the community, the matrix A is a collection of independent Bernoulli random variables.

Task: to recover the community assignments. As we will see, when the probabilities p and q are close to each other it is hard to solve this problem. We note that:

$$\mathbb{E}\,[A] = P = \Theta B \Theta^\top - \text{diag}\left(\Theta B \Theta^\top\right)$$

where:

$$B = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$$

where the entry B_{ij} represents the probability of an agent in community i to form an edge with an agent belonging to community j.

We have matrix $\Theta \in \{0,1\}^{n \times k}$, where each row has only 1 as non zero entry in the position corresponding to the index of the community to which node i belongs, i.e.:

$$\Theta_{ik} = \begin{cases} 1 & \text{if node } i \text{ is in community } k \\ 0 & \text{otherwise} \end{cases}$$

Example: let $n = 4$ and the community assignments be $\{1, 2\}$ and $\{3, 4\}$. Then:

$$\Theta B \Theta^\top = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right]$$

Example: Let n be an even number, $k = 2$ and let $\left\{1, 2, \ldots, \frac{n}{2}\right\}$ and $\left\{\frac{n}{2}+1, \frac{n}{2}+2, \ldots, n\right\}$ be the community assignments. It can be shown that $rank(P) = k = 2$, and that the leading eigenvalue of matrix P, with the corresponding leading eigenvector, are:

$$\lambda_1 = \left(\frac{p+q}{2}\right) n\,, \qquad \nu_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \frac{1}{\sqrt{n}}$$

While the second leading eigenvalue of matrix P and its corresponding eigenvector are equal to:

$$\lambda_2 = \left(\frac{p-q}{2}\right) n\,, \qquad \nu_2 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix} \frac{1}{\sqrt{n}} \tag{23.3}$$

where in the second leading eigenvector in (23.3) the first $\frac{n}{2}$ entries have positive sign while the last $\frac{n}{2}$ have negative sign. We note that the sign of the entries of the second leading eigenvector exactly correspond to the community assignments.

Rule: Target the second leading eigenvector of matrix P and cluster the nodes on the basis of the sign of the corresponding index.

In order to compute a bound on the number of mistakes that we make by clustering the nodes according to the above rule, we proceed as follows. Let the adjacency matrix $A = P + E$, where matrix E is a noise matrix. We have:

$$\mathbb{E}\left[A\right] = P$$

Moreover, from a result in a previous homework, with high probability, we have:

$$\left\|E\right\|_{op} \lesssim \sqrt{n}$$

Also:

$$\left\|P\right\|_{op} = \left(\frac{p+q}{2}\right) n - p \asymp \left(\frac{p+q}{2}\right) n$$

We note that the signal contained in P is stronger than the noise because the operator norm of P is of the order n while the operator norm of E is of the order $\sqrt{n}$. This means that when n is large enough we may use matrix A, instead of P, to recover community assignments.

By Davis–Kahan theorem, targeting the 2^{nd} leading eigenvector of matrix A, i.e. $\hat{\nu}_2\left(A\right)$, we get:

$$\min_{\varepsilon \in \{-1,\,1\}} \left\|\varepsilon\, \hat{\nu}_2\left(A\right) - \nu_2\left(P\right)\right\| \lesssim \frac{\left\|E\right\|_{op}}{\left(\frac{p-q}{2}\right) n} \leq \frac{C}{\left(\frac{p-q}{2}\right)\sqrt{n}}$$

where $C > 0$ is an universal constant and $\left(\frac{p-q}{2}\right) n$ is the eigengap. It follows that, with high probability:

$$n \sum_{i=1}^{n} \left(\varepsilon\, \hat{\nu}_2\left(A\right)_i - \nu_2\left(P\right)_i\right)^2 \leq \frac{C^{\,2}}{\left(\frac{p-q}{2}\right)^2}$$

where $\hat{\nu}_2(A)_i$ and $\nu_2(P)_i$ are the ith entry of the second leading eigenvector of, respectively, matrix A and matrix P. Since $\sqrt{n}\,\nu_2(P) \in \{-1, 1\}^n$, if $\operatorname{sign}(\varepsilon\,\hat{\nu}_2(A)_i) \neq \operatorname{sign}(\nu_2(P)_i)$, that means we are misclustering node i, then we have:

$$n\,(\varepsilon\,\hat{\nu}_2(A)_i - \nu_2(P)_i)^2 \geq 1$$

So that, with high probability, the bound on the number of misclustered nodes is:

$$|\{\, i \in \{1, 2, \ldots, n\} \text{ s.t. } \operatorname{sign}(\varepsilon\,\hat{\nu}_2(A)_i) \neq \operatorname{sign}(\nu_2(P)_i)\}| \leq \frac{C^2}{\left(\frac{p-q}{2}\right)^2}$$

The above result suggests that, in order to recover the communities in a consistent manner, i.e. the number of mistakes vanishes as n increases, we need:

$$\frac{C^2}{\left(\frac{p-q}{2}\right)^2}\frac{1}{n} \xrightarrow{n\to\infty} 0$$

Therefore, in order for the above condition to hold, we need $\frac{p-q}{2} \gg \frac{1}{\sqrt{n}}$. In other words, if the probabilities p and q are too close to each other then it is not possible to produce an estimator that consistently recovers the communities.

In [EBW17] the authors provide a bound for the quantity:

$$\min_{\varepsilon\in\{-1,\,1\}} \|\varepsilon\,\hat{\nu} - \nu\|_\infty$$

that leads to sharper results in the Stochastic Block Model.

23.2 Sparse PCA

We now turn to the topic of Sparse PCA. Assume the following sparse spiked covariance model:

$$\Sigma = \theta\nu\nu^\top + I_d \tag{23.4}$$

where $\theta > 0$, $\nu \in \mathbb{S}^{d-1}$ and $\|\nu\|_0 = k \leq \frac{d}{2}$. We want to estimate ν with $\hat{\nu}$, the latter being defined as follows:

$$\hat{\nu}^\top\hat{\Sigma}_n\hat{\nu} = \max_{\substack{u\in\mathbb{S}^{d-1}\\ \|u\|_0=k'}} u^\top\hat{\Sigma}_n u \tag{23.5}$$

where we know that $k \leq k' \leq \frac{d}{2}$. In order to be solved, the optimization problem in (23.5) requires a combinatorial search over all possible sparse leading eigenvectors of $\hat{\Sigma}_n$. In practice, this is infeasible due to high computational costs. Nonetheless, we may establish the following bound.

Theorem 23.1 *Let $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} (0, \Sigma)$, where $X_i \in SG_d(\|\Sigma\|_{op})$ for all $i = 1, 2, \ldots, n$ where, in the settings of (23.4), $\|\Sigma\|_{op} = 1 + \theta$. Let $\hat{\nu}$ be defined as in (23.5). Then, with probability at least $1 - \delta$, $[\,\delta \in (0, 1)\,]$:*

$$\min_{\varepsilon\in\{-1,\,1\}} \|\varepsilon\,\hat{\nu} - \nu\| \leq C\frac{1+\theta}{\theta}\max\left\{\sqrt{\frac{(k+k')\log\left(\frac{ed}{k+k'}\right) + \log\left(\frac{1}{\delta}\right)}{n}},\ \frac{(k+k')\log\left(\frac{ed}{k+k'}\right) + \log\left(\frac{1}{\delta}\right)}{n}\right\} \tag{23.6}$$

By comparing this result with (23.2), we note that the term $(k+k')\log\left(\frac{ed}{k+k'}\right)$ is present instead of the term d. The rate in (23.6) is better when, as we are assuming, $(k+k') \leq d$.

We will prove the result (23.6) in the next lecture. In the meanwhile, here some important facts that we are going to use in the proof.

Fact:

$$\begin{aligned}
\theta \sin^2\left(\angle\left(\nu, \hat{\nu}\right)\right) &= \nu^\top \Sigma \nu - \hat{\nu}^\top \Sigma \hat{\nu} \\
&= \nu^\top \hat{\Sigma}_n \nu - \hat{\nu}^\top \Sigma \hat{\nu} - \nu^\top \left(\hat{\Sigma}_n - \Sigma\right) \nu \\
&\overset{(i)}{\leq} \hat{\nu}^\top \left(\hat{\Sigma}_n - \Sigma\right) \hat{\nu} - \nu^\top \left(\hat{\Sigma}_n - \Sigma\right) \nu \\
&= \left\langle\, \hat{\Sigma}_n - \Sigma,\, \hat{\nu}\hat{\nu}^\top - \nu\nu^\top \,\right\rangle
\end{aligned}$$

where for inequality (i) we used (23.5). Note that for any two matrices A, B of the same order, we have:

$$\left\langle\, A,\, B \,\right\rangle = \operatorname{tr}\left(A^\top B\right)$$

Let $S \subseteq \{1, 2, \ldots, d\}$ be such that $S = \operatorname{supp}(\nu) \bigcup \operatorname{supp}(\hat{\nu})$. Then $|S| \leq k + k'$. Let Σ_S be the submatrix of Σ with dimensions $|S| \times |S|$, with rows and columns corresponding to the indexes in S. Similarly we define $\hat{\Sigma}_S$ as submatrix of $\hat{\Sigma}_n$ and the vectors ν_S and $\hat{\nu}_s$, both belonging to $\mathbb{R}^{|S|}$. Then:

$$\theta \sin^2\left(\angle\left(\nu, \hat{\nu}\right)\right) \leq \left\langle\, \hat{\Sigma}_S - \Sigma_S,\, \hat{\nu}_S \hat{\nu}_S^\top - \nu_S \nu_S^\top \,\right\rangle$$

References

[EBW17] J. Eldridge, M. Belkin and Y. Wang "Unperturbed: spectral analysis beyond Davis–Kahan", arXiv: 1706.06516 [stat. ML], 2017

[RV18] R. Vershynin, "High-Dimensional Probability. An Introduction with Applications in Data Science", *Cambridge University Press*, 2018.

[YWS14] Y. Yu, T. Wang and R.,J. Samworth, "A useful variant of the Davis–Kahan theorem for statisticians", *Biometrika*, 2015, Vol. 102(2), pp. 315–323.