EE290 Mathematics of Data Science Lecture 9-10 - 09/26/2019

Lecture 9-10: Spectral Method for Exact Recovery of SSBM

Lecturer: Jiantao Jiao *Scribe: Sara Fridovich-Keil and Han Feng*

In this lecture, we introduce the spectral method for exact recovery of the communities in the symmetric stochastic block model. Before [2], it was believed that the spectral method might require an additional cleaning step.

1 Notation

We introduce and use the following objects:

1. $SSBM(n, 2, \frac{a \log n}{n}, \frac{b \log n}{n})$ is a random graph on n vertices with 2 communities (of either exactly equal or expected equal size), where in-community edges exist with probability $p = \frac{a \log n}{n}$ and between-community edges exist with probability $q = \frac{b \log n}{n}$. By assumption, $p > q$.
2. $\tilde{A}$ is the adjacency matrix of the graph, with the first $n/2$ indices corresponding to the first community and the latter $n/2$ indices corresponding to the second community.
3. A' is $\tilde{A}$ with self-loops added with probability p (i.e. $A' = \tilde{A} + \text{diag}(\mathsf{B}(p))$). Note that:
$$\mathbb{E}A' = n\frac{p+q}{2}\bar{\phi_1}\bar{\phi_1}^\top + n\frac{p-q}{2}\bar{\phi_2}\bar{\phi_2}^\top.$$
4. $\bar{\phi_1} = \frac{1}{\sqrt{n}}\mathbf{1}$ is the average degree, which carries no community information.
5. $\bar{\phi_2} = \frac{1}{\sqrt{n}}(1 \ldots -1 \ldots)^\top$ captures all of the available community information. Later, we will remove $\bar{\phi_1}$ and simplify notation to use $\bar{\phi} = \bar{\phi_2}$.
6. $A = A' - n\frac{p+q}{2}\bar{\phi_1}\bar{\phi_1}^\top$, constructed to retain only the community-relevant information. Note that:
$$\mathbb{E}A = n\frac{p-q}{2}\bar{\phi_2}\bar{\phi_2}^\top = \bar{\lambda}\bar{\phi}\bar{\phi}^\top,$$
where $\bar{\lambda} = \frac{a-b}{2}\log n$, the largest (and only nonzero) eigenvalue of $\mathbb{E}A$.
7. In general, $\bar{\cdot}$ denotes the expected value on the population, and plain $\cdot$ denotes the value on the sample we observe. For instance, $\bar{A} = \mathbb{E}A$ and $\bar{\lambda} = \mathbb{E}\lambda$.
8. Unless otherwise specified, $\|\cdot\|$ denotes the operator norm (which for a symmetric matrix equals its largest absolute eigenvalue).

2 Spectral Method for exact Recovery of SSBM

Theorem 1. *$SSBM(n, 2, \frac{a \log n}{n}, \frac{b \log n}{n})$ efficiently solvable if $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$ using spectral method.*

We will show it for the case $a > b$.

Algorithm 2. *Solve the largest eigenpair (λ, ϕ) such that $A\phi = \lambda\phi$. Return $X_{spec}(i) = 1$, if $\phi(i) \geq 0$ and 2 otherwise.*

We start by considering some relevant results.

Lemma 3.

$$\mathbb{P}\left(\left\|A-\bar{A}\right\| \geq c_1\sqrt{\log n}\right) \leq c_2 n^{-3}. \tag{1}$$

Note that c_1, c_2 depend on a, b. We also have a more general lemma [3, Theorem 9]:

Lemma 4. *A symmetric, zero diagonal matrix, $\{A_{ij}, i<j\}$ independent, [0,1] valued. Assume: $\mathbb{E}A_{ij} \leq p$, $\frac{c_0 \log n}{n} \leq p \leq 1-c_1$. Then, for any $c>0$, $\exists c'>0$, such that*

$$\mathbb{P}\left(\|A-\mathbb{E}A\| \leq c'\sqrt{np}\right) \geq 1-n^{-c}.$$

Note that this lemma does not follow from Latala's result, because that requires the row and column second moments to be $\leq k^2 n$ and the fourth moment to be $\leq k^4 n^2$ in order to get $\|A-\bar{A}\| \leq k\sqrt{n}$, so it would give us a bound with $p^{1/4}$ instead of $p^{1/2}$.

From Davis-Kahan, $|\langle \phi, \bar{\phi}\rangle| = 1-o(1)$. Since $|\bar{\phi}_i| = \frac{1}{\sqrt{n}}$, $|\phi_i - \bar{\phi}_i| < \frac{1}{\sqrt{n}}$ for each i would be a sufficient condition for exact recovery (ie it would guarantee that ϕ would have all correct signs). Unfortunately, this condition provably doesn't hold. The problem is that this condition is two-sided, whereas actually ϕ_i can be arbitrarily large in magnitude (with the same sign as $\bar{\phi}_i$) and the spectral method will still recover correctly (so the condition is unnecessarily strong).

Instead, we'll compare ϕ with $A\bar{\phi}/\bar{\lambda}$, instead of $\bar{\phi} = \bar{A}\bar{\phi}/\bar{\lambda}$. The intuition here is that ϕ is an eigenvector of A, not $\bar{A}$. We will prove Theorem 1, following [1, §4] and [2].

We start by introducing the key lemma, Lemma 5, showing how it proves Theorem 1.

Lemma 5. *$\exists$ constant $C(a,b)$ such that as $n \to \infty$,*

$$\mathbb{P}\left(\min_{s\in\pm 1}\left\|s\phi - A\bar{\phi}/\bar{\lambda}\right\|_\infty \leq \frac{C}{\sqrt{n}\log\log n}\right) \geq 1-\frac{C}{n^2}.$$

Assuming Lemma 5, we can prove that the spectral method succeeds in exact recovery with high probability. Define two "good" events

$$\mathcal{E}_1 = \left\{\min_{i\in[1:n/2]}(A\bar{\phi}/\bar{\lambda})_i \geq \frac{2\epsilon}{(a-b)\sqrt{n}}, \max_{i\in[n/2+1:n]}(A\bar{\phi}/\bar{\lambda})_i \leq \frac{-2\epsilon}{(a-b)\sqrt{n}}\right\},$$
$$\mathcal{E}_2 = \left\{\min_{s\in\pm 1}\|s\phi - A\bar{\phi}/\bar{\lambda}\|_\infty \leq \frac{c}{\sqrt{n}\log\log n}\right\}.$$

Intuitively, $\mathcal{E}_1$ captures the event that the signs of $A\bar{\phi}/\bar{\lambda}$ are correct, with sufficient margin, and $\mathcal{E}_2$ captures the event that ϕ is sufficiently close to $A\bar{\phi}/\bar{\lambda}$. Then, $\mathcal{E}_1 \cap \mathcal{E}_2$ captures the event that the signs of ϕ are correct, ie that the spectral method succeeds in exact recovery, so we just need that $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \to 1$, which holds because $(A\bar{\phi})_i \sim \frac{1}{\sqrt{n}}\left(\mathsf{B}(n/2,p) - \mathsf{B}(n/2,q)\right)$ for $i \in [1:n/2]$ and ([2, Lemma 8])

$$\mathbb{P}\left(\mathsf{B}(n/2,p) - \mathsf{B}(n/2,q) \leq \epsilon\log n\right) \leq n^{-\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2 + \epsilon\log(a/b)/2},$$

and we apply the union bound.

The remainder of the lecture is devoted to proving Lemma 5.

We start by choosing the sign of ϕ so that $\phi^\top\bar{\phi} \geq 0$.

$$\begin{aligned}\|\phi - A\bar{\phi}/\bar{\lambda}\|_\infty &\leq \|\phi - A\phi/\bar{\lambda}\|_\infty + \|A\phi/\bar{\lambda} - A\bar{\phi}/\bar{\lambda}\|_\infty \\ &= \|\phi - \lambda\phi/\bar{\lambda}\|_\infty + \|(A/\bar{\lambda})(\phi-\bar{\phi})\|_\infty \\ &= \frac{|\lambda-\bar{\lambda}|}{\bar{\lambda}}\|\phi\|_\infty + \frac{1}{\bar{\lambda}}\|A(\phi-\bar{\phi})\|_\infty.\end{aligned}$$

By Weyl's inequality [1], $|\lambda - \bar{\lambda}| \lesssim \|A - \bar{A}\| \lesssim \sqrt{\log n}$, so we will now also condition on $\|A - \bar{A}\| \lesssim \sqrt{\log n}$, which gives us:

$$\|\phi - A\bar{\phi}/\bar{\lambda}\|_\infty \leq \frac{\|\phi\|_\infty}{\sqrt{\log n}} + \frac{1}{\bar{\lambda}}\|A(\phi - \bar{\phi})\|_\infty.$$

We'll continue by bounding the second term $\|A(\phi - \bar{\phi})\|_\infty$ using a leave one out analysis, which decomposes the dependence structure. Define n auxiliary matrices

$$A^{(m)}_{ij} = A_{ij}\delta_{i\neq m, j\neq m}, \text{ for } m = 1, \dots n,$$

so $A^{(m)}$ is just A with the mth row and column replaced with zero. Let $\phi^{(m)}$ be the leading eigenvector of $A^{(m)}$ with $\phi^{(m)\top}\bar{\phi} \geq 0$. Let A_m be the m-th row of A. We have

$$(A(\phi - \bar{\phi}))_m = A_m(\phi - \bar{\phi}) = A_m(\phi - \phi^{(m)}) + A_m(\phi^{(m)} - \bar{\phi}).$$

2.1 Analysis of $A_m(\phi - \phi^{(m)})$

We can bound the first term as follows:

$$\begin{aligned} A_m(\phi - \phi^{(m)}) &\leq ||A_m|| \cdot ||\phi^{(m)} - \phi|| \\ &\leq ||A||_{2\to\infty}||\phi^{(m)} - \phi|| \\ &\lesssim \sqrt{\log n}||\phi||_\infty. \end{aligned} \tag{2}$$

The above steps require some justification, provided below. First we used

$$\|A\|_{2\to\infty} = \sup_x \frac{\|Ax\|_\infty}{\|x\|_2} = \max_m \|A_m\|_2 \leq \|A\|_2,$$

because $\|Ax\|_\infty \leq \|Ax\|_2$. This further leads to the justification for (2) below

$$\begin{aligned} \|A\|_{2\to\infty} &\leq \left\|A - \bar{A}\right\|_{2\to\infty} + \left\|\bar{A}\right\|_{2\to\infty} \\ &\lesssim \left\|A - \bar{A}\right\|_2 + \log n/\sqrt{n} \\ &\lesssim \sqrt{\log n} + \log n/\sqrt{n}, \end{aligned} \tag{3}$$

where we used the concentration of A. In order to show that $||\phi - \phi^{(m)}|| \lesssim ||\phi||_\infty$, we invoke the stronger version of Davis-Kahan.

Theorem 6. *(Davis-Kahan) Let u, v be maximal eigenvectors of real symmetric matrices A, B. Then:*

$$\min_{s\in\pm 1} \|su - v\| \lesssim \frac{\|(A - B)u\|}{\lambda_1(A) - \lambda_2(B)}$$

The theorem implies

$$\left\|\phi - \phi^{(m)}\right\|_2 \leq \frac{\left\|(A - A^{(m)})\phi\right\|_2}{\lambda_1(A) - \lambda_2(A^{(m)})}.$$

The denominator can be bounded with Weyl's inequality, which requires the following estimate

$$\begin{aligned} ||A - A^{(m)}|| &\leq ||A - A^{(m)}||_F \\ &\leq \sqrt{2}||A||_{2\to\infty} && (A^{(m)} \text{ differs from } A \text{ by one row and one column}) \\ &\lesssim \sqrt{\log n} && \text{From (3)} \end{aligned}$$

[1] $|\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|_2$.

This combined with $\left\|A-\bar{A}\right\| \lesssim \sqrt{\log n}$ implies $\left\|A^{(m)}-\bar{A}\right\| \lesssim \sqrt{\log n}$. Hence $\lambda_2(A^{(m)}) \sim \sqrt{\log n}$, which is dominated by $\lambda_1(A) \sim \log n$ (which again from the concentration of A and Weyl's inequality). To bound the numerator in Davis-Kahan, we compute

$$\left(\left(A-A^{(m)}\right)\phi\right)_i = \begin{cases} \lambda\phi_m, & \text{if } i=m \\ A_{im}\phi_m, & \text{if } i\neq m. \end{cases}$$

As a result

$$\begin{aligned} \left\|(A^{(m)}-A)\phi\right\| &= \sqrt{\lambda^2|\phi_m|^2+\sum_{i\neq m}A_{im}^2\phi_m^2} \\ &\leq |\phi_m|\sqrt{\lambda^2+||A||_{2\to\infty}^2} \\ &\lesssim \log n|\phi_m|. \end{aligned}$$

where the last inequality comes from the concentration of A and the bound on λ. This completes the justification of (2).

2.2 Analysis of the term $A_m(\phi^{(m)}-\bar{\phi})$

We need a lemma from [2, Lemma 7] that captures the concentration of Bernoulli random variables with an unbalanced weight w.

Lemma 7. *Let $w\in\mathbb{R}^n$ be fixed and $X_i\sim B(p_i)$ independent. Assume that $\max_i p_i\leq p$. Given $\alpha>0$,*

$$\mathbb{P}\left(|\sum_{i=1}^n w_iX_i-\mathbb{E}X_i| \geq \frac{(2+\alpha)pn}{1\vee\log\left(\frac{\sqrt{n}\|w\|_\infty}{\|w\|_2}\right)}\|w\|_\infty\right) \leq 2e^{-\alpha np}.$$

Setting $w=\phi^{(m)}-\bar{\phi}$, $X_i=A'_{mi}$, $p=(a\vee b)\frac{\log n}{n}$, $\alpha=3(a\vee b)$. [2] Conditioning on w:

$$\mathbb{P}\left(\left|\sum_{i=1}^n(A_{mi}-\bar{A}_{mi})(\phi^{(m)}-\bar{\phi})_i\right| \lesssim \frac{\log n\left\|\phi^{(m)}-\bar{\phi}\right\|_\infty}{1\vee\log\frac{\sqrt{n}\left\|\phi^{(m)}-\bar{\phi}\right\|_\infty}{\left\|\phi^{(m)}-\bar{\phi}\right\|_2}}\right) > 1-\frac{2}{n^3}.$$

Define the good event

$$\mathcal{E}_1^{(m)} := \left\{\left|A_m(\phi^{(m)}-\bar{\phi})\right| \lesssim \underbrace{\frac{\log n}{\sqrt{n}}\left(\left\|\phi^{(m)}-\bar{\phi}\right\|_2+\frac{\sqrt{n}\left\|\phi^{(m)}-\bar{\phi}\right\|_\infty}{1\vee\log\frac{\sqrt{n}\left\|\phi^{(m)}-\bar{\phi}\right\|_\infty}{\left\|\phi^{(m)}-\bar{\phi}\right\|_2}}\right)}_{T_3}\right\}$$

Since

$$\left|\sum_{i=1}^n\bar{A}_{mi}(\phi^{(m)}-\bar{\phi})_i\right| \leq \left\|\bar{A}\right\|_{2\to\infty}\left\|\phi^{(m)}-\bar{\phi}\right\|_2 \lesssim \frac{\log n}{\sqrt{n}}\left\|\phi^{(m)}-\bar{\phi}\right\|_2,$$

[2] Be careful that A'_{mi} is different from A, but $A'_{mi}-\mathbb{E}A'_{mi}=A_{mi}-\mathbb{E}A_{mi}$.

We have

$$\mathbb{P}(\mathcal{E}_1^{(m)}) \geq 1 - \frac{2}{n^3},$$

and as a result

$$\mathbb{P}\left(\cap_{m=1}^n \mathcal{E}_1^{(m)}\right) \geq 1 - \frac{2}{n^2}.$$

We further bound the term T_3. Applying Davis-Kahan to $A^{(m)} - \bar{A}$ (whose norm has been shown to be the order $\sqrt{\log n}$), we have $\left\|\phi^{(m)} - \bar{\phi}\right\|_2 \lesssim 1/\sqrt{\log n}$, and

$$\begin{aligned} T_3 &\lesssim \log n \frac{\left\|\phi^{(m)} - \bar{\phi}\right\|_\infty \vee \sqrt{\frac{1}{n}}}{\log(\sqrt{\log n})} \\ &\lesssim \frac{\log n \left\|\phi\right\|_\infty}{\log\log n}. \end{aligned}$$

The first inequality above considers the following two cases: when $\left\|\phi^{(m)} - \bar{\phi}\right\|_\infty < 1/\sqrt{n}$ the second term dominates T_3; when $\left\|\phi^{(m)} - \bar{\phi}\right\|_\infty \geq 1/\sqrt{n}$, the first term dominates T_3. The next inequality comes from

$$\begin{aligned} \left\|\phi^{(m)} - \bar{\phi}\right\|_\infty &\leq \left\|\phi^{(m)} - \phi\right\|_\infty + \left\|\phi - \bar{\phi}\right\|_\infty \\ &\leq \left\|\phi^{(m)} - \phi\right\|_2 + \left\|\phi\right\|_\infty + \left\|\bar{\phi}\right\|_\infty \\ &\lesssim \left\|\phi\right\|_\infty \end{aligned}$$

where we used the previous bound in (2) and the fact that the unit vector ϕ satisfies $\left\|\phi\right\|_\infty \geq 1/\sqrt{n} = \left\|\bar{\phi}\right\|$.

This completes the proof that $A_m(\phi^{(m)} - \bar{\phi}) \lesssim \frac{\log n \|\phi\|_\infty}{\log\log n}$ with high probability.

2.3 Bounding $\left\|\phi\right\|_\infty$

Summarizing the bounds above

$$\begin{aligned} \left\|\phi - A\bar{\phi}/\bar{\lambda}\right\|_\infty &\lesssim \frac{||\phi||_\infty}{\sqrt{\log n}} + \frac{1}{\bar{\lambda}}||A(\phi - \bar{\phi})||_\infty \\ &\lesssim \frac{||\phi||_\infty}{\sqrt{\log n}} + \frac{1}{\log n}\left(\sqrt{\log n}\left\|\phi\right\|_\infty + \frac{\log n \left\|\phi\right\|_\infty}{\log\log n}\right) \\ &\lesssim \frac{\left\|\phi\right\|_\infty}{\log\log n}. \end{aligned} \tag{4}$$

It remains to bound the L_∞ norm of ϕ. We claim $\left\|\phi\right\|_\infty \lesssim \frac{1}{\sqrt{n}}$:

$$\begin{aligned} \left\|\phi\right\|_\infty &\leq \left\|\phi - A\bar{\phi}/\bar{\phi}\right\|_\infty + \left\|A\bar{\phi}/\bar{\lambda}\right\|_\infty \\ &\lesssim \left\|A\bar{\phi}\right\|_\infty/\bar{\lambda}, \end{aligned}$$

where we used (4). It suffices to show that $\left\|A\bar{\phi}\right\|_\infty \lesssim \log n/\sqrt{n}$. This comes from direct calculation.

$$\left\|A_m\bar{\phi}\right\| \leq \frac{1}{\sqrt{n}}\sum_{i=1}^m |A_{mi}|,$$

where

$$|A_{mi}| = |A'_{mi} - \frac{a+b}{2n}\log n| \le |A_{mi}|^2 + \frac{a+b}{2n}\log n.$$

Hence,

$$\begin{aligned}\left\|A_m\bar{\phi}\right\| &\le \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{m}|A_{mi}|^2\right) + \frac{1}{\sqrt{n}}\frac{a+b}{2}\log n \\ &\le \frac{1}{\sqrt{n}}\|A\|_{2\to\infty}^2 + \frac{1}{\sqrt{n}}\frac{a+b}{2}\log n \\ &\lesssim \frac{1}{\sqrt{n}}\log n.\end{aligned}$$

References

[1] Emmanuel Abbe. Community Detection and Stochastic Block Models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, June 2018.

[2] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise Eigenvector Analysis of Random Matrices with Low Expected Rank. *arXiv:1709.09565*, September 2017.

[3] B. Hajek, Y. Wu, and J. Xu. Achieving Exact Cluster Recovery Threshold via Semidefinite Programming: Extensions. *IEEE Transactions on Information Theory*, 62(10):5918–5937, October 2016.