36-710 Advanced Statistical Theory		Fall 2018
	Lecture 21: Nov 12	
Lecturer: Alessandro Rinaldo		*Scribes: Wanshan Li*

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

21.1 Stochastic Block Model

Suppose $A = [A_{ij}] \in \mathbb{R}^{n\times n}$ is the adjacency matrix of a random graph. In this lecture we only consider undirected graph with no self-edges, so A is symmetric. A common way to model A is to assume that A_{ij} are independent Bernoulli variables for $i < j$, i.e.,

$$A_{ij} \sim \text{Bernoulli}(p_{ij}), A_{ji} = A_{ij},\ p_{ij} \in (0,1),\ i < j,\ A_{ii} = 0.$$

When $p_{ij} = p$ for all $i < j$, this degenerates to the well-known Erdös-Rényi model, which could be too simple to be practical. A practical generalization is the Stochastic Block Model (SBM). SBM assumes that there is a symmetric matrix $B \in \mathbb{R}^{k\times k}$, for $k \ll n$, and a map $C : \{1, \cdots, n\} \to \{1, \cdots, k\}$, such that

$$p_{ij} = B_{C(i),C(j)}.$$

By definition, there are $k(k+1)/2$ parameters to estimate in SBM. In SBM, n nodes are grouped in k groups, with group labels given by map C, and we call these groups *communities*.

Example 21.1. *Suppose $B = (p - q)I_k + q\mathbf{1}_k\mathbf{1}_k^\top$, with $p \in (0,1)$, $q \in (0,p)$ and $\mathbf{1}_k = [1,\cdots,1]^\top \in \mathbb{R}^k$. When $k = 3$, this gives*

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

In words, $p_{ij} = p$ if $C(i) = C(j)$, and q otherwise.

Usually we denote $P = \mathbb{E}[A]$, so

$$P = [p_{ij}]_{i,j=1,\cdots,n}.$$

Define a matrix $\Theta \in \mathbb{R}^{n\times k}$ by

$$\Theta_{i,j} = \begin{cases} 1,\ C(i) = j \text{ (i.e., node } i \text{ is in community } j), \\ 0,\ \text{otherwise.} \end{cases}$$

Then matrix $P = \mathbb{E}[A]$ can be expressed as

$$P = \Theta B \Theta^\top - \text{diag}(\Theta B \Theta^\top). \tag{21.1}$$

Remark

- One should notice that each row of Θ has only 1 non-zero entries, and the number of non-zero entries in column j of Θ is the number of nodes in community j.
- We do not know Θ based on the observation A, because we do not know the underlying map C giving community labels, and our goal is to estimate the map C, or the partition of nodes into communities. Estimation of C or Θ is called *community detection*.
- *We assume that we know* k when doing community detection, though it's not the case in practice. In practice people will use some methods to choose k.

Example 21.2. *Suppose $n = 2m$, $k = 2$, and communities are $\{1, \cdots, m\}$, $\{m+1, \cdots, 2m\}$. For $0 < q < p < 1$, assume that*

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

Then

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

So $\operatorname{rank}(\Theta B \Theta^\top) = 2$, *with top-2 eigenvalues and eigenvectors*

$$\lambda_1 = \left(\frac{p+q}{2}\right) \cdot n, \ U_1 = \frac{1}{\sqrt{n}}[1, \cdots, 1] \in \mathbb{R}^n$$

and

$$\lambda_2 = \left(\frac{p-q}{2}\right) \cdot n, \ U_2 = \frac{1}{\sqrt{n}}[1, \cdots, 1, -1, \cdots, -1] \in \mathbb{R}^n.$$

From now on let's consider community detection under the setting of Example 21.2. Think of A as

$$A = P + E.$$

Since diagonal entries of P are zero, we have

$$\|P\| \sim \left(\frac{p+q}{2}\right)^n.$$

Also, $\|E\| \sim \sqrt{n}$ with high probability. To estimate Θ, we introduce the following algorithm of spectral clustering.

Spectral clustering algorithm

1 Compute the second eigenvector U_2 of A.

2 Cluster nodes based on the sign of the entries of U_2.

By Davis-Khan theorem, we can show that this algorithm works well. The eigengap in Davis-Khan theorem is

$$\delta = \min\{\lambda_2, \lambda_1 - \lambda_2\} = \min\{q, \frac{p-q}{2}\} \times n \triangleq \mu n.$$

Then, by Davis-Khan theorem,

$$\min_{\varepsilon \in \{1,-1\}} \|\varepsilon U_2(A) - U_2(P)\| \leq \frac{2^{3/2}\|E\|}{\mu n}.$$

Since $\|E\| \lesssim \sqrt{n}$, we know that with high probability,

$$\min_{\varepsilon \in \{1,-1\}} \|\varepsilon U_2(A) - U_2(P)\| \lesssim \frac{C}{\mu \sqrt{n}},$$

which is equivalent to

$$\min_{\varepsilon \in \{1,-1\}} \|\varepsilon \sqrt{n} U_2(A) - \sqrt{n} U_2(P)\| \lesssim \frac{C}{\mu}$$

with high probability. Since $\sqrt{n}U_2(P) \in \{1, -1\}^n$, if $\text{sign}(\varepsilon U_2(A)_i) \neq \text{sign}(\varepsilon U_2(P)_i)$, where $U_2(A)_i$ is the i-th entry of A, then

$$n\left(\varepsilon U_2(A)_i - \varepsilon U_2(P)_i\right)^2 \geq 1.$$

Therefore,

$$\#\{i \in \{1, \cdots, n\} : \text{sign}(\varepsilon U_2(A)_i) \neq \text{sign}(\varepsilon U_2(P)_i)\} \leq \frac{C^2}{\mu^2}.$$

Thus, if $\frac{C^2}{\mu^2} \cdot \frac{1}{n} \to 0$ as $n \to \infty$, then the spectral clustering algorithm is correct over all nodes except for a vanishing fraction. In conclusion, the condition we need is

$$\frac{C^2}{\mu^2} \cdot \frac{1}{n} \to 0 \Leftrightarrow \min\{q, \frac{p-q}{2}\} \gg \frac{1}{\sqrt{n}}.$$

Remark

- It's reasonable that successful community detection requires $\frac{p-q}{2} \gg \frac{1}{\sqrt{n}}$, since a larger $\frac{p-q}{2}$ implies a larger signal-to-noise ratio.
- However, the condition $q \gg \frac{1}{\sqrt{n}}$ seems strange, since $q = 0$ will lead to two dis-connected communities, and make it trivial to do community detection. The reason why q cannot vanish is that the method, spectral clustering on the second eigenvector, is too restricted. But this is not a big problem in practice, because usually people only use community detection method on a connected graph, otherwise it would be more reasonable to take different components as different (groups of) communities.
- In the case $q = 0$, if we consider clustering two eigenvectors together, we can still use spectral clustering to do community detection well. In fact,

  $$\lambda_1 = \lambda_2 = pm, \lambda_3 = \cdots = \lambda_n = 0,\ U_1 = [\mathbf{1}, \mathbf{0}]', U_2 = [\mathbf{0}, \mathbf{1}]',$$

  where $\mathbf{1} = [1, \cdots, 1]$ and $\mathbf{0} = [0, \cdots, 0]$. Therefore, the rows of $U = [U_1, U_2]$ only take values of $(0, 1)$ and $(1, 0)$. Similar to what we do above, we can prove that rows of $U(A) = [U_1(A), U_2(A)]$ concentrate around two centroids when p is not too small. There are many general discussions of spectral clustering, for example, [LR2015].

21.2 Uniform Law of Large Numbers

Suppose $\{X_1, \cdots, X_n\} \overset{i.i.d.}{\sim} P$ are real-valued random variables. Let $F_X(x) = \mathbb{P}(X_1 \leq x)$ denote the cumulative distribution function. We can estimate $F_X(x)$ by

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1} X_i \leq x \sim \frac{1}{n}\text{Binomial}(n, F_X(x)).$$

Using, e.g., Hoeffding's inequality, one can show that for any fixed $x \in \mathbb{R}$,

$$\hat{F}_n(x) \xrightarrow{P} F_X(x).$$

The next question would be, what if we want to construct an estimator $\hat{F}_n(x)$, such that

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_X(x)| \xrightarrow{P} 0?$$

To answer this question and more general ones, we first introduce a more general framework. Let P be a probability distribution over some space $(\mathcal{X}, \mathcal{B})$, and $\mathcal{F}$ be a collection of real-valued functions on $\mathcal{X}$. Suppose $\{X_1, \cdots, X_n\} \overset{i.i.d.}{\sim} P$. Based on this sample we can construct P_n, called empirical measure, as a random probability measure on $(\mathcal{X}, \mathcal{B})$, by

$$P_n : A \in \mathcal{X} \mapsto P_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \in A}.$$

Then for any $f \in \mathcal{F}$, let

$$\mathbb{P}f = \mathbb{E}_{X\sim P}[f(X)] = \int_{\mathcal{X}} f(x)\mathrm{d}P(x),$$
$$\mathbb{P}_n f = \mathbb{E}_{X\sim P_n}[f(X)] = \int_{\mathcal{X}} f(x)\mathrm{d}P_n(x),$$

We are interested in the behaviour of

$$\begin{aligned}\|P - P_n\|_{\mathcal{F}} &= \sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_n f| \\ &= \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_i)])\right|.\end{aligned}$$

As an example, can we prove or disprove that $\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_X(x)| \xrightarrow{P} 0$? The class of results like $\|P - P_n\|_{\mathcal{F}} \xrightarrow{P} 0$ are called uniform law of large numbers, and we will discuss it in details next time.

References

[LR2015] Lei, Jing and Rinaldo, Alessandro. Consistency of spectral clustering in stochastic block models. Ann. Statist. 43 (2015), no. 1, 215–237. doi:10.1214/14-AOS1274.