MATH888: High-dimensional probability and statistics Fall 2021

Lecture 40 — December 10, 2021

Sebastien Roch, UW-Madison *Scribe: Aidan Howells*

1 Overview

In this and the previous lecture we discussed the stochastic block model. This lecture was given from slides, available on the course website. The first half of the slides were review of last time, and the rest was new material based on Sections 3.1 and 3.4 in [1]. These lecture notes cover the new material, with notation changed to match last time.

2 Notation

Recall that we are concerned with detecting communities in the stochastic block model (SBM). Here n, the number of vertices, is a positive even integer. The vertices are partitioned uniformly at random into two communities of size $n/2$. Each pair of vertices in the same community has an edge between them with probability $q_{\text{in}} \in [0,1]$, and each pair in different communities has an edge with probability $q_{\text{out}} \in [0, q_{\text{in}}]$, all independently. Let $\Pi_2^n = \{\mathbf{x} \in \{\pm 1\}^n : \mathbf{x}^\intercal \mathbf{1} = 0\}$ be the set of possible partitions (we think of i as being in the $+$ community when $x_i = +1$, and in the $-$ community otherwise). Let (X, A) be drawn from the SBM, so that $X \in \Pi_2^n$ is uniformly random and A is the (random) adjacency matrix of the resulting graph.

3 Spectral Clustering

Last time we discussed estimating the partition into communities, given the adjacency matrix A, by an $\mathbf{x} \in \Pi_2^n$ which maximized $\mathbf{x}^\intercal A \mathbf{x}$. We suggested that the correct approach was to relax this optimization problem. The relaxed optimization problem that we wish to consider is the following:

$$\max_{\substack{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2^2 = n \\ \mathbf{x}^\intercal \mathbf{1} = 0}} \mathbf{x}^\intercal A \mathbf{x}.$$

Consider the matrix

$$M := A - \frac{q_{\text{in}} + q_{\text{out}}}{2} \mathbf{1}\mathbf{1}^\intercal + q_{\text{in}} I$$

Let $M^* = \mathbb{E}[M|X]$. (Alternatively, you can assume that M is indexed so that all members of the first community come before those of the second community, and just look at $\mathbb{E}[M]$). One can check that M^*_{ij} is $(q_{\text{in}} - q_{\text{out}})/2$ if i and j are in the same community and $-(q_{\text{in}} - q_{\text{out}})/2$ otherwise. It follows that the leading eigenvalue of M^* is

$$\lambda^* = \frac{(q_{\text{in}} - q_{\text{out}})n}{2},$$

with associated unit eigenvalue $\mathbf{u}^* = (u_i^*)$ given by

$$u_i^* = \begin{cases} n^{-1/2} & X_i = +1 \\ -n^{-1/2} & X_i = -1 \end{cases}$$

If we knew $\mathbf{u}^*$, we would be done. So our strategy to estimate X is the following spectral clustering algorithm:

1. Compute the leading eigenvector $\mathbf{u} = (u_i)$ of M.
2. Estimate X via the vector $\hat{X} = (\mathrm{sgn}(u_i))_{i=1}^n$.

The intuition is that w.h.p. $\mathbf{u}^*$ should be close to $\mathbf{u}$. Indeed, next week we intend to sketch the proof of the following theorem:

Theorem 1 ([1, Thm 3.8])**.** *Suppose that* $q_{\text{in}} \gtrsim \log(n)/n$ *and* $\sqrt{q_{\text{in}}/n} = o(q_{\text{in}} - q_{\text{out}})$*. With probability exceeding* $1 - O(n^{-8})$*, the spectral method achieves*

$$\max_{s \in \{\pm 1\}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i - s\hat{X}_i\} = 1 - o(1).$$

Note the following special cases:

Special Case 1: $q_{\text{in}} - q_{\text{out}} \gg \sqrt{\log n}/n$ if $q_{\text{in}} \asymp \log(n)/n$

Special Case 2: $q_{\text{in}} - q_{\text{out}} \gg 1/\sqrt{n}$ if $q_{\text{in}} \asymp 1$

The proof of the theorem will use the following theorem, which we will not prove:

Theorem 2 ([1, Thm 3.4])**.** *Consider a symmetric random matrix* $Y = [Y_{ij}] \in \mathbb{R}^{n \times n}$*, whose entries are independently generated, mean zero, and uniformly bounded by* $B \in (0, \infty)$*. Define*

$$v = \max_i \sum_j \mathbb{E}[Y_{ij}^2].$$

There there exists some universal constant $c > 0$ *such that for any* $t \geq 0$*,*

$$\mathbb{P}\left\{\|Y\| \geq 4\sqrt{\nu} + t\right\} \leq n \exp\left(-\frac{t^2}{cB^2}\right).$$

In addition to this theorem, the proof will also use a version of the Davis–Kahan theorem. It is also true, though the Davis–Kahan theorem will not suffice to prove, that the spectral method will give us exact recovery:

Theorem 3 ([1, Thm 4.6])**.** *Fix* $\varepsilon > 0$*. Suppose that* $q_{\text{in}} = \alpha \log(n)/n$ *and* $q_{\text{out}} = \beta \log(n)/n$ *for some sufficiently large constants* $\alpha > \beta > 0$*. In addition, assume that*

$$(\sqrt{q_{\text{in}}} - \sqrt{q_{\text{out}}})^2 \geq 2(1 - \varepsilon)\frac{\log n}{n}.$$

With probability $1 - o(1)$*, the spectral method yields* $X = s\hat{X}$ *for some* $s \in \{\pm 1\}$*.*

References

[1] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, *Spectral Methods for Data Science: A Statistical Perspective*, `https://arxiv.org/pdf/2012.08496.pdf`