Topics in High-Dimensional Probability and Statistics*

Lecture 7: Community detection in random graphs

Contents

1 Stochastic block model

The stochastic block model (SBM) is a general model for generating non-directed and simple[1] random graphs exhibiting a certain community structure. In these few pages, we'll cover a small aspect of the theory developed in the context of this model. For more details in this direction, we refer the reader to [1] and the references therein.

1.1 Stochastic block model with two classes

Consider $V = \{1, \dots, n\}$ as a vertex set. Consider a fixed collection of labels

$$\{x_i\}_{i=1}^n, \quad x_i \in \{-1, +1\},$$

assigned to these vertices. The vertices are then split into two communities

$$V_- = \{i : x_i = -1\} \quad \text{and} \quad V_+ = \{i : x_i = +1\},$$

of respective sizes

$$n_- = |V_-| \quad \text{and} \quad n_+ = |V_+|.$$

Now, fix $q < p \in (0, 1)$, and consider the random graph G on these vertices whose adjacency matrix $A = (A_{i,j})_{i,j=1}^n$ has random and independent entries, has $A_{i,i} = 0$, and satisfies

$$\mathbb{P}(A_{i,j} = 1) = \begin{cases} p & \text{if } x_i x_j = +1, \\ q & \text{if } x_i x_j = -1. \end{cases}$$

In other words, two vertices in the same community (resp. different communities) are connected with probability p (resp. q) independently of other connections. Since $q < p$, a typical realization of graph G will display a community structure since vertices in the same community will tend to be more densely connected.

*Teaching material can be found at https://www.qparis-math.com/teaching.

[1] At most one edge between two vertices and no edge from a vertex to itself.

1.2 Recovering communities

A statistical question of interest is the following: Observing only one realization of the adjacency matrix A, and knowing connection probabilities p and q, can we recover the two communities V_- and V_+ ? The problem is equivalent to recovering the (unobserved) labels $\{x_i\}_{i=1}^n$ up to a sign flip and the goal is therefore to construct

$$\{\hat{x}_i\}_{i=1}^n, \quad \hat{x}_i \in \{-1, +1\},$$

based only on A, such that the proportion of misclassified points

$$\min_{\varepsilon \in \{-1,+1\}} \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\hat{x}_i \neq \varepsilon x_i\},$$

is as small as possible (with high probability).

2 Spectral clustering

In this section, we present the main insight we'll use to construct $\{\hat{x}_i\}_{i=1}^n$. First, for technical reasons, let us introduce the modified adjacency matrix A° defined as follows. Let $\{\xi_i\}_{i=1}^n$ be independent Bernoulli random variables with parameter p, i.e., such that

$$\mathbb{P}(\xi_i = 0) = 1 - p \quad \text{and} \quad \mathbb{P}(\xi_i = 1) = p,$$

and let

$$A^\circ = A + \mathrm{diag}(\xi_1, \dots, \xi_n).$$

Matrix A° corresponds to the adjacency matrix of the graph G where loops are added to each vertex with probability p, and can be easily constructed given A. Then, observe that

$$\mathbb{E}[A^\circ] = \frac{p+q}{2} \mathbf{1}_n \mathbf{1}_n^\top + \frac{p-q}{2} x x^\top,$$

where $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector with each entry equal to 1 and $x = (x_1, \dots, x_n)^\top$ is the vector of labels. As a result, denoting

$$M = \frac{p-q}{2} x x^\top \quad \text{and} \quad \hat{M} = A^\circ - \frac{p+q}{2} \mathbf{1}_n \mathbf{1}_n^\top,$$

we obtain

$$\hat{M} = M + (A^\circ - \mathbb{E}[A^\circ]),$$

so that, in particular,

$$\mathbb{E}[\hat{M}] = M.$$

Remark 2.1. *The above representation is interesting for the following reason. Suppose we could access to matrix M. Then we'd be able to reconstruct exactly the two communities. Indeed, matrix M has rank 1 and the label vector $x/\sqrt{n}$ is (up to a sign flip) the unique unit eigenvector of M associated with its non-zero eigenvalue $n(p-q)/2$. In practice, we can only access to one realization of the adjacency matrix A. But from A, and the knowledge of p and q, we can easily construct $\hat{M}$ which is an unbiased estimator of M.*

As result, we can consider the following simple strategy: Let

$$\hat{M} = \sum_{i=1}^{n} \lambda_i u_i u_i^\top,$$

be a spectral decomposition of $\hat{M}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the eigenvalues of $\hat{M}$ and $u_1, \ldots, u_n \in \mathbb{R}^n$ are associated orthonormal eigenvectors. Then, for $1 \leq i \leq n$, we define

$$\hat{x}_i = \mathrm{sign}(u_1^i) = \begin{cases} +1 & \text{if } u_1^i \geq 0, \\ -1 & \text{if } u_1^i < 0, \end{cases} \tag{2.1}$$

where u_1^i denotes the i-th coordinate of u_1.

3 Performance evaluation

In the next section, we are going to prove the following result.

Theorem 3.1. *Suppose $\hat{x}_i$ is constructed as in (2.1). Then for any $\delta \in (0,1)$, the proportion of misclassified points satisfies,*

$$\min_{\varepsilon\in\{-1,+1\}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{x}_i \neq \varepsilon x_i\} \leq \frac{128}{n^2(p-q)^2} \max\left\{ v_{p,q} \log\left(\frac{2n}{\delta}\right), \frac{4}{9}\log^2\left(\frac{2n}{\delta}\right)\right\},$$

with probability at least $1-\delta$, where

$$v_{p,q} = \frac{3n_+ - n_-}{2} v_p + \sqrt{\frac{(n_+ - n_-)^2}{4} v_p^2 + n_+ n_- v_q^2},$$

and $v_u := u(1-u)$.

Let us discuss the implications of this result in the simple situation where

$$n_- = n_+ = \frac{n}{2}.$$

In this case we get

$$v_{p,q} = n\left(\frac{v_p + v_q}{2}\right) \leq n\left(\frac{p+q}{2}\right),$$

and the upper bound in Theorem 3.1 is less than

$$\max\left\{ \frac{c_1(p+q)}{(p-q)^2 n} \log\left(\frac{2n}{\delta}\right), \frac{c_2}{(p-q)^2 n^2}\log^2\left(\frac{2n}{\delta}\right)\right\}. \tag{3.1}$$

The above expression goes to 0 with the size n of the graph whenever $q < p$ are independent of n which corresponds to the so called *dense regime*. The *sparse regime* corresponds to

$$p = \frac{a_n}{n} \quad \text{and} \quad q = \frac{b_n}{n},$$

where $b_n \leq a_n$ and $a_n, b_n \ll n$. One easily checks that, for any $\alpha > 0$ and any $0 < b < a$, if

$$a_n = a \log^{1+\alpha} n \quad \text{and} \quad b_n = b \log^{1+\alpha} n,$$

the expression (3.1) tends to 0 with n. Using slightly more sophisticated proofs than the one we present next, we can actually show that the proportion of missclassified vertices tends to zero with high probability provided

4 Proof of Theorem 3.1

We start with a general statement connecting the proportion of misclassified points and the operator norm of

$$A^\circ - \mathbb{E}[A^\circ].$$

Lemma 4.1. *Suppose $\hat{x}_i$ is constructed as in (2.1), then the proportion of misclassified points satisfies,*

$$\min_{\varepsilon\in\{-1,+1\}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{x}_i \neq \varepsilon x_i\} \leq \frac{32}{n^2(p-q)^2} \|A^\circ - \mathbb{E}[A^\circ]\|_{\mathrm{op}}^2.$$

Proof. For all $1 \leq i \leq n$,

$$\begin{aligned} \mathbf{1}\{\hat{x}_i \neq \varepsilon x_i\} &= \mathbf{1}\{\mathrm{sign}(u_1^i) \neq \varepsilon x_i\} \\ &= \mathbf{1}\{\mathrm{sign}(\sqrt{n} u_1^i) \neq \varepsilon x_i\} \\ &\leq (\sqrt{n} u_1^i - \varepsilon x_i)^2 \\ &= n(u_1^i - \varepsilon \frac{x_i}{\sqrt{n}})^2. \end{aligned}$$

As a result, the proportion of misclassified points satisfies

$$\min_{\varepsilon\in\{-1,+1\}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{x}_i \neq \varepsilon x_i\} \leq \min_{\varepsilon\in\{-1,+1\}} \|u_1 - \varepsilon \frac{x}{\sqrt{n}}\|_2^2.$$

Since u_1 and $x/\sqrt{n}$ are both unit eigenvectors associated to the largest eigenvalue of $\hat{M}$ and M respectively, and since the largest eigenvalue of M is $n(p-q)/2$, the Davis-Kahane $sin(\theta)$ theorem (Theorem A.1) implies that

$$\begin{aligned} \min_{\varepsilon\in\{-1,+1\}} \|u_1 - \varepsilon \frac{x}{\sqrt{n}}\|_2^2 &\leq \frac{32}{n^2(p-q)^2} \|\hat{M} - M\|_{\mathrm{op}}^2 \\ &= \frac{32}{n^2(p-q)^2} \|A^\circ - \mathbb{E}[A^\circ]\|_{\mathrm{op}}^2, \end{aligned}$$

which concludes the proof. □

We now provide a bound for the operator norm of

$$A^\circ - \mathbb{E}[A^\circ].$$

Lemma 4.2. *For any $\delta \in (0,1)$,*

$$\|A^\circ - \mathbb{E}[A^\circ]\|_{\mathrm{op}} \leq \max\left\{ \sqrt{4 v_{p,q} \log\left(\frac{2n}{\delta}\right)}, \frac{4}{3}\log\left(\frac{2n}{\delta}\right)\right\},$$

with probability at least $1-\delta$, where

$$v_{p,q} = \frac{3n_+ - n_-}{2} v_p + \sqrt{\frac{(n_+ - n_-)^2}{4} v_p^2 + n_+ n_- v_q^2},$$

and $v_u := u(1-u)$.

Proof. We are going to control $\|A^\circ - \mathbb{E}[A^\circ]\|_{\mathrm{op}}$ using the Matrix Bernstein inequality (Theorem A.2). First, note that

$$A^\circ - \mathbb{E}[A^\circ] = \sum_{i\leq j} X^{i,j},$$

where $X^{i,j} = (X_{k,l}^{i,j}) \in \mathbb{R}^{n\times n}$ denotes the random matrix defined by $X_{k,l}^{i,j} = 0$ if $(k,l) \notin \{(i,j),(j,i)\}$ and

$$X_{i,j}^{i,j} = X_{j,i}^{i,j} = A_{i,j}^\circ - \mathbb{E}[A_{i,j}^\circ].$$

Note that all matrices $(X^{i,j})_{i\leq j}$ are all independent and satisfy

$$\|X^{i,j}\|_{\mathrm{op}} = |A^\circ_{i,j} - \mathbb{E}[A^\circ_{i,j}]| \leq 1.$$

Note finally that, for all $i \leq j$, $\mathbb{E}[(X^{i,j})^2] \in \mathbb{R}^{n\times n}$ is the matrix whose (k,l)-entry is 0 if $(k,l) \notin \{(i,j),(j,i)\}$ and whose (i,j) and (j,i) entries are both equal to

$$\mathrm{Var}(A^\circ_{i,j}) = \left\{ \begin{array}{ll} p(1-p) & \text{if } x_i x_j = +1, \\ q(1-q) & \text{if } x_i x_j = -1. \end{array} \right.$$

As a result, for some permutation matrix Q,

$$\sum_{i\leq j} \mathbb{E}[(X^{i,j})^2] = Q \begin{pmatrix} p(1-p)\mathbf{1}_{n_+,n_+} & q(1-q)\mathbf{1}_{n_+,n_-} \\ q(1-q)\mathbf{1}_{n_-,n_+} & p(1-p)\mathbf{1}_{n_-,n_-} \end{pmatrix} Q^\top,$$

where $\mathbf{1}_{n,m}$ denotes the $n \times m$ matrix with all entries equal to 1. In particular, we deduce (and leave it as an exercise) that

$$\begin{aligned} v_{p,q}(n) &:= \|\sum_{i\leq j} \mathbb{E}[(X^{i,j})^2]\|_{\mathrm{op}} \\ &= \frac{3n_+ - n_-}{2} v_p + \sqrt{\frac{(n_+ - n_-)^2}{4} v_p^2 + n_+ n_- v_q^2}, \end{aligned}$$

where

$$v_p := p(1-p) \quad \text{and} \quad v_q = q(1-q).$$

A direct application of Theorem A.2 therefore implies that, for every $\delta \in (0,1)$,

$$\|A^\circ - \mathbb{E}[A^\circ]\|_{\mathrm{op}} \leq \max\left\{ \sqrt{4v(n)\log\left(\frac{2n}{\delta}\right)}, \frac{4}{3}\log\left(\frac{2n}{\delta}\right) \right\},$$

with probability at least $1-\delta$. □

A Appendix

Theorem A.1 (Davis-Kahan $\sin(\theta)$ theorem)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices. Consider the spectral decompositions*

$$A = \sum_{i=1}^n \lambda_i u_i u_i^\top \quad \textit{and} \quad B = \sum_{i=1}^n \mu_i v_i v_i^\top,$$

where $\lambda_1 \geq \lambda_2 \geq \dots$ (resp. $\mu_1 \geq \mu_2 \geq \dots$) are the eigenvalues of A (resp. B) and u_i (resp. v_i) is a unit eigenvector of A (resp. B) associated to eigenvalue λ_i (resp. μ_i). Then, for all $1 \leq i \leq n$,

$$\min_{\varepsilon\in\{-1,+1\}} \|u_i - \varepsilon v_i\|_2 \leq \frac{2\sqrt{2}\|A-B\|_{\mathrm{op}}}{\min_{j\neq i}\{|\lambda_i - \lambda_j|\}}.$$

Theorem A.2 (Matrix Bernstein)**.** *Let $X_1, \dots, X_m \in \mathbb{R}^{d\times d}$ be independent random symmetric matrices. Suppose in addition that there exists $B > 0$ such that, for all $1 \leq i \leq m$,*

$$\mathbb{E}[X_i] = 0_{d\times d} \quad \textit{and} \quad \|X_i\|_{\mathrm{op}} \leq B.$$

Then, for all $t > 0$,

$$\mathbb{P}\left(\|\sum_{i=1}^m X_i\|_{\mathrm{op}} \geq t \right) \leq 2d \exp\left(-\frac{t^2}{2v(m) + \frac{2Bt}{3}} \right),$$

where

$$v(m) := \|\sum_{i=1}^m \mathbb{E}[X_i^2]\|_{\mathrm{op}}.$$

In particular, for all $\delta \in (0,1)$,

$$\|\sum_{i=1}^m X_i\|_{\mathrm{op}} \leq \max\left\{ \sqrt{4v(m)\log\left(\frac{2d}{\delta}\right)}, \frac{4B}{3}\log\left(\frac{2d}{\delta}\right) \right\},$$

with probability at least $1-\delta$.

References

[1] E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.