Stochastic Block Model

Introduced by Holland, Laskey, Leinhardt in 1983

① each of n nodes assigned a community

$$\pi : V \to [k]$$

② edges sampled independently

$$\mathbb{P}[(i,j) \in E] = W_{\pi(i),\pi(j)}$$

Goal: recover the planted community structure

Special case: π is a bisection and

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

where $p > q$ is called assortative

def: The sparsity of a cut $(U, V \setminus U)$ is

$$\phi(U) = \frac{|E(U, V \setminus U)|}{\min(|U|, |V \setminus U|)}$$

Easy to see the bisection that minimizes sparsity in expectation is the planted one

claim: [informal] Under various assumptions on the parameters, whp the sparsest bisection

is the planted one too

① how do we bound the deviations?

② how do we find the minimizer?

Finding the minimum bisection is NP-hard. Nevertheless, many approaches succeed in this generative model

Combinatorial Approach: Given $i, j \in V$, are they in same or diff communities?

same: expected number of common neighbors

$$\sim p^2 \frac{n}{2} + \frac{q^2 n}{4}$$

different: $\sim pqn$

If $p = \frac{1}{2}$, $q = \frac{1}{4}$, nodes in same community will have many more common neighbors

Spectral Approach: Following [McSherry]

Recall the adjacency matrix A is $V \times V$ with

$$A_{ij} = 1 \text{ iff } (i,j) \in E$$

The main idea is to write

$$A = \mathbb{E}[A] + \text{noise}$$

After permuting the rows and columns based on community structure

$$\mathbb{E}[A] = \underbrace{\begin{bmatrix} pJ & qJ \\ qJ & pJ \end{bmatrix}}_{M} - pI$$

where J = all ones matrix

Proposition: M is rank two and its non-zero e-vals are

$$\left(\frac{p+q}{2}\right)n \quad \text{and} \quad \left(\frac{p-q}{2}\right)n$$

Proof: We can guess the eigenvectors

$$\vec{1} \quad \text{and} \quad \begin{bmatrix} \vec{1} \\ -\vec{1} \end{bmatrix} \begin{matrix} \} \frac{n}{2} \\ \} \frac{n}{2} \end{matrix}$$

Intuitively, if the noise is small enough we expect the eigenvectors of M and A to be close

Thus our algorithm is:

Spectral Partitioning

① Compute the adjacency matrix A

② Compute the second eigenvector v

③ threshold based on v

Put i in community #1 $\iff v_i \geq 0$

Main Questions:

① How do we bound the noise?

② How do we bound the distance between v and the second eigenvector of M?

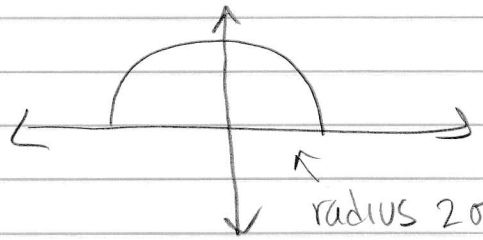
Some Random Matrix Theory

Suppose we take a Gaussian random matrix

$$R_{ij} = \begin{cases} N(0,\sigma^2) & \text{for } i>j \\ 0 & \text{for } i=j \\ R_{ji} & \text{for } i<j \end{cases}$$

What do its eigenvalues look like?

They converge to a semicircle



This is called the semicircle law

To be more precise, well-behaved functions f on the spectrum satisfy

$$\sum_i f(\lambda_i) \to \int f(x)\, d\mu$$

μ: measure on semicircle

Even more beautifully, there is universality

i.e if off-diagonals are iid from any distribution with mean zero, variance σ^2, same result holds

But we need different sorts of bounds

Let $\hat{A} = A + pI$, same eigenvalues, easier to work with

Let $R = \hat{A} - M$

Thm [Vu] There are constants C_1, C_2 s.t.

$$\|R\| \leq 2\sqrt{pn} + C_1 (pn)^{1/4} \ln n$$

with prob $1 - o(1)$, as long as $p \geq C_2 \frac{\ln^4 n}{n}$

Matrix Perturbation Bounds

* than what we used for Jennrich

we need sharper* tools to bound e-vecs

Thm [Davis-Kahan] If $\hat{A}$ and M are symmetric and $R = M - \hat{A}$ then

$$\theta = \angle(v_i, w_i) \leq \frac{2\|R\|}{\min_{j \neq i} |\mu_i - \mu_j|}$$

v_i, w_i: ith eigenvectors of $\hat{A}$ and M, sorted by eigenvalue

μ_i, μ_j: e-vals of M

Now let's bound how many nodes we misclassify

Let $\delta = v_2 - w_2$ (unit vectors). If i is misclassified, then

$$|\delta_i| \geq \frac{1}{\sqrt{n}}$$

Hence if we misclassify k vertices then

$$\|\delta\| \geq \sqrt{\frac{k}{n}}$$

Now since v_2 and w_2 are unit vectors

$$\|\delta\| \leq \sqrt{2} \sin\theta$$

And combining our bounds on $\|R\|$ and Davis-Kahan we have

$$\|\delta\| \leq C \frac{\sqrt{pn}}{\frac{n}{2}(p-q)} \Rightarrow k = O\left(\frac{p}{(p-q)^2}\right)$$

Eg. if $p = \frac{1}{2}$ and $a = p - \frac{c}{\sqrt{n}}$ we misclassify only a constant fraction

Finally, let's prove Davis-Kahan

Proof: WLOG we may assume $\mu_i = 0$ by shifting

$$M \leftarrow M - \mu_i I$$
$$\hat{A} \leftarrow \hat{A} - \mu_i I$$

Now $\|R\| = \|M - \hat{A}\| \geq |\lambda_i|$ (Weyl's inequality)

Moreover expand $v_i = i^{th}$ e-vector of $\hat{A}$ in terms of e-vectors of M

$$v_i = \sum_j c_j w_j$$

Since $\hat{A}$ and M are symmetric, the eigenvectors are orthogonal which means

$$c_j = w_j^T v_i \quad \text{and} \quad c_i = w_i^T v_i = \cos\theta$$

Finally let $\delta = \min_{j \neq i} |\mu_j|$

Now we have

$$\|M v_i\|^2 = \sum_j c_j^2 \mu_j^2$$
$$\geq \sum_{j \neq i} c_j^2 \delta^2 = \delta^2 (1 - c_i^2)$$

$$= \delta^2 \sin^2\theta$$

Also we have

$$\|Mv_i\| \leq \|\hat{A}v_i\| + \|Rv_i\|$$

$$= \lambda_i + \|Rv_i\| \leq 2\|R\|$$

Putting it all together completes the proof. ∎

Observe that the dependence on δ is necessary, e.g. if we have

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

it has an e-value of one with multiplicity two

But tiny perturbations can drastically change e-vectors

When there is no gap, can instead bound angles between subspaces, e.g. perturbations of

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

do not change $\text{span}(w_1, w_2)$ much

Sharp Thresholds

Bounds from spectral clustering are qualitatively tight in many regimes

Main Question: what are the fundamental limits for community detection?

e.g. exact recovery

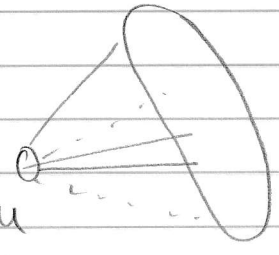
Thm [Abbe, Bandeira, Hall] If $p = \frac{a \log n}{n}$, $q = \frac{b \log n}{n}$ for binary symmetric SBM exact recovery with prob $1-o(1)$ is possible if:

$$(\sqrt{a} - \sqrt{b})^2 > 2$$

and impossible if $(\sqrt{a} - \sqrt{b})^2 < 2$.

Made algorithmic by [Hajek, Wu, Xu]

Simple heuristic for where it comes from



Given the colors of $V \setminus \{u\}$ and edges $E(\{u\}, V \setminus \{u\})$, what is the color of u?

Take S = # edges to red − # edges to blue

If $\pi(u) = \text{red}$ then

$$S \sim \text{Bin}(p, \tfrac{n}{2}-1) - \text{Bin}(q, \tfrac{n}{2})$$

else $S \sim \text{Bin}(q, \tfrac{n}{2}) - \text{Bin}(p, \tfrac{n}{2}-1)$

claim $\mathbb{P}[S > 0] > 1 - o(\frac{1}{n})$ if $(\sqrt{a}-\sqrt{b})^2 > 2$

Thus we can union bound over all n nodes and still get failure prob $o(1)$

[Abbe, Bandeira, Hall] analyzed MLE, i.e. minimum bisection which is NP-hard

[Hajek, Wu, Xu] analyzed the SDP relaxation for minimum bisection (later)

Many other sharp thresholds, often related to optimal error exponents of large deviations (e.g. f-divergences)