Planted partition models

Daniel Hsu

COMS 4772

Planted partition models

- Also called "stochastic block models" in statistics.
- Regarded as model for "community structure" in networks.
- Extremely fashionable, not very realistic.
- Interesting to study.

Planted bisection

- n people, partition into two groups of $n/2$ each.
- Appearance of edges (e.g., links, friendship, interaction) between people are random and independent.
  - Two people in same group have edge with probability p.
  - Two people in different groups have edge with probability $q < p$.
- Only observe edges (adjacency matrix); partition is "hidden".
- **Goal**: recover the groups.

Random adjacency matrix

- Random adjacency matrix $\boldsymbol{A}$ in $\{0,1\}^{n \times n}$
- Expected value:

$$\mathbb{E}(\boldsymbol{A}) \;=\; \left(\begin{array}{ccc|ccc} p & p & p & q & q & q \\ p & p & p & q & q & q \\ p & p & p & q & q & q \\ \hline q & q & q & p & p & p \\ q & q & q & p & p & p \\ q & q & q & p & p & p \end{array}\right)$$

(Assuming people are ordered so first group is $1, 2, \ldots, n/2$.)

Spectral analysis

- $\mathbb{E}(\boldsymbol{A})$ has rank 2:

$$\mathbb{E}(\boldsymbol{A}) \;=\; \frac{p+q}{2}\left(\begin{array}{ccc|ccc} +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 & +1 & +1 \\ \hline +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 & +1 & +1 \end{array}\right)$$
$$+\;\frac{p-q}{2}\left(\begin{array}{ccc|ccc} +1 & +1 & +1 & -1 & -1 & -1 \\ +1 & +1 & +1 & -1 & -1 & -1 \\ +1 & +1 & +1 & -1 & -1 & -1 \\ \hline -1 & -1 & -1 & +1 & +1 & +1 \\ -1 & -1 & -1 & +1 & +1 & +1 \\ -1 & -1 & -1 & +1 & +1 & +1 \end{array}\right).$$

Spectral clustering

- Top eigenvalue and eigenvector of $\mathbb{E}(\boldsymbol{A})$:

$$\lambda_1 \;=\; \frac{p+q}{2}\cdot n\,, \qquad \boldsymbol{v}_1 \;=\; \frac{1}{\sqrt{n}}\boldsymbol{1}\,.$$

- Second eigenvalue and eigenvector of $\mathbb{E}(\boldsymbol{A})$:

$$\lambda_2 \;=\; \frac{p-q}{2}\cdot n\,, \qquad v_{2,i} = \begin{cases} +\frac{1}{\sqrt{n}} & \text{if person } i \text{ in group } 1\,, \\ -\frac{1}{\sqrt{n}} & \text{if person } i \text{ in group } 2\,. \end{cases}$$

- **Spectral clustering**: extract second eigenvector $\hat{\boldsymbol{v}}_2$ of $\boldsymbol{A}$, and partition people based on sign of corresponding entry in $\hat{\boldsymbol{v}}_2$.

Noise

- $\boldsymbol{A} = \mathbb{E}(\boldsymbol{A}) + \boldsymbol{Z}$ for some zero-mean random matrix $\boldsymbol{Z}$.
- Using Matrix Bernstein inequality: with high probability,

$$\|\boldsymbol{Z}\|_2 \;\leq\; O\Big(\sqrt{pn\log n} + \log n\Big)\,.$$

- Sharper result (Vu, 2007): with high probability,

$$\|\boldsymbol{Z}\|_2 \;\leq\; C\sqrt{pn}$$

  whenever $p \geq \frac{C' \log^4 n}{n}$.
- Now relate eigenvectors of $\boldsymbol{A}$ to that of $\mathbb{E}(\boldsymbol{A})$.

Perturbation analysis

- Pretend we already know $(p+q)/2$.
- Let $\boldsymbol{v}^*$ be top eigenvector of $\mathbb{E}(\boldsymbol{A}) - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top$
  - $v_i^* = \pm\frac{1}{\sqrt{n}}$, corresponding eigenvalue $\lambda^* = \frac{p-q}{2} \cdot n$.
- Let $\hat{\boldsymbol{v}}$ be top eigenvector of $\boldsymbol{A} - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top$.
  - Using Weyl's inequality: corresponding eigenvalue

$$\hat{\lambda} \;\geq\; \frac{p-q}{2} \cdot n - C\sqrt{pn}\,.$$

  - **Assume**

$$\frac{p-q}{\sqrt{p}} \;\gg\; \frac{1}{\sqrt{n}}$$

    so

$$\hat{\lambda} \;\geq\; \frac{p-q}{2} \cdot n - C\sqrt{pn} \;\geq\; \frac{p-q}{4} \cdot n\,.$$

- Using Davis-Kahan:

$$\varepsilon \;:=\; \|(\boldsymbol{I} - \hat{\boldsymbol{v}}\hat{\boldsymbol{v}}^\top)\boldsymbol{v}^*\|_2 \;\leq\; \frac{C\sqrt{pn}}{\frac{p-q}{4} \cdot n} \;=\; \frac{\sqrt{p}}{p-q} \cdot \frac{4C}{\sqrt{n}} \;\ll\; 1\,.$$

Comparing unit vectors

- $\|(\boldsymbol{I} - \hat{\boldsymbol{v}}\hat{\boldsymbol{v}}^\top)\boldsymbol{v}^*\|_2^2 = 1 - \langle \hat{\boldsymbol{v}}, \boldsymbol{v}^* \rangle^2$, so

$$\min\left\{\|\boldsymbol{v}^* - \hat{\boldsymbol{v}}\|_2^2, \|\boldsymbol{v}^* - (-\hat{\boldsymbol{v}})\|_2^2\right\} = 2(1 - \sqrt{1-\varepsilon^2}) \leq 2\varepsilon^2 .$$

  - (WLOG assume min achieved by $\|\boldsymbol{v}^* - \hat{\boldsymbol{v}}\|_2^2$.)
- **Classification error rate**: since $v_i^* = \pm\frac{1}{\sqrt{n}}$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\mathrm{sign}(v_i^*) \neq \mathrm{sign}(\hat{v}_i)\} &\leq \frac{1}{n}\sum_{i=1}^{n}(1 - n v_i^* \hat{v}_i)^2 \\ &= \sum_{i=1}^{n}(v_i^* - \hat{v}_i)^2 \\ &= \|\boldsymbol{v}^* - \hat{\boldsymbol{v}}\|_2^2 \\ &\leq 2\varepsilon^2 . \end{aligned}$$

Boosting accuracy

- Suppose $2\varepsilon^2 \approx 1/3$, but you really want perfect partitioning.
- Say $\hat{S} \subseteq \{1, 2, \ldots, n\}$ is estimate of first group; about $1/3$ of them actually belong to second group.
- People who are *really* in first group will have more edges with people in $\hat{S}$ than people who are *really* in second group.
  - Use this fact to *very* accurately classify people.
  - (Technically, need independence, but can achieve this by "sample splitting".)