Applications of spectral methods (ℓ_2 theory)

Cong Ma

University of Chicago, Autumn 2021

What we have learned so far

- Classical ℓ_2 matrix perturbation theory:
  - Davis-Kahan's $\sin \boldsymbol{\Theta}$ theorem
  - Wedin's $\sin \boldsymbol{\Theta}$ theorem
  - Eigenvector perturbation of probability transition matrices

- Matrix concentration inequalities:
  - Matrix Bernstein inequality

What we have learned so far

- Classical ℓ_2 matrix perturbation theory:
  - Davis-Kahan's $\sin \boldsymbol{\Theta}$ theorem
  - Wedin's $\sin \boldsymbol{\Theta}$ theorem
  - Eigenvector perturbation of probability transition matrices

- Matrix concentration inequalities:
  - Matrix Bernstein inequality

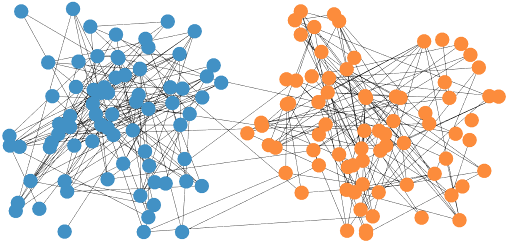
— *we will check their applications today*

Outline

- Community recovery in stochastic block model
- Low-rank matrix completion
- Ranking from pairwise comparisons

Community recovery in stochastic block model

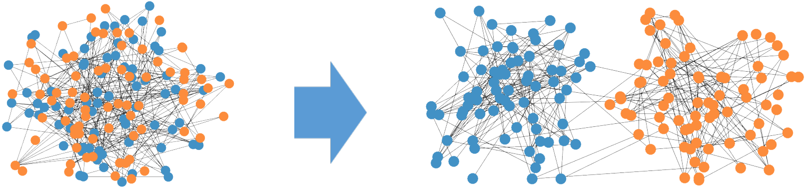
Stochastic block model (SBM)



$x_i^\star = 1$: 1st community $\qquad$ $x_i^\star = -1$: 2nd community

- n nodes $\{1, \ldots, n\}$
- 2 communities
- n unknown variables: $x_1^\star, \ldots, x_n^\star \in \{1, -1\}$
  - encode community memberships

A simple model: stochastic block model (SBM)



$\mathcal{G}$

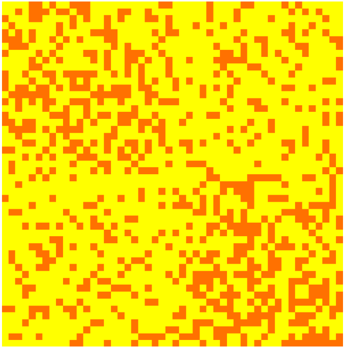
- observe a graph $\mathcal{G}$

$$(i,j) \in \mathcal{G} \text{ with prob. } \begin{cases} p, & \text{if } i \text{ and } j \text{ are from same community} \\ q, & \text{else} \end{cases}$$

  Here, $p > q$

- **Goal:** recover community memberships of all nodes, i.e., $\{x_i^\star\}$

Adjacency matrix

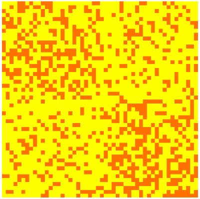
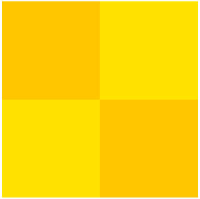
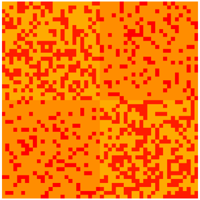


Consider the adjacency matrix $\boldsymbol{A} \in \{0,1\}^{n\times n}$ of $\mathcal{G}$: (assume $A_{ii} = p$)

$$A_{i,j} = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{G} \\ 0, & \text{else} \end{cases}$$

- WLOG, suppose $x_1^\star = \cdots = x_{n/2}^\star = 1$; $x_{n/2+1}^\star = \cdots = x_n^\star = -1$

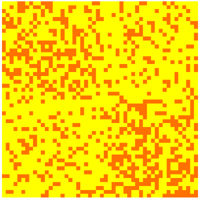
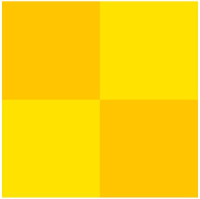
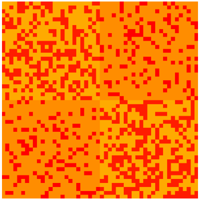
Adjacency matrix

$$\boldsymbol{A} \quad = \quad \underbrace{\mathbb{E}[\boldsymbol{A}]}_{\text{rank 2}} \quad + \quad \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$$

$$\mathbb{E}[\boldsymbol{A}] = \left[\begin{array}{cc} p\mathbf{1}\mathbf{1}^\top & q\mathbf{1}\mathbf{1}^\top \\ q\mathbf{1}\mathbf{1}^\top & p\mathbf{1}\mathbf{1}^\top \end{array}\right] = \underbrace{\frac{p+q}{2}\mathbf{1}\mathbf{1}^\top}_{\text{uninformative bias}} + \frac{p-q}{2} \underbrace{\left[\begin{array}{c} \mathbf{1} \\ -\mathbf{1} \end{array}\right]}_{=\boldsymbol{x}^\star = [x_i]_{1\le i\le n}} \left[\mathbf{1}^\top, -\mathbf{1}^\top\right]$$

Spectral clustering

$$\boldsymbol{A} \quad = \quad \underbrace{\mathbb{E}[\boldsymbol{A}]}_{\text{rank 2}} \quad + \quad \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$$

1. computing the leading eigenvector $\boldsymbol{u} = [u_i]_{1 \le i \le n}$ of $\boldsymbol{A} - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top$
2. rounding: output $x_i = \begin{cases} 1, & \text{if } u_i > 0 \\ -1, & \text{if } u_i < 0 \end{cases}$

Apply Davis-Kahan's result

Let $\quad \boldsymbol{M}^{\star} := \mathbb{E}[\boldsymbol{A}] - \frac{p+q}{2}\mathbf{1}\mathbf{1}^{\top} = \frac{p-q}{2}\begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix}\begin{bmatrix} \mathbf{1}^{\top} & -\mathbf{1}^{\top} \end{bmatrix}$,

$\boldsymbol{M} := \boldsymbol{A} - \frac{p+q}{2}\mathbf{1}\mathbf{1}^{\top}$, and $\boldsymbol{u}^{\star} := \frac{1}{\sqrt{n}}\begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix}$

Then the Davis-Kahan $\sin\boldsymbol{\Theta}$ Theorem yields

$$\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^{\star}) \leq \frac{\|\boldsymbol{M} - \boldsymbol{M}^{\star}\|}{\lambda_1(\boldsymbol{M}^{\star}) - \|\boldsymbol{M} - \boldsymbol{M}\|} = \frac{\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|}{\frac{(p-q)n}{2} - \|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|} \quad (5.1)$$

as long as $\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| < \lambda_1(\boldsymbol{M}^{\star}) = \frac{(p-q)n}{2}$

Bounding $\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|$

Matrix concentration inequalities tell us that

Lemma 5.1

Consider SBM with $p > q$ and $p \gtrsim \frac{\log n}{n}$. Then with high prob.

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| \lesssim \sqrt{np \log n} \tag{5.2}$$

— *better concentration yields $\sqrt{np}$ bound*

Statistical accuracy of spectral clustering

Substitute (5.2) into (5.1) to reach

$$\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \leq \frac{\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|}{\frac{(p-q)n}{2} - \|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|} \lesssim \frac{\sqrt{np\log n}}{(p-q)n}$$

provided that $(p-q)n \gg \sqrt{np\log n}$

Thus, under condition $\frac{p-q}{\sqrt{p}} \gg \sqrt{\frac{\log n}{n}}$, with high prob. one has

$$\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \ll 1 \qquad \Longrightarrow \qquad \text{nearly perfect clustering}$$

Statistical accuracy of spectral clustering

$$\frac{p-q}{\sqrt{p}} \gg \sqrt{\frac{\log n}{n}} \quad \Longrightarrow \quad \text{nearly perfect clustering}$$

- **dense regime:** if $p \asymp q \asymp 1$, then this condition reads

$$p - q \gg \sqrt{\frac{\log n}{n}}$$

- **"sparse" regime:** if $p = \frac{a \log n}{n}$ and $q = \frac{b \log n}{n}$ for $a, b \asymp 1$, then

$$a - b \gg \sqrt{a}$$

This condition is information-theoretically optimal (up to log factor)
— *Mossel, Neeman, Sly '15, Abbe '18*

Proof of Lemma 5.1

To simplify presentation, assume $A_{i,j}$ and $A_{j,i}$ are independent

(check: why this assumption does not change our bounds)

Proof of Lemma 5.1

Write $\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$ as $\sum_{i,j} \boldsymbol{X}_{i,j}$, where $\boldsymbol{X}_{i,j} = \big(A_{i,j} - \mathbb{E}[A_{i,j}]\big)\boldsymbol{e}_i\boldsymbol{e}_j^\top$

- Since $\mathsf{Var}(A_{i,j}) \leq p$, one has $\mathbb{E}\left[\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\right] \preceq p\boldsymbol{e}_i\boldsymbol{e}_i^\top$, which gives

$$\sum\nolimits_{i,j} \mathbb{E}\left[\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\right] \preceq \sum\nolimits_{i,j} p\boldsymbol{e}_i\boldsymbol{e}_i^\top \preceq np\,\boldsymbol{I}$$

  Similarly, $\sum_{i,j} \mathbb{E}\left[\boldsymbol{X}_{i,j}^\top\boldsymbol{X}_{i,j}\right] \preceq np\,\boldsymbol{I}$. As a result,

$$v = \max\left\{\left\|\sum\nolimits_{i,j} \mathbb{E}\left[\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\right]\right\|, \left\|\sum\nolimits_{i,j} \mathbb{E}\left[\boldsymbol{X}_{i,j}^\top\boldsymbol{X}_{i,j}\right]\right\|\right\} \leq np$$

- In addition, $\|\boldsymbol{X}_{i,j}\| \leq 1 =: B$

- Take the matrix Bernstein inequality to conclude that with high prob.,

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| \lesssim \sqrt{v\log n} + B\log n \lesssim \sqrt{np\log n} \quad (\text{since } p \gtrsim \frac{\log n}{n})$$

Low-rank matrix completion

Low-rank matrix completion

$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

figure credit: Candès

- consider a low-rank matrix $\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$
- each entry $M_{i,j}^\star$ is observed independently with prob. p
- **intermediate goal:** estimate $\boldsymbol{U}^\star, \boldsymbol{V}^\star$

Spectral method for matrix completion

1. identify the key matrix $\boldsymbol{M}^\star$
2. construct surrogate matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ as

$$M_{i,j} = \begin{cases} \frac{1}{p} M_{i,j}^\star, & \text{if } M_{i,j}^\star \text{ is observed} \\ 0, & \text{else} \end{cases}$$

   - **rationale for rescaling:** ensures $\mathbb{E}[\boldsymbol{M}] = \boldsymbol{M}^\star$

3. compute the rank-r SVD $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ of $\boldsymbol{M}$, and return $(\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V})$

Statistical accuracy of spectral estimate

Let's analyze a simple case where $\boldsymbol{M} = \boldsymbol{u}\boldsymbol{v}^\top$ with

$$\boldsymbol{u} = \frac{1}{\|\tilde{\boldsymbol{u}}\|_2}\tilde{\boldsymbol{u}}, \quad \boldsymbol{v} = \frac{1}{\|\tilde{\boldsymbol{v}}\|_2}\tilde{\boldsymbol{v}}, \quad \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$$

From Wedin's Theorem: if $p \gg \log^3 n/n$, then with high prob.

$$\max\{\mathsf{dist}(\hat{\boldsymbol{u}}, \boldsymbol{u}), \mathsf{dist}(\hat{\boldsymbol{v}}, \boldsymbol{v})\} \le \frac{\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|}{\sigma_1(\boldsymbol{M}) - \|\hat{\boldsymbol{M}} - \boldsymbol{M}\|} \asymp \underbrace{\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|}_{\text{controlled by Bernstein}}$$

$$\ll 1 \quad \text{(nearly accurate estimates)} \qquad (5.3)$$

Sample complexity

For rank-1 matrix completion,

$$p \gg \frac{\log^3 n}{n} \qquad \Longrightarrow \qquad \text{nearly accurate estimates}$$

Sample complexity needed to yield reliable spectral estimates is

$$\underbrace{n^2 p \asymp n \log^3 n}_{\text{optimal up to log factor}}$$

Proof of (5.3)

Write $\hat{\boldsymbol{M}} - \boldsymbol{M} = \sum_{i,j} \boldsymbol{X}_{i,j}$, where $\boldsymbol{X}_{i,j} = (\hat{M}_{i,j} - M_{i,j})\boldsymbol{e}_i \boldsymbol{e}_j^\top$

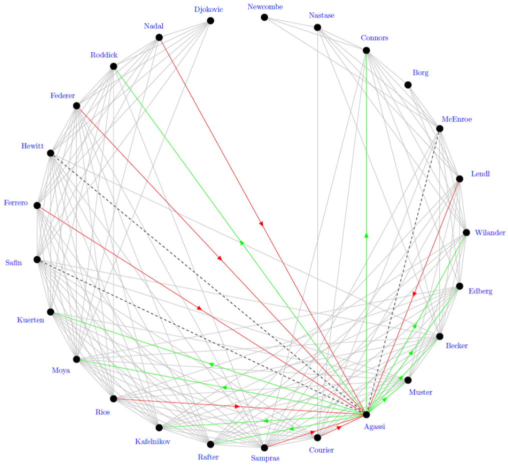
- First,
$$\|\boldsymbol{X}_{i,j}\| \leq \frac{1}{p} \max_{i,j} |M_{i,j}| \lesssim \frac{\log n}{pn} := B \quad \text{(check)}$$

- Next, $\mathbb{E}\big[\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\big] = \mathsf{Var}(\hat{M}_{i,j})\boldsymbol{e}_i\boldsymbol{e}_i^\top$ and hence
$$\mathbb{E}\big[\sum\nolimits_{i,j} \boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\big] \preceq \Big\{\max_{i,j} \mathsf{Var}\big(\hat{M}_{i,j}\big)\Big\} n\boldsymbol{I} \preceq \Big\{\frac{n}{p} \max_{i,j} M_{i,j}^2\Big\}\boldsymbol{I}$$
$$\Longrightarrow \qquad \big\|\mathbb{E}\big[\sum\nolimits_{i,j} \boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\big]\big\| \leq \frac{n}{p} \max_{i,j} M_{i,j}^2 \lesssim \frac{\log^2 n}{np} \quad \text{(check)}$$
Similar bounds hold for $\big\|\mathbb{E}\big[\sum_{i,j} \boldsymbol{X}_{i,j}^\top \boldsymbol{X}_{i,j}\big]\big\|$. Therefore,
$$v := \max\Big\{\big\|\mathbb{E}\big[\sum\nolimits_{i,j} \boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\big]\big\|, \big\|\mathbb{E}\big[\sum\nolimits_{i,j} \boldsymbol{X}_{i,j}^\top\boldsymbol{X}_{i,j}\big]\big\|\Big\} \lesssim \frac{\log^2 n}{np}$$

- Take the matrix Bernstein inequality to yield: if $p \gg \log^3 n/n$, then
$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \lesssim \sqrt{v \log n} + B \log n \ll 1$$

Ranking from pairwise comparisons

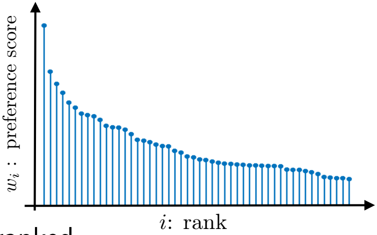
Ranking from pairwise comparisons



pairwise comparisons for ranking tennis players

figure credit: Bozóki, Csató, Temesi

Bradley-Terry-Luce (logistic) model



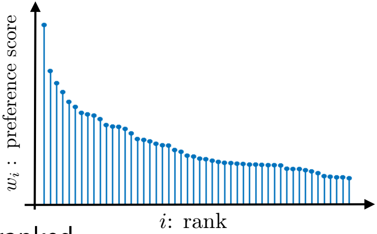
- n items to be ranked
- assign a latent score $\{w_i^\star\}_{1 \leq i \leq n}$ to each item, so that

$$\text{item } i \succ \text{item } j \quad \text{if} \quad w_i^\star > w_j^\star$$

- each pair of items (i, j) is compared independently

$$\mathbb{P}\left\{\text{item } j \text{ beats item } i\right\} = \frac{w_j^\star}{w_i^\star + w_j^\star}$$

Bradley-Terry-Luce (logistic) model



- n items to be ranked
- assign a latent score $\{w_i^\star\}_{1 \leq i \leq n}$ to each item, so that

$$\text{item } i \succ \text{item } j \quad \text{if} \quad w_i^\star > w_j^\star$$

- each pair of items (i, j) is compared independently

$$y_{i,j} \overset{\text{ind.}}{=} \begin{cases} 1, & \text{with prob. } \frac{w_j^\star}{w_i^\star + w_j^\star} \\ 0, & \text{else} \end{cases}$$

- **intermediate goal:** estimate score vector $\boldsymbol{w}^\star$ (up to scaling)

Spectral ranking

1. identify key matrix $\boldsymbol{P}^\star$—probability transition matrix

$$P^\star_{i,j} = \begin{cases} \frac{1}{n} \cdot \frac{w_j^\star}{w_i^\star + w_j^\star}, & \text{if } i \neq j \\ 1 - \sum_{l: l \neq i} P^\star_{i,l}, & \text{if } i = j \end{cases}$$

Rationale:

- $\boldsymbol{P}^\star$ obeys

$$w_i^\star P^\star_{i,j} = w_j^\star P^\star_{j,i} \qquad \text{(detailed balance)}$$

- Thus, the stationary distribution $\boldsymbol{\pi}^\star$ of $\boldsymbol{P}^\star$ obeys

$$\boldsymbol{\pi}^\star \; = \; \frac{1}{\sum_l w_l^\star} \boldsymbol{w}^\star \qquad \text{(reveals true scores)}$$

Spectral ranking

2. construct a surrogate matrix $\boldsymbol{P}$ obeying

$$P_{i,j} = \begin{cases} \frac{1}{n} y_{i,j}, & \text{if } i \neq j \\ 1 - \sum_{l : l \neq i} P_{i,l}, & \text{if } i = j \end{cases}$$

3. return leading left eigenvector $\boldsymbol{\pi}$ of $\boldsymbol{P}$ as score estimate

— closely related to PageRank

Spectral ranking

2. construct a surrogate matrix $\boldsymbol{P}$ obeying

$$P_{i,j} = \begin{cases} \frac{1}{n} y_{i,j}, & \text{if } i \neq j \\ 1 - \sum_{l:l \neq i} P_{i,l}, & \text{if } i = j \end{cases}$$

3. return leading left eigenvector $\boldsymbol{\pi}$ of $\boldsymbol{P}$ as score estimate

— closely related to PageRank

Key: stability of eigenspace against perturbation $\boldsymbol{M} - \boldsymbol{M}^{\star}$?

Statistical guarantees for spectral ranking

— Negahban, Oh, Shah '16, Chen, Fan, Ma, Wang '19

Suppose $\max_{i,j} \frac{w_i}{w_j} \lesssim 1$. Then with high prob.

$$\frac{\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2}{\|\boldsymbol{\pi}\|_2} \asymp \frac{\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{\boldsymbol{\pi}}}{\|\boldsymbol{\pi}\|_2} \lesssim \underbrace{\frac{1}{\sqrt{n}} \to 0}_{\text{nearly perfect estimate}}$$

- a consequence of Theorem **??** and matrix Bernstein (exercise)