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 $\sin \mathbf{\Theta}$ theorem
  - Wedin's $\sin \mathbf{\Theta}$ 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 see their applications today*

Outline

- Community recovery in stochastic block model

  — *application of Davis-Kahan's theorem*

- Low-rank matrix completion

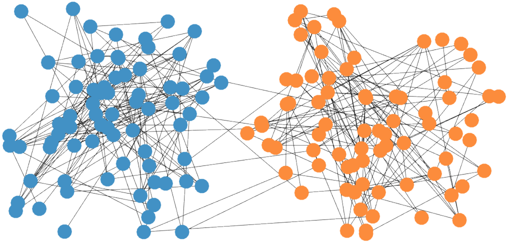
  — *application of Wedin's theorem*

- Ranking from pairwise comparisons

  — *application of eigenvector perturbation of prob. transition matrix*

Community recovery in stochastic block model

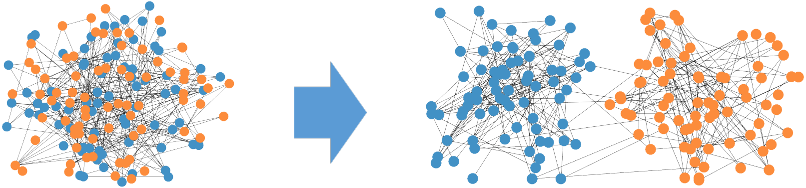
Stochastic block model (SBM)



$x_i^\star = 1$: 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

Stochastic block model (SBM)



$\mathcal{G}$

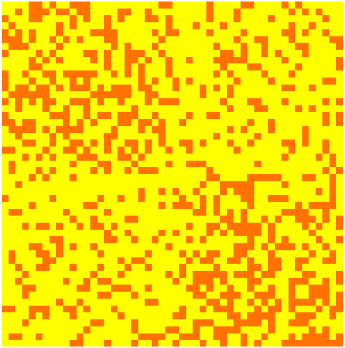
- observe a graph $\mathcal{G}$

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

Here, $p > q$

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

Adjacency matrix

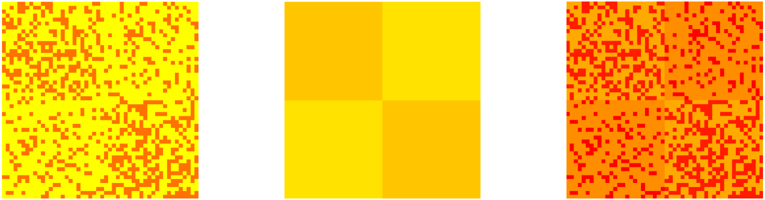


Consider the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ of $\mathcal{G}$: (assume $A_{ii} = p$)

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

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

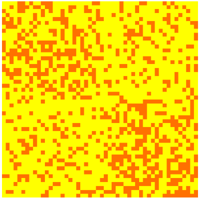
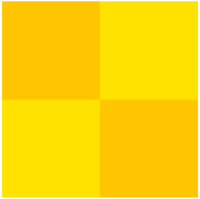
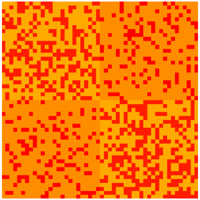
Adjacency matrix



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

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

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 $x_i = \begin{cases} 1, & \text{if } u_i \ge 0 \\ -1, & \text{if } u_i < 0 \end{cases}$

Analysis of spectral clustering

Consider "ground-truth" matrix

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

which obeys

$$\lambda_1(\boldsymbol{M}^{\star}) := \frac{(p-q)n}{2}, \quad \text{and} \quad \boldsymbol{u}^{\star} := \frac{1}{\sqrt{n}}\begin{bmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{bmatrix}.$$

Also, we have perturbed matrix

$$\boldsymbol{M} := \boldsymbol{A} - \frac{p+q}{2}\mathbf{1}\mathbf{1}^{\top}$$

Davis-Kahan implies if $\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| < \lambda_1(\boldsymbol{M}^{\star}) = \frac{(p-q)n}{2}$, then

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

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

Matrix Bernstein inequality tells 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*

- with high probability in this course often means "with probability at least $1 - O(n^{-8})$"

Statistical accuracy of spectral clustering

Substitute ineq. (5.2) into ineq. (5.1) to reach

$$\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \leq \frac{\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|}{\frac{(p-q)n}{2} - \|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|} \lesssim \frac{\sqrt{np \log n}}{(p-q)n} = o(1)$$

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

Now question is

— *how to transfer from estimation error to mis-clustering error*

From estimation error to mis-clustering error

WLOG assume that $\|\boldsymbol{u} - \boldsymbol{u}^\star\|_2 = \mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star)$. Consider the set

$$\mathcal{N} := \{i \mid |u_i - u_i^\star| \geq 1/\sqrt{n}\}$$

We claim that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \neq x_i^\star\} \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{|u_i - u_i^\star| \geq \frac{1}{\sqrt{n}}\right\} = \frac{|\mathcal{N}|}{n}$$

To see this, observe that for any i obeying $x_i \neq x_i^\star$, one has $\mathsf{sgn}(u_i) \neq \mathsf{sgn}(u_i^\star)$, thus indicating that $|u_i - u_i^\star| \geq |u_i^\star| = 1/\sqrt{n}$
In the end, we have

$$|\mathcal{N}| \leq \frac{\|\boldsymbol{u} - \boldsymbol{u}^\star\|_2^2}{(1/\sqrt{n})^2} = o(n)$$

Statistical accuracy of spectral clustering

$$\frac{p-q}{\sqrt{p}} \gg \sqrt{\frac{\log n}{n}} \quad \Longrightarrow \quad \text{almost exact recovery}$$

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

$$p - q \gg \sqrt{\frac{\log n}{n}} \qquad \text{(extremely small gap)}$$

- **"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 5.2

We write $\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$ as sum of independent random matrices

$$\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}] = \sum_{i<j} \big(A_{i,j} - \mathbb{E}[A_{i,j}]\big)\big(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top\big)$$

We only need to consider $\boldsymbol{A}_{\text{upper}} := \sum_{i<j} \underbrace{\big(A_{i,j} - \mathbb{E}[A_{i,j}]\big)\boldsymbol{e}_i\boldsymbol{e}_j^\top}_{=:\boldsymbol{X}_{i,j}}$

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

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

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

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

Proof of Lemma 5.2 (cont.)

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

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| \lesssim \sqrt{v \log n} + B \log n \lesssim \sqrt{np \log n}$$

— *as long as* $p \gtrsim \frac{\log n}{n}$

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 $M^\star_{i,j}$ 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

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

   - **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 $\boldsymbol{M}^\star = \boldsymbol{u}^\star \boldsymbol{v}^{\star\top}$ with

$$\boldsymbol{u}^\star = \frac{1}{\|\tilde{\boldsymbol{u}}\|_2}\tilde{\boldsymbol{u}}, \quad \boldsymbol{v}^\star = \frac{1}{\|\tilde{\boldsymbol{v}}\|_2}\tilde{\boldsymbol{v}}, \quad \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}} \overset{\text{indep.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$$

From Wedin's Theorem: if $\|\boldsymbol{M} - \boldsymbol{M}^\star\| \leq \frac{1}{2}\sigma_1(\boldsymbol{M}^\star) = \frac{1}{2}$, then

$$\max\left\{\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star), \mathsf{dist}(\boldsymbol{v}, \boldsymbol{v}^\star)\right\} \lesssim \frac{\|\boldsymbol{M} - \boldsymbol{M}^\star\|}{\sigma_1(\boldsymbol{M}^\star)} \asymp \|\boldsymbol{M} - \boldsymbol{M}^\star\| \quad (5.3)$$

Bounding $\|\boldsymbol{M} - \boldsymbol{M}^\star\|$

Matrix Bernstein inequality tells us that

Lemma 5.2

Consider matrix completion with $p \gg \frac{\log^3 n}{n}$. *Then with high prob.*

$$\|\boldsymbol{M} - \boldsymbol{M}^\star\| \lesssim \sqrt{\frac{\log^3 n}{n}} = o(1) \tag{5.4}$$

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

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

— *sub-optimal accuracy though*

Proof of inequality (5.4)

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

- First, based on Gaussianity, we have

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

- Next, $\mathbb{E}[\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top] = \mathsf{Var}(M_{i,j})\boldsymbol{e}_i\boldsymbol{e}_i^\top$ and hence

$$\mathbb{E}\big[\sum\nolimits_{i,j}\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\big] \preceq \Big\{\max_{i,j}\mathsf{Var}\big(M_{i,j}\big)\Big\}n\boldsymbol{I} \preceq \Big\{\frac{n}{p}\max_{i,j}(M^\star_{i,j})^2\Big\}\boldsymbol{I}$$

$$\Longrightarrow \qquad \big\|\mathbb{E}\big[\sum\nolimits_{i,j}\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\big]\big\| \leq \frac{n}{p}\max_{i,j}(M^\star_{i,j})^2 \lesssim \frac{\log^2 n}{np} \quad \text{(check)}$$

Similar bounds hold for $\big\|\mathbb{E}\big[\sum_{i,j}\boldsymbol{X}_{i,j}^\top\boldsymbol{X}_{i,j}\big]\big\|$. Therefore,

$$v := \max\Big\{\big\|\mathbb{E}\big[\sum\nolimits_{i,j}\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\big]\big\|, \big\|\mathbb{E}\big[\sum\nolimits_{i,j}\boldsymbol{X}_{i,j}^\top\boldsymbol{X}_{i,j}\big]\big\|\Big\} \lesssim \frac{\log^2 n}{np}$$

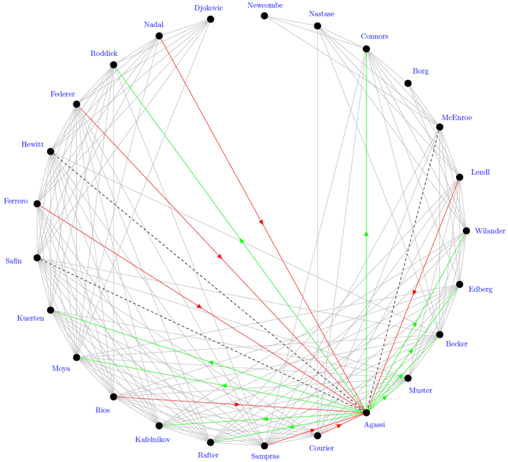
Proof of inequality (5.4) (cont.)

Take the matrix Bernstein inequality to yield: if $p \gg (\log^3 n)/n$, then

$$\|\boldsymbol{M} - \boldsymbol{M}^\star\| \lesssim \sqrt{v \log n} + B \log n \asymp \sqrt{\frac{\log^3 n}{n}} \ll 1$$

Ranking from pairwise comparisons

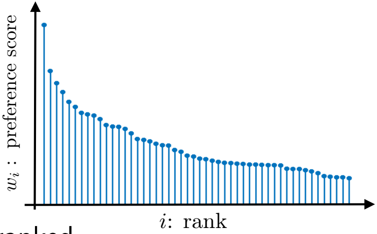
Ranking from pairwise comparisons



pairwise comparisons for ranking tennis players

figure credit: Bozóki, Csató, Temesi

Bradley-Terry-Luce (logistic) model



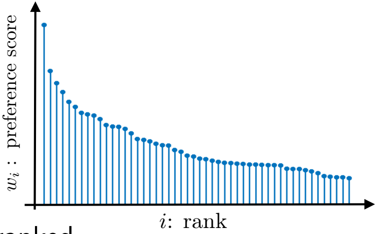
- n items to be ranked
- assign a latent positive score $\{w_i^\star\}_{1 \le i \le n}$ to each item, so that

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

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

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

Bradley-Terry-Luce (logistic) model



- n items to be ranked
- assign a latent positive 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

Analysis of spectral ranking

Apply our perturbation bound to see

$$\|\boldsymbol{\pi}-\boldsymbol{\pi}^\star\|_{\boldsymbol{\pi}^\star} \le \frac{\left\|\boldsymbol{\pi}^{\star\top}\boldsymbol{E}\right\|_{\boldsymbol{\pi}^\star}}{1-\max\left\{\lambda_2(\boldsymbol{P}^\star),-\lambda_n(\boldsymbol{P}^\star)\right\}-\|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star}}$$

provided that

$$1-\max\left\{\lambda_2(\boldsymbol{P}^\star),-\lambda_n(\boldsymbol{P}^\star)\right\}-\|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star}>0 \tag{5.5}$$

Analysis of spectral ranking

Apply our perturbation bound to see

$$\|\boldsymbol{\pi} - \boldsymbol{\pi}^\star\|_{\boldsymbol{\pi}^\star} \leq \frac{\left\|\boldsymbol{\pi}^{\star\top}\boldsymbol{E}\right\|_{\boldsymbol{\pi}^\star}}{1 - \max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\} - \|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star}}$$

provided that

$$1 - \max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\} - \|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star} > 0 \qquad (5.5)$$

— *need to understand spectral gap and noise size*

Spectral gap of Markov chain

Define condition number

$$\kappa := \frac{\max_{1\leq i\leq n} w_i^\star}{\min_{1\leq i\leq n} w_i^\star}$$

Lemma 5.3

It follows that

$$1 - \max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\} \geq \frac{1}{2\kappa^2}.$$

- We omit the proof; it's based on comparison between two reversible Markov chains

Bound $\|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star}$

Recall that $\boldsymbol{E} := \boldsymbol{P} - \boldsymbol{P}^\star$

Lemma 5.4

With probability at least $1 - O(n^{-8})$,

$$\|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star} \leq \sqrt{\kappa}\,\|\boldsymbol{E}\| \lesssim \sqrt{\frac{\kappa \log n}{n}}.$$

Analysis of spectral ranking (cont.)

Recall perturbation bound

$$\|\boldsymbol{\pi} - \boldsymbol{\pi}^\star\|_{\boldsymbol{\pi}^\star} \leq \frac{\left\|\boldsymbol{\pi}^{\star\top}\boldsymbol{E}\right\|_{\boldsymbol{\pi}^\star}}{1 - \max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\} - \left\|\boldsymbol{E}\right\|_{\boldsymbol{\pi}^\star}}$$
$$\leq 4\kappa^2\left\|\boldsymbol{\pi}^{\star\top}\boldsymbol{E}\right\|_{\boldsymbol{\pi}^\star} \qquad (\text{provided that } n \gg \kappa^5 \log n)$$

Note that for any $\boldsymbol{v}$, one has

$$\|\boldsymbol{v}\|_{\boldsymbol{\pi}^\star} \leq \sqrt{\pi^\star_{\max}}\,\|\boldsymbol{v}\|_2, \qquad \text{and} \qquad \|\boldsymbol{v}\|_2 \leq \frac{1}{\sqrt{\pi^\star_{\min}}}\,\|\boldsymbol{v}\|_{\boldsymbol{\pi}^\star}$$

As a result, one has

$$\|\boldsymbol{\pi} - \boldsymbol{\pi}^\star\|_2 \leq \frac{1}{\sqrt{\pi^\star_{\min}}}\|\boldsymbol{\pi} - \boldsymbol{\pi}^\star\|_{\boldsymbol{\pi}^\star} \leq \frac{4\kappa^2}{\sqrt{\pi^\star_{\min}}}\|\boldsymbol{\pi}^{\star\top}\boldsymbol{E}\|_{\boldsymbol{\pi}^\star}$$
$$\leq 4\kappa^{2.5}\|\boldsymbol{\pi}^{\star\top}\boldsymbol{E}\|_2 \leq 4\kappa^{2.5}\|\boldsymbol{E}\|\,\|\boldsymbol{\pi}^\star\|_2$$

Proof of Lemma 5.4

By construction of $\boldsymbol{P}$ and $\boldsymbol{P}^\star$, we see that

$$E_{i,j} = P_{i,j} - P^\star_{i,j} = \frac{1}{n}\big(y_{i,j} - \mathbb{E}[y_{i,j}]\big) \tag{5.6}$$

for any $i \neq j$. In addition, for all $1 \leq i \leq n$, it follows that

$$E_{i,i} = P_{i,i} - P^\star_{i,i} = -\sum_{j:j\neq i} E_{i,j} = -\frac{1}{n}\sum_{j:j\neq i}\big(y_{i,j} - \mathbb{E}[y_{i,j}]\big). \tag{5.7}$$

We shall decompose the matrix $\boldsymbol{E}$ into three parts: upper triangular part, diagonal part, and lower triangular part:

$$\|\boldsymbol{E}\| \leq \|\boldsymbol{E}_{\mathsf{upper}}\| + \|\boldsymbol{E}_{\mathsf{diag}}\| + \|\boldsymbol{E}_{\mathsf{lower}}\| \tag{5.8}$$

— *we will upper bound* $\|\boldsymbol{E}_{\mathsf{upper}}\|$

Control $\|\boldsymbol{E}_{\mathsf{diag}}\|$

Note that

$$\|\boldsymbol{E}_{\mathsf{diag}}\| = \max_{1\leq i\leq n} |E_{i,i}| = \max_{1\leq i\leq n} \frac{1}{n}\Big| \underbrace{\sum_{j:j\neq i} \big(y_{i,j} - \mathbb{E}[y_{i,j}]\,\big)}_{=: X_j}\Big|$$

- First, we have $|X_j| \leq 1 =: B$
- Second, one has

$$\sum_{j:j\neq i} \mathbb{E}[X_j^2] = \sum_{j:j\neq i} \mathsf{Var}(y_{i,j}) \leq n =: v$$

By Bernstein's inequality and union bound, we have w.h.p.

$$\max_i |E_{i,i}| \lesssim \frac{1}{n} \cdot (\sqrt{v \log n} + B \log n) \asymp \sqrt{\frac{\log n}{n}}$$

Control $\|\boldsymbol{E}_{\mathsf{upper}}\|$

First of all, we have

$$\boldsymbol{E}_{\mathsf{upper}} = \sum_{i<j} E_{i,j}\boldsymbol{e}_i\boldsymbol{e}_j^\top = \sum_{i<j} \underbrace{\tfrac{1}{n}\big(y_{i,j} - \mathbb{E}[y_{i,j}]\big)\boldsymbol{e}_i\boldsymbol{e}_j^\top}_{=:\boldsymbol{X}_{i,j}}$$

Then

- $\|\boldsymbol{X}_{i,j}\| \leq \frac{1}{n} =: B$
- Since $\mathsf{Var}(y_{i,j}) \leq 1$, one has $\mathbb{E}\left[\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\right] \preceq \frac{1}{n^2}\boldsymbol{e}_i\boldsymbol{e}_i^\top$, which gives

$$\sum\nolimits_{i<j} \mathbb{E}\left[\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\right] \preceq \sum\nolimits_{i<j} \frac{1}{n^2}\boldsymbol{e}_i\boldsymbol{e}_i^\top \preceq \frac{1}{n}\,\boldsymbol{I}_n$$

Similarly, $\sum_{i<j} \mathbb{E}\left[\boldsymbol{X}_{i,j}^\top\boldsymbol{X}_{i,j}\right] \preceq \frac{1}{n}\,\boldsymbol{I}_n$. As a result,

$$v := \max\left\{\Big\|\sum\nolimits_{i,j} \mathbb{E}\left[\boldsymbol{X}_{i,j}\boldsymbol{X}_{i,j}^\top\right]\Big\|, \Big\|\sum\nolimits_{i,j} \mathbb{E}\left[\boldsymbol{X}_{i,j}^\top\boldsymbol{X}_{i,j}\right]\Big\|\right\} \leq \frac{1}{n}$$

Control $\|\boldsymbol{E}_{\mathsf{upper}}\|$ (cont.)

Invoke matrix Bernstein to obtain

$$\|\boldsymbol{E}_{\mathsf{upper}}\| \lesssim \sqrt{v \log n} + B \log n \asymp \sqrt{\frac{\log n}{n}}$$

— *same bound holds for* $\|\boldsymbol{E}_{\mathsf{lower}}\|$

Putting pieces together

Assuming $\kappa = O(1)$, we have

$$\|\boldsymbol{\pi} - \boldsymbol{\pi}^\star\|_2 \lesssim \sqrt{\frac{\log n}{n}} \|\boldsymbol{\pi}^\star\|_2$$

- vanishing relative error when n goes to infinity
- optimal error up to a log factor

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