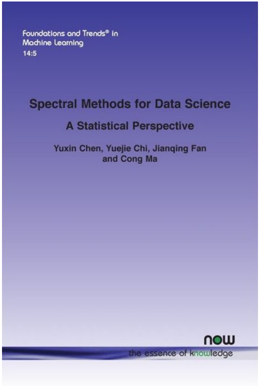
Spectral methods

Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

Outline

- A motivating application: graph clustering
- Distance and angles between two subspaces
- ℓ_2 eigen-space perturbation theory
- Extension: perturbation theory for singular subspaces
- Extension: eigen-space perturbation for asymmetric transition matrices



Spectral methods for data science: a statistical perspective

— Y. Chen, Y. Chi, J. Fan, C. Ma '21

A motivating application: graph clustering

Graph clustering / community detection

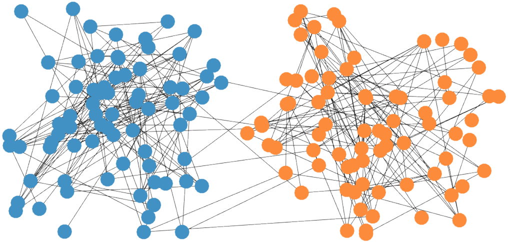
Community structures are common in many social networks

figure credit: The Future Buzz

figure credit: S. Papadopoulos

Goal: partition users into several clusters based on their friendships / similarities

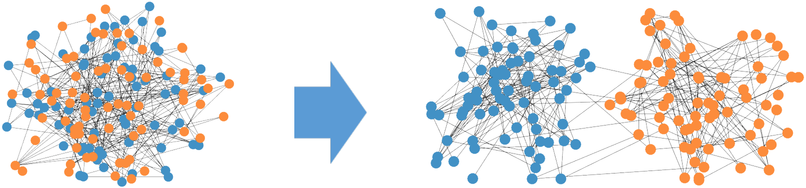
A simple model: stochastic block model (SBM)



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

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

A simple model: stochastic block model (SBM)



$\mathcal{G}$

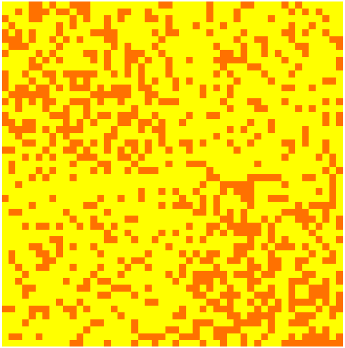
- observe a graph $\mathcal{G}$

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

  Here, $p > q$ and $p, q \gtrsim \log n / n$

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

Adjacency matrix

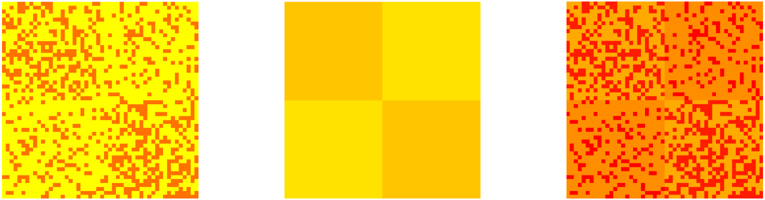


Consider the adjacency matrix $\boldsymbol{A} \in \{0,1\}^{n \times n}$ of $\mathcal{G}$:

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

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

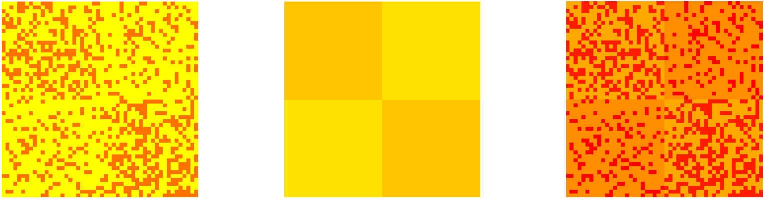
Adjacency matrix



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

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

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

Spectral clustering

Rationale: recovery is reliable if $\underbrace{\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]}_{\text{perturbation}}$ is sufficiently small

- if $\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}] = \boldsymbol{0}$, then

$$\hat{\boldsymbol{u}} \propto \pm \begin{bmatrix} \boldsymbol{1} \\ -\boldsymbol{1} \end{bmatrix} \quad \Longrightarrow \quad \text{perfect clustering}$$

Question: how to quantify the effect of perturbation $\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$ on $\hat{\boldsymbol{u}}$?

Distance and angles between two subspaces

Setup and notation

Consider 2 symmetric matrices $\boldsymbol{M}$, $\hat{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{H} \in \mathbb{R}^{n \times n}$ with eigen-decompositions

$$\boldsymbol{M} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top \qquad \text{and} \qquad \hat{\boldsymbol{M}} = \sum_{i=1}^{n} \hat{\lambda}_i \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i^\top$$

where $\lambda_1 \geq \cdots \geq \lambda_n$; $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$. For simplicity, write

$$\boldsymbol{M} = [\boldsymbol{U}_0, \boldsymbol{U}_1] \begin{bmatrix} \boldsymbol{\Lambda}_0 & \\ & \boldsymbol{\Lambda}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_0^\top \\ \boldsymbol{U}_1^\top \end{bmatrix}$$

$$\hat{\boldsymbol{M}} = [\hat{\boldsymbol{U}}_0, \hat{\boldsymbol{U}}_1] \begin{bmatrix} \hat{\boldsymbol{\Lambda}}_0 & \\ & \hat{\boldsymbol{\Lambda}}_1 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{U}}_0^\top \\ \hat{\boldsymbol{U}}_1^\top \end{bmatrix}$$

Here, $\boldsymbol{U}_0 = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$, $\boldsymbol{\Lambda}_0 = \mathrm{diag}\left([\lambda_1, \cdots, \lambda_r]\right)$, $\cdots$

Setup and notation

$$M = [\underbrace{u_1 \;\cdots\; u_r}_{U_0} \; \underbrace{u_{r+1} \;\cdots\; u_n}_{U_1}] \cdot \begin{bmatrix} \underbrace{\begin{matrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{matrix}}_{\Lambda_0} & \\ & \underbrace{\begin{matrix} \lambda_{r+1} & & \\ & \ddots & \\ & & \lambda_n \end{matrix}}_{\Lambda_1} \end{bmatrix} \begin{bmatrix} u_1^\top \\ \vdots \\ u_r^\top \\ u_{r+1}^\top \\ \vdots \\ u_n^\top \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} U_0^\top \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} U_1^\top \end{matrix}$$

Setup and notation

- $\|\boldsymbol{M}\|$: spectral norm (largest singular value of $\boldsymbol{M}$)

- $\|\boldsymbol{M}\|_{\mathrm{F}}$: Frobenius norm ($\|\boldsymbol{M}\|_{\mathrm{F}} = \sqrt{\mathsf{tr}(\boldsymbol{M}^\top \boldsymbol{M})} = \sqrt{\sum_{i,j} M_{i,j}^2}$)

Eigen-space perturbation theory

Main focus: how does the perturbation $\boldsymbol{H}$ affect the distance between $\boldsymbol{U}$ and $\hat{\boldsymbol{U}}$?

Question #0: how to define distance between two subspaces?

- $\|\boldsymbol{U} - \hat{\boldsymbol{U}}\|_{\mathrm{F}}$ and $\|\boldsymbol{U} - \hat{\boldsymbol{U}}\|$ are not appropriate, since they fall short of accounting for $\underbrace{\text{global orthonormal transformation}}_{\forall \text{ orthonormal } \boldsymbol{R} \in \mathbb{R}^{r\times r},\ \boldsymbol{U} \text{ and } \boldsymbol{U}\boldsymbol{R} \text{ represent same subspace}}$

Distance between two eigen-spaces

One metric that takes care of global orthonormal transformation is

$$\mathrm{dist}(\boldsymbol{X}_0, \boldsymbol{Z}_0) := \|\boldsymbol{X}_0 \boldsymbol{X}_0^\top - \boldsymbol{Z}_0 \boldsymbol{Z}_0^\top\| \tag{2.1}$$

This metric has several equivalent expressions:

Lemma 2.1

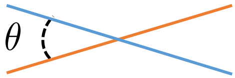
Suppose $\boldsymbol{X} := [\boldsymbol{X}_0, \underbrace{\boldsymbol{X}_1}_{\text{complement subspace}}]$ *and* $\boldsymbol{Z} := [\boldsymbol{Z}_0, \underbrace{\boldsymbol{Z}_1}_{\text{complement subspace}}]$ *are square orthonormal matrices. Then*

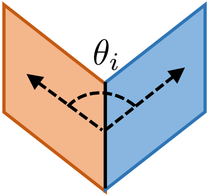
$$\mathrm{dist}(\boldsymbol{X}_0, \boldsymbol{Z}_0) = \|\boldsymbol{X}_0^\top \boldsymbol{Z}_1\| = \|\boldsymbol{Z}_0^\top \boldsymbol{X}_1\|$$

- sanity check: if $\boldsymbol{X}_0 = \boldsymbol{Z}_0$, then $\mathrm{dist}(\boldsymbol{X}_0, \boldsymbol{Z}_0) = \|\boldsymbol{X}_0^\top \boldsymbol{Z}_1\| = 0$
- proof: see Slide 2-22

Principal angles between two eigen-spaces

In addition to "distance", one might also be interested in "angles"





We can quantify the similarity between two lines (represented resp. by unit vectors $\boldsymbol{x}_0$ and $\boldsymbol{z}_0$) by an angle between them

$$\theta = \arccos\langle \boldsymbol{x}_0, \boldsymbol{z}_0 \rangle$$

Principal angles between two eigen-spaces

For r-dimensional subspaces, one needs r angles

Specifically, given $\|\boldsymbol{X}_0^\top \boldsymbol{Z}_0\| \le 1$, we write the singular value decomposition (SVD) of $\boldsymbol{X}_0^\top \boldsymbol{Z}_0 \in \mathbb{R}^{r \times r}$ as

$$\boldsymbol{X}_0^\top \boldsymbol{Z}_0 = \boldsymbol{U} \underbrace{\begin{bmatrix} \cos\theta_1 & & \\ & \ddots & \\ & & \cos\theta_r \end{bmatrix}}_{=: \cos \boldsymbol{\Theta}} \boldsymbol{V}^\top =: \boldsymbol{U} \cos \boldsymbol{\Theta}\, \boldsymbol{V}^\top$$

where $\{\theta_1, \cdots, \theta_r\}$ are called the principal angles between $\boldsymbol{X}_0$ and $\boldsymbol{Z}_0$

Relations between principal angles and $\mathsf{dist}(\cdot, \cdot)$

As expected, principal angles and distances are closely related

Lemma 2.2

Suppose $\boldsymbol{X} := [\boldsymbol{X}_0, \boldsymbol{X}_1]$ *and* $\boldsymbol{Z} := [\boldsymbol{Z}_0, \boldsymbol{Z}_1]$ *are square orthonormal matrices. Then*

$$\|\boldsymbol{X}_0^\top \boldsymbol{Z}_1\| = \|\sin \boldsymbol{\Theta}\| = \max\{|\sin\theta_1|, \cdots, |\sin\theta_r|\}$$

Lemmas 2.1 and 2.2 taken collectively give

$$\mathsf{dist}(\boldsymbol{X}_0, \boldsymbol{Z}_0) = \max\{|\sin\theta_1|, \cdots, |\sin\theta_r|\} \tag{2.2}$$

Proof of Lemma 2.2

$$
\begin{aligned}
\|\boldsymbol{X}_0^\top \boldsymbol{Z}_1\| &= \|\boldsymbol{X}_0^\top \underbrace{\boldsymbol{Z}_1 \boldsymbol{Z}_1^\top}_{=\boldsymbol{I}-\boldsymbol{Z}_0\boldsymbol{Z}_0^\top} \boldsymbol{X}_0\|^{\frac{1}{2}} \\
&= \|\boldsymbol{X}_0^\top \boldsymbol{X}_0 - \boldsymbol{X}_0^\top \boldsymbol{Z}_0 \boldsymbol{Z}_0^\top \boldsymbol{X}_0\|^{\frac{1}{2}} \\
&= \|\boldsymbol{I} - \boldsymbol{U} \cos^2 \boldsymbol{\Theta}\, \boldsymbol{U}^\top\|^{\frac{1}{2}} \qquad \text{(since } \boldsymbol{X}_0^\top \boldsymbol{Z}_0 = \boldsymbol{U} \cos \boldsymbol{\Theta}\, \boldsymbol{V}^\top\text{)} \\
&= \|\boldsymbol{I} - \cos^2 \boldsymbol{\Theta}\|^{\frac{1}{2}} \\
&= \|\sin \boldsymbol{\Theta}^2\|^{\frac{1}{2}} \\
&= \|\sin \boldsymbol{\Theta}\|
\end{aligned}
$$

Proof of Lemma 2.1

We first claim that the SVD of $\boldsymbol{X}_1^\top \boldsymbol{Z}_0$ can be written as

$$\boldsymbol{X}_1^\top \boldsymbol{Z}_0 = \tilde{\boldsymbol{U}} \sin \boldsymbol{\Theta}\, \boldsymbol{V}^\top \tag{2.3}$$

for some orthonormal $\tilde{\boldsymbol{U}}$ (to be proved later). With this claim in place, one has

$$\boldsymbol{Z}_0 = [\boldsymbol{X}_0, \boldsymbol{X}_1] \begin{bmatrix} \boldsymbol{X}_0^\top \\ \boldsymbol{X}_1^\top \end{bmatrix} \boldsymbol{Z}_0 = [\boldsymbol{X}_0, \boldsymbol{X}_1] \begin{bmatrix} \boldsymbol{U} \cos \boldsymbol{\Theta}\, \boldsymbol{V}^\top \\ \tilde{\boldsymbol{U}} \sin \boldsymbol{\Theta}\, \boldsymbol{V}^\top \end{bmatrix}$$

$$\Longrightarrow \quad \boldsymbol{Z}_0 \boldsymbol{Z}_0^\top = [\boldsymbol{X}_0, \boldsymbol{X}_1] \begin{bmatrix} \boldsymbol{U} \cos^2 \boldsymbol{\Theta}\, \boldsymbol{U}^\top & \boldsymbol{U} \cos \boldsymbol{\Theta} \sin \boldsymbol{\Theta}\, \tilde{\boldsymbol{U}}^\top \\ \tilde{\boldsymbol{U}} \cos \boldsymbol{\Theta} \sin \boldsymbol{\Theta}\, \boldsymbol{U}^\top & \tilde{\boldsymbol{U}} \sin^2 \boldsymbol{\Theta}\, \tilde{\boldsymbol{U}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_0^\top \\ \boldsymbol{X}_1^\top \end{bmatrix}$$

As a consequence,

$$\begin{aligned} &\boldsymbol{X}_0 \boldsymbol{X}_0^\top - \boldsymbol{Z}_0 \boldsymbol{Z}_0^\top \\ &\quad = [\boldsymbol{X}_0, \boldsymbol{X}_1] \begin{bmatrix} \boldsymbol{I} - \boldsymbol{U} \cos^2 \boldsymbol{\Theta}\, \boldsymbol{U}^\top & -\boldsymbol{U} \cos \boldsymbol{\Theta} \sin \boldsymbol{\Theta}\, \tilde{\boldsymbol{U}}^\top \\ -\tilde{\boldsymbol{U}} \cos \boldsymbol{\Theta} \sin \boldsymbol{\Theta}\, \boldsymbol{U}^\top & -\tilde{\boldsymbol{U}} \sin^2 \boldsymbol{\Theta}\, \tilde{\boldsymbol{U}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_0^\top \\ \boldsymbol{X}_1^\top \end{bmatrix} \end{aligned}$$

Proof of Lemma 2.1 (cont.)

This further gives

$$
\begin{aligned}
&\left\| \boldsymbol{X}_0 \boldsymbol{X}_0^\top - \boldsymbol{Z}_0 \boldsymbol{Z}_0^\top \right\| \\
&= \left\| \begin{bmatrix} \boldsymbol{U} & \\ & \tilde{\boldsymbol{U}} \end{bmatrix} \begin{bmatrix} \sin^2 \boldsymbol{\Theta} & -\cos \boldsymbol{\Theta} \sin \boldsymbol{\Theta} \\ -\cos \boldsymbol{\Theta} \sin \boldsymbol{\Theta} & -\sin^2 \boldsymbol{\Theta} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^\top & \\ & \tilde{\boldsymbol{U}}^\top \end{bmatrix} \right\| \\
&= \Bigg\| \underbrace{\begin{bmatrix} \sin^2 \boldsymbol{\Theta} & -\cos \boldsymbol{\Theta} \sin \boldsymbol{\Theta} \\ -\cos \boldsymbol{\Theta} \sin \boldsymbol{\Theta} & -\sin^2 \boldsymbol{\Theta} \end{bmatrix}}_{\text{each block is a diagonal matrix}} \Bigg\| \qquad (\|\cdot\| \text{ is rotationally invariant}) \\
&= \max_{1 \le i \le r} \left\| \begin{bmatrix} \sin^2 \theta_i & -\cos \theta_i \sin \theta_i \\ -\cos \theta_i \sin \theta_i & -\sin^2 \theta_i \end{bmatrix} \right\| \\
&= \max_{1 \le i \le r} \left\| \sin \theta_i \begin{bmatrix} \sin \theta_i & -\cos \theta_i \\ -\cos \theta_i & -\sin \theta_i \end{bmatrix} \right\| \\
&= \max_{1 \le i \le r} |\sin \theta_i| \; = \; \|\sin \boldsymbol{\Theta}\|
\end{aligned}
$$

Proof of Lemma 2.1 (cont.)

It remains to justify (2.3). To this end, observe that

$$\begin{aligned} \boldsymbol{Z}_0^\top \boldsymbol{X}_1 \boldsymbol{X}_1^\top \boldsymbol{Z}_0 &= \boldsymbol{Z}_0^\top \boldsymbol{Z}_0 - \boldsymbol{Z}_0^\top \boldsymbol{X}_0 \boldsymbol{X}_0^\top \boldsymbol{Z}_0 \\ &= \boldsymbol{I} - \boldsymbol{V} \cos^2 \boldsymbol{\Theta}\, \boldsymbol{V}^\top \\ &= \boldsymbol{V} \sin^2 \boldsymbol{\Theta}\, \boldsymbol{V}^\top \end{aligned}$$

and hence the right singular space (resp. singular values) of $\boldsymbol{X}_1^\top \boldsymbol{Z}_0$ is given by $\boldsymbol{V}$ (resp. $\sin \boldsymbol{\Theta}$). This immediately implies (2.3).

Eigen-space perturbation theory

Davis-Kahan $\sin\Theta$ Theorem: a simple case

— *recall the setup in Page 2-13*

Chandler Davis

William Kahan

Theorem 2.3

Suppose $\boldsymbol{M} \succeq \mathbf{0}$ *and has rank* r. *If* $\|\boldsymbol{H}\| < \lambda_r(\boldsymbol{M})$, *then*

$$\mathsf{dist}\big(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0\big) \leq \frac{\|\boldsymbol{H}\boldsymbol{U}_0\|}{\lambda_r(\boldsymbol{M}) - \|\boldsymbol{H}\|} \leq \frac{\|\boldsymbol{H}\|}{\lambda_r(\boldsymbol{M}) - \|\boldsymbol{H}\|}$$

- depends on $\underbrace{\text{smallest non-zero eigenvalue of } \boldsymbol{M}}_{\text{eigen-gap between } \lambda_r(\boldsymbol{M}) \text{ and } \lambda_{r+1}(\boldsymbol{M})}$ and perturbation size

Proof of Theorem 2.3

We intend to control $\hat{\boldsymbol{U}}_1^\top \boldsymbol{U}_0$ by studying their interactions through $\boldsymbol{H}$:

$$\begin{aligned}
\left\|\hat{\boldsymbol{U}}_1^\top \boldsymbol{H} \boldsymbol{U}_0\right\| &= \left\|\hat{\boldsymbol{U}}_1^\top \Big(\underbrace{\hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{U}}^\top}_{\boldsymbol{M}+\boldsymbol{H}} - \underbrace{\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top}_{\boldsymbol{M}}\Big)\boldsymbol{U}_0\right\| \\
&= \left\|\hat{\boldsymbol{\Lambda}}_1 \hat{\boldsymbol{U}}_1^\top \boldsymbol{U}_0 - \hat{\boldsymbol{U}}_1^\top \boldsymbol{U}_0 \boldsymbol{\Lambda}_0\right\| \qquad (\text{since } \boldsymbol{U}_1^\top \boldsymbol{U}_0 = \hat{\boldsymbol{U}}_1^\top \hat{\boldsymbol{U}}_0 = \boldsymbol{0}) \\
&\geq \left\|\hat{\boldsymbol{U}}_1^\top \boldsymbol{U}_0 \boldsymbol{\Lambda}_0\right\| - \left\|\hat{\boldsymbol{\Lambda}}_1 \hat{\boldsymbol{U}}_1^\top \boldsymbol{U}_0\right\| \qquad (\text{triangle inequality}) \\
&\geq \left\|\hat{\boldsymbol{U}}_1^\top \boldsymbol{U}_0\right\| \lambda_r - \left\|\hat{\boldsymbol{U}}_1^\top \boldsymbol{U}_0\right\| \|\hat{\boldsymbol{\Lambda}}_1\| \qquad (2.4)
\end{aligned}$$

In view of Weyl's Theorem, $\|\hat{\boldsymbol{\Lambda}}_1\| \leq \|\boldsymbol{H}\|$, which combined with (2.4) gives

$$\left\|\hat{\boldsymbol{U}}_1^\top \boldsymbol{U}_0\right\| \leq \frac{\left\|\hat{\boldsymbol{U}}_1^\top \boldsymbol{H}\boldsymbol{U}_0\right\|}{\lambda_r - \|\boldsymbol{H}\|} \leq \frac{\|\hat{\boldsymbol{U}}_1\| \cdot \left\|\boldsymbol{H}\boldsymbol{U}_0\right\|}{\lambda_r - \|\boldsymbol{H}\|} = \frac{\left\|\boldsymbol{H}\boldsymbol{U}_0\right\|}{\lambda_r - \|\boldsymbol{H}\|}$$

This together with Lemma 2.1 completes the proof

Davis-Kahan $\sin\Theta$ Theorem: more general case

Theorem 2.4 (Davis-Kahan $\sin\Theta$ Theorem)

Suppose $\lambda_r(\boldsymbol{M}) \geq a$ *and* $\lambda_{r+1}(\hat{\boldsymbol{M}}) \leq a - \Delta$ *for some* $\Delta > 0$. *Then*

$$\mathrm{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0) \leq \frac{\|\boldsymbol{H}\boldsymbol{U}_0\|}{\Delta} \leq \frac{\|\boldsymbol{H}\|}{\Delta}$$

- immediate consequence: if $\lambda_r(\boldsymbol{M}) > \lambda_{r+1}(\boldsymbol{M}) + \|\boldsymbol{H}\|$, then

$$\mathrm{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0) \leq \frac{\|\boldsymbol{H}\|}{\underbrace{\lambda_r(\boldsymbol{M}) - \lambda_{r+1}(\boldsymbol{M})}_{\text{spectral gap}} - \|\boldsymbol{H}\|} \tag{2.5}$$

Back to stochastic block model ...

Let $\boldsymbol{M} = \underbrace{\mathbb{E}[\boldsymbol{A}] - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top}_{=\frac{p-q}{2}\left[\begin{array}{c}\mathbf{1}\\ -\mathbf{1}\end{array}\right]\left[\mathbf{1}^\top, -\mathbf{1}^\top\right]}$, $\hat{\boldsymbol{M}} = \boldsymbol{A} - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top$ and $\boldsymbol{u} = \frac{1}{\sqrt{n}}\left[\begin{array}{c}\mathbf{1}\\ -\mathbf{1}\end{array}\right]$

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

$$\mathrm{dist}(\hat{\boldsymbol{u}}, \boldsymbol{u}) \leq \frac{\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|}{\lambda_1(\boldsymbol{M}) - \|\hat{\boldsymbol{M}} - \boldsymbol{M}\|} = \frac{\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|}{\frac{(p-q)n}{2} - \|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|} \tag{2.6}$$

Question: how to bound $\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|$?

A hammer: matrix Bernstein inequality

Consider a sequence of independent random matrices $\{\boldsymbol{X}_l \in \mathbb{R}^{d_1 \times d_2}\}$

- $\mathbb{E}[\boldsymbol{X}_l] = \boldsymbol{0}$
- $\|\boldsymbol{X}_l\| \leq B$ for each l
- variance statistic:

$$v := \max\left\{\left\|\mathbb{E}\left[\sum_l \boldsymbol{X}_l \boldsymbol{X}_l^\top\right]\right\|, \left\|\mathbb{E}\left[\sum_l \boldsymbol{X}_l^\top \boldsymbol{X}_l\right]\right\|\right\}$$

Theorem 2.5 (Matrix Bernstein inequality)

For all $\tau \geq 0$,

$$\mathbb{P}\left\{\left\|\sum_l \boldsymbol{X}_l\right\| \geq \tau\right\} \leq (d_1 + d_2) \exp\left(\frac{-\tau^2/2}{v + B\tau/3}\right)$$

A hammer: matrix Bernstein inequality

$$\mathbb{P}\left\{\left\|\sum_l \boldsymbol{X}_l\right\| \geq \tau\right\} \leq (d_1 + d_2) \exp\left(\frac{-\tau^2/2}{v + B\tau/3}\right)$$

- **moderate-deviation regime** (τ is small):
  — sub-Gaussian tail behavior $\exp(-\tau^2/2v)$

- **large-deviation regime** (τ is large):
  — sub-exponential tail behavior $\exp(-3\tau/2B)$ (slower decay)

- **user-friendly form** (exercise): with prob. $1 - O((d_1 + d_2)^{-10})$

$$\left\|\sum_l \boldsymbol{X}_l\right\| \lesssim \sqrt{v \log(d_1 + d_2)} + B \log(d_1 + d_2) \qquad (2.7)$$

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

The matrix Bernstein inequality yields

Lemma 2.6

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

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| \lesssim \sqrt{np \log n} \tag{2.8}$$

Statistical accuracy of spectral clustering

Substitute (2.8) into (2.6) to reach

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

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

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

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 2.6

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 2.6

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.,

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

Extension: singular subspaces

Singular value decomposition

Consider two matrices $\boldsymbol{M}, \hat{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{H} \in \mathbb{R}^{n_1 \times n_2}$ with SVD

$$\boldsymbol{M} = [\boldsymbol{U}_0, \boldsymbol{U}_1] \begin{bmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_1 \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_0^\top \\ \boldsymbol{V}_1^\top \end{bmatrix}$$

$$\hat{\boldsymbol{M}} = \left[\hat{\boldsymbol{U}}_0, \hat{\boldsymbol{U}}_1\right] \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_0 & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{\Sigma}}_1 \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{V}}_0^\top \\ \hat{\boldsymbol{V}}_1^\top \end{bmatrix}$$

where $\boldsymbol{U}_0$ (resp. $\hat{\boldsymbol{U}}_0$) and $\boldsymbol{V}_0$ (resp. $\hat{\boldsymbol{V}}_0$) represent the top-r singular subspaces of $\boldsymbol{M}$ (resp. $\hat{\boldsymbol{M}}$)

Wedin $\sin\Theta$ Theorem

The Davis-Kahan Theorem generalizes to singular subspace perturbation:

Theorem 2.7 (Wedin $\sin\Theta$ Theorem)

Suppose $\underbrace{\sigma_r(\boldsymbol{M})}_{r\text{th singular value}} \geq a$ *and* $\sigma_{r+1}(\hat{\boldsymbol{M}}) \leq a - \Delta$ *for some* $\Delta > 0$. *Then*

$$\max\left\{\mathsf{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0), \mathsf{dist}(\hat{\boldsymbol{V}}_0, \boldsymbol{V}_0)\right\} \leq \underbrace{\frac{\max\left\{\|\boldsymbol{H}\boldsymbol{V}_0\|, \|\boldsymbol{H}^\top \boldsymbol{U}_0\|\right\}}{\Delta}}_{\text{two-sided interactions}}$$

$$\leq \frac{\|\boldsymbol{H}\|}{\Delta}$$

Example: low-rank matrix completion

- Netflix challenge: Netflix provides highly incomplete ratings from 0.5 million users for & 17,770 movies
- How to predict unseen user ratings for movies?

Example: low-rank matrix completion

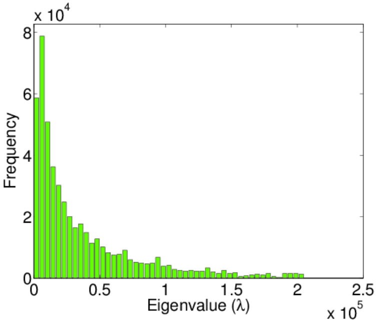
In general, we cannot infer missing ratings

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

— this is an underdetermined system (more unknowns than observations)

Example: low-rank matrix completion

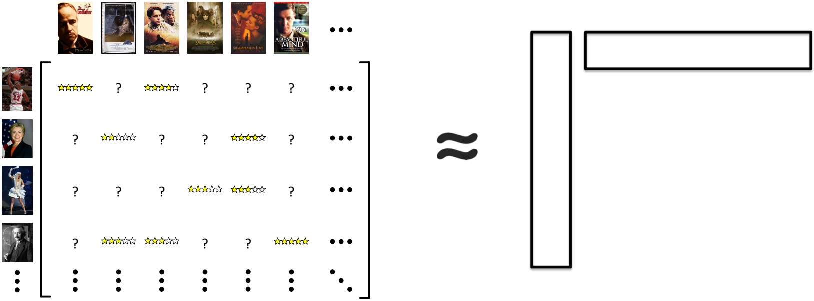
... unless rating matrix has other structure



A few factors explain most of the data

Example: low-rank matrix completion

... unless rating matrix has other structure



A few factors explain most of the data $\longrightarrow$ low-rank approximation

How to exploit (approx.) low-rank structure in prediction?

Model for low-rank matrix completion

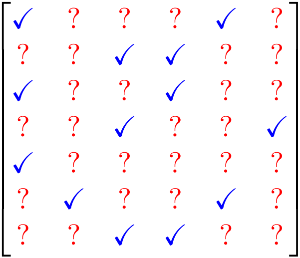


figure credit: Candès

- consider a low-rank matrix $\boldsymbol{M}$
- each entry $M_{i,j}$ is observed independently with prob. p
- **goal:** fill in missing entries

Spectral estimate for matrix completion

1. set $\hat{\boldsymbol{M}} \in \mathbb{R}^{n \times n}$ as

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

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

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

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.

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

$$\ll 1 \quad \text{(nearly accurate estimates)} \qquad (2.9)$$

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 (2.9)

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
$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \lesssim \sqrt{v \log n} + B \log n \ll 1$$

Extension: eigen-space for asymmetric transition matrices

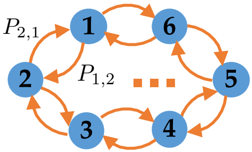
Eigen-decomposition for asymmetric matrices

Eigen-decomposition for asymmetric matrices is much more tricky:

1. both eigenvalues and eigenvectors might be complex-valued
2. eigenvectors might not be orthogonal to each other

This lecture focuses on a special case: **probability transition matrices**

Probability transition matrices



Consider a Markov chain $\{X_t\}_{t\geq 0}$

- n states
- transition probability $\mathbb{P}\{X_{t+1} = j \mid X_t = i\} = P_{i,j}$
- transition matrix $\boldsymbol{P} = [P_{i,j}]_{1\leq i,j\leq n}$
- stationary distribution $\underbrace{\boldsymbol{\pi} = [\pi_1, \cdots, \pi_n]}_{\pi_1+\cdots+\pi_n=1}$ is 1st eigenvector of $\boldsymbol{P}$

$$\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$$

- $\{X_t\}_{t\geq 0}$ is said to be reversible if $\pi_i P_{i,j} = \pi_j P_{j,i}$ for all i, j

Eigenvector perturbation for transition matrices

Define $\|\boldsymbol{a}\|_{\boldsymbol{\pi}} := \sqrt{\pi_1 a_1^2 + \cdots + \pi_n a_n^2}$

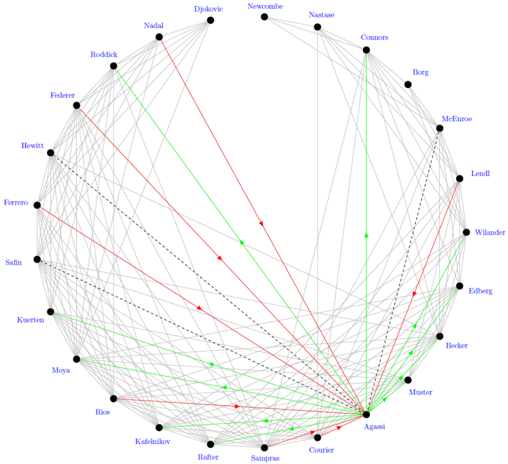
Theorem 2.8 (Chen, Fan, Ma, Wang '17)

Suppose $\boldsymbol{P}$, $\hat{\boldsymbol{P}}$ *are transition matrices with stationary distributions* $\boldsymbol{\pi}$, $\hat{\boldsymbol{\pi}}$, *respectively. Assume* $\boldsymbol{P}$ *induces a reversible Markov chain. If* $1 > \max\{\lambda_2(\boldsymbol{P}), -\lambda_n(\boldsymbol{P})\} + \|\hat{\boldsymbol{P}} - \boldsymbol{P}\|_{\boldsymbol{\pi}}$, *then*

$$\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{\boldsymbol{\pi}} \leq \frac{\|\boldsymbol{\pi}(\hat{\boldsymbol{P}} - \boldsymbol{P})\|_{\boldsymbol{\pi}}}{\underbrace{1 - \max\{\lambda_2(\boldsymbol{P}), -\lambda_n(\boldsymbol{P})\}}_{\text{spectral gap}} - \underbrace{\|\hat{\boldsymbol{P}} - \boldsymbol{P}\|_{\boldsymbol{\pi}}}_{\text{perturbation}}}$$

- $\hat{\boldsymbol{P}}$ does not need to induce a reversible Markov chain

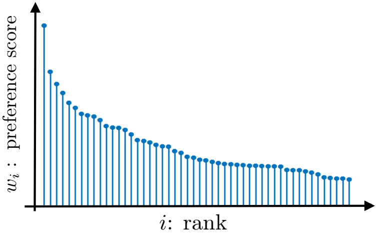
Example: ranking from pairwise comparisons



pairwise comparisons for ranking tennis players

figure credit: Bozóki, Csató, Temesi

Bradley-Terry-Luce (logistic) model



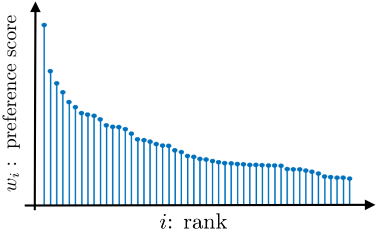
- n items to be ranked
- assign a latent score $\{w_i\}_{1 \leq i \leq n}$ to each item, so that

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

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

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

Bradley-Terry-Luce (logistic) model



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

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

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

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

Spectral ranking method

- construct a probability transition matrix $\hat{\boldsymbol{P}}$ obeying

$$\hat{P}_{i,j} = \begin{cases} \frac{1}{2n} y_{i,j}, & \text{if } i \neq j \\ 1 - \sum_{l:l\neq i} \hat{P}_{i,l}, & \text{if } i = j \end{cases}$$

- return the score estimate as the leading left eigenvector $\hat{\boldsymbol{\pi}}$ of $\hat{\boldsymbol{P}}$

— closely related to PageRank!

Rationale behind spectral method

$$\mathbb{E}[\hat{P}_{i,j}] \;=\; \frac{1}{2n}\cdot\frac{w_j}{w_i+w_j}, \qquad i \neq j$$

- $\boldsymbol{P} := \mathbb{E}[\hat{\boldsymbol{P}}]$ obeys

$$w_i P_{i,j} = w_j P_{j,i} \qquad \text{(detailed balance)}$$

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

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

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 2.8 and matrix Bernstein (exercise)

Reference

- "*Spectral methods for data science: A statistical perspective*," Y. Chen, Y. Chi, J. Fan, C. Ma, *Foundations and Trends in Machine Learning*, 2021.
- "*The rotation of eigenvectors by a perturbation*," C. Davis, W. Kahan, *SIAM Journal on Numerical Analysis*, 1970.
- "*Perturbation bounds in connection with singular value decomposition*," P. Wedin, *BIT Numerical Mathematics*, 1972.
- "*Inference, estimation, and information processing, EE 378B lecture notes*," A. Montanari, Stanford University.
- "*COMS 4772 lecture notes*," D. Hsu, Columbia University.
- "*Community detection and stochastic block models*," E. Abbe, *Foundations and Trends in Communications and Information Theory*, 2018.

Reference

- "*Consistency thresholds for the planted bisection model*," E. Mossel, J. Neeman, A. Sly, *ACM Symposium on Theory of Computing*, 2015.
- "*Matrix completion from a few entries*," R. Keshavan, A. Montanari, S. Oh, *IEEE Transactions on Information Theory*, 2010.
- "*The PageRank citation ranking: bringing order to the web*," L. Page, S. Brin, R. Motwani, T. Winograd, 1999.
- "*Rank centrality: ranking from pairwise comparisons*," S. Negahban, S. Oh, D. Shah, *Operations Research*, 2017.
- "*Spectral method and regularized MLE are both optimal for top-K ranking*," Y. Chen, J. Fan, C. Ma, K. Wang, *Annals of Statistics*, 2019.