

Spectral methods



Yuxin Chen

Princeton University, Fall 2018

Outline

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

A motivating application: graph clustering

Graph clustering / community detection

Community structures are common in many social networks

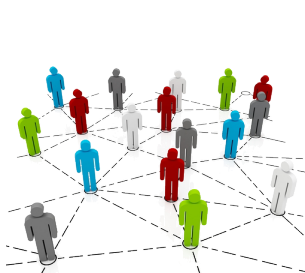


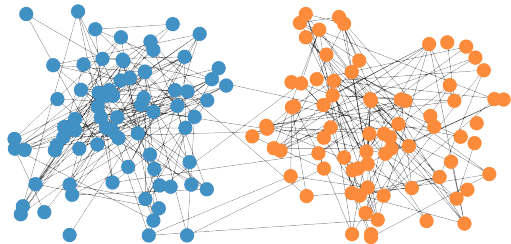
figure credit: The Future Buzz



figure credit: S. Papadopoulos

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

A simple model: stochastic block model (SBM)

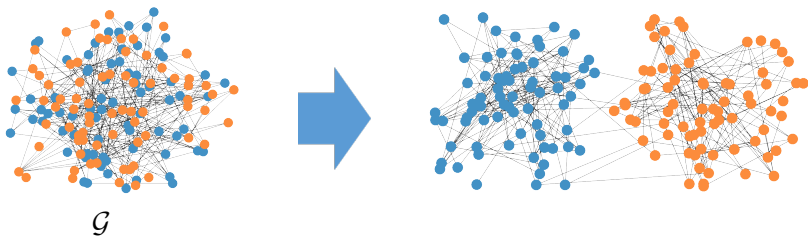


$x_i = 1$: 1st community

$x_i = -1$: 2nd community

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

A simple model: stochastic block model (SBM)



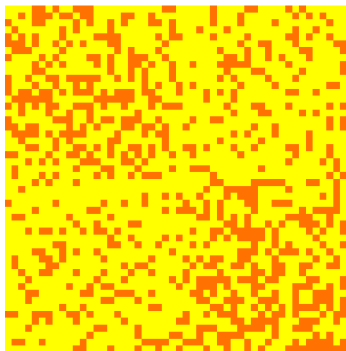
- observe a graph \mathcal{G}

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

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

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

Adjacency matrix

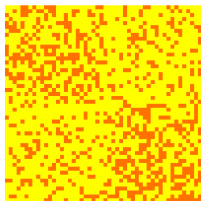
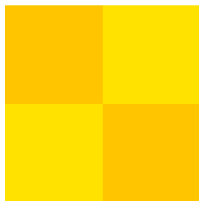


Consider the adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$ of \mathcal{G} :

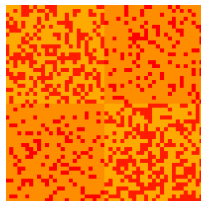
$$A_{i,j} = \begin{cases} 1, & \text{if } (i, j) \in \mathcal{G} \\ 0, & \text{else} \end{cases}$$

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

Adjacency matrix


 \mathbf{A}
 $=$

 $\mathbb{E}[\mathbf{A}]$

rank 2

 $+$

 $\mathbf{A} - \mathbb{E}[\mathbf{A}]$

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

Spectral clustering



The diagram shows three heatmaps representing matrices. The first heatmap, labeled A , is a noisy matrix with a yellow and orange pixelated pattern. The second heatmap, labeled $\mathbb{E}[A]$, is a smooth matrix with a clear 2x2 block structure in shades of yellow and orange. Below this heatmap is a blue bracket with the text "rank 2" underneath it. The third heatmap, labeled $A - \mathbb{E}[A]$, is a noisy matrix with a red and orange pixelated pattern. The equation $A = \mathbb{E}[A] + A - \mathbb{E}[A]$ is represented by the arrangement of the heatmaps and the equals sign between them.

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

1. computing the leading eigenvector $\hat{\mathbf{u}} = [\hat{u}_i]_{1 \leq i \leq n}$ of $A - \frac{p+q}{2} \mathbf{1}\mathbf{1}^\top$
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 $\underbrace{\mathbf{A} - \mathbb{E}[\mathbf{A}]}_{\text{perturbation}}$ is sufficiently small

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

$$\hat{\mathbf{u}} \propto \pm \begin{bmatrix} 1 \\ -1 \end{bmatrix} \implies \text{perfect clustering}$$

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

Distance and angles between two subspaces

Setup and notation

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

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

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

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

Here, $\mathbf{U}_0 = [\mathbf{u}_1, \dots, \mathbf{u}_r]$, $\mathbf{\Lambda}_0 = \text{diag}([\lambda_1, \dots, \lambda_r]), \dots$

Setup and notation

$$M = \left[\underbrace{u_1 \cdots u_r}_{U_0} \underbrace{u_{r+1} \cdots u_n}_{U_1} \right]$$

$$\cdot \left[\begin{array}{ccc} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \\ \underbrace{\hspace{10em}}_{\Lambda_0} & & \\ & \lambda_{r+1} & \\ & & \ddots \\ & & & \lambda_n \\ & & & \underbrace{\hspace{10em}}_{\Lambda_1} \end{array} \right] \left[\begin{array}{c} u_1^\top \\ \vdots \\ u_r^\top \\ u_{r+1}^\top \\ \vdots \\ u_n^\top \end{array} \right] \left\{ \begin{array}{l} U_0^\top \\ U_1^\top \end{array} \right.$$

Setup and notation

- $\|M\|$: spectral norm (largest singular value of M)
- $\|M\|_F$: Frobenius norm ($\|M\|_F = \sqrt{\text{tr}(M^\top M)} = \sqrt{\sum_{i,j} M_{i,j}^2}$)

Eigen-space perturbation theory

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

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

- $\|U - \hat{U}\|_F$ and $\|U - \hat{U}\|$ are not appropriate, since they fall short of accounting for global orthonormal transformation

\forall orthonormal $R \in \mathbb{R}^{r \times r}$, U and UR represent same subspace

Distance between two eigen-spaces

One metric that takes care of global orthonormal transformation is

$$\text{dist}(\mathbf{X}_0, \mathbf{Z}_0) := \|\mathbf{X}_0 \mathbf{X}_0^\top - \mathbf{Z}_0 \mathbf{Z}_0^\top\| \quad (2.1)$$

This metric has several equivalent expressions:

Lemma 2.1

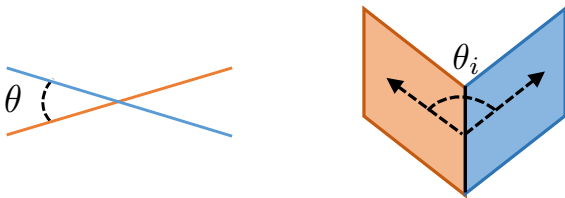
Suppose $\mathbf{X} := [\mathbf{X}_0, \underbrace{\mathbf{X}_1}_{\text{complement subspace}}]$ and $\mathbf{Z} := [\mathbf{Z}_0, \underbrace{\mathbf{Z}_1}_{\text{complement subspace}}]$ are orthonormal matrices. Then

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

- sanity check: if $\mathbf{X}_0 = \mathbf{Z}_0$, then $\text{dist}(\mathbf{X}_0, \mathbf{Z}_0) = \|\mathbf{X}_0^\top \mathbf{Z}_1\| = 0$

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 x_0 and z_0) by an angle between them

$$\theta = \arccos\langle x_0, z_0 \rangle$$

Principal angles between two eigen-spaces

For r -dimensional subspaces, one needs r angles

Specifically, given $\|X_0^\top Z_0\| \leq 1$, we write SVD of $X_0^\top Z_0 \in \mathbb{R}^{r \times r}$ as

$$X_0^\top Z_0 = U \underbrace{\begin{bmatrix} \cos \theta_1 & & \\ & \ddots & \\ & & \cos \theta_r \end{bmatrix}}_{:= \cos \Theta} V^\top := U \cos \Theta V^\top$$

where $\{\theta_1, \dots, \theta_r\}$ are called the **principal angles** between X_0 and Z_0

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

As expected, principal angles and distance are closely related

Lemma 2.2

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

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

Lemmas 2.1 and 2.2 taken collectively give

$$\text{dist}(\mathbf{X}_0, \mathbf{Z}_0) = \max\{|\sin \theta_1|, \dots, |\sin \theta_r|\} \quad (2.2)$$

Proof of Lemma 2.2

$$\begin{aligned}\|X_0^\top Z_1\| &= \|X_0^\top \underbrace{Z_1 Z_1^\top}_{=I - Z_0 Z_0^\top} X_0\|^{\frac{1}{2}} \\&= \|X_0^\top X_0 - X_0^\top Z_0 Z_0^\top X_0\|^{\frac{1}{2}} \\&= \|I - U \cos^2 \Theta U^\top\|^{\frac{1}{2}} \quad (\text{since } X_0^\top Z_0 = U \cos \Theta V^\top) \\&= \|I - \cos^2 \Theta\|^{\frac{1}{2}} \\&= \|\sin \Theta\|\end{aligned}$$

Proof of Lemma 2.1

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

$$\mathbf{X}_1^\top \mathbf{Z}_0 = \tilde{\mathbf{U}} \sin \Theta \mathbf{V}^\top \quad (2.3)$$

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

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

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

As a consequence,

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

Proof of Lemma 2.1 (cont.)

This further gives

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

Proof of Lemma 2.1 (cont.)

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

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

and hence right singular space (resp. singular values) of $\mathbf{X}_1^\top \mathbf{Z}_0$ is given by \mathbf{V} (resp. $\sin \Theta$)

Eigen-space perturbation theory

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



Chandler Davis



William Kahan

Theorem 2.3

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

$$\text{dist}(\hat{U}_0, U_0) \leq \frac{\|HU_0\|}{\lambda_r(M) - \|H\|} \leq \frac{\|H\|}{\lambda_r(M) - \|H\|}$$

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

Proof of Theorem 2.3

We intend to control $\hat{U}_1^\top U_0$ by studying their interactions through \mathbf{H} :

$$\begin{aligned}
 \|\hat{U}_1^\top \mathbf{H} U_0\| &= \left\| \hat{U}_1^\top \left(\underbrace{\hat{U} \hat{\Lambda} \hat{U}^\top}_{M+H} - \underbrace{U \Lambda U^\top}_M \right) U_0 \right\| \\
 &= \left\| \hat{\Lambda}_1 \hat{U}_1^\top U_0 - \hat{U}_1^\top U_0 \Lambda_0 \right\| && \text{(since } U_1^\top U_0 = \hat{U}_1^\top \hat{U}_0 = \mathbf{0} \text{)} \\
 &\geq \left\| \hat{U}_1^\top U_0 \Lambda_0 \right\| - \left\| \hat{\Lambda}_1 \hat{U}_1^\top U_0 \right\| && \text{(triangle inequality)} \\
 &\geq \left\| \hat{U}_1^\top U_0 \right\| \lambda_r - \left\| \hat{U}_1^\top U_0 \right\| \left\| \hat{\Lambda}_1 \right\| && (2.4)
 \end{aligned}$$

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

$$\left\| \hat{U}_1^\top U_0 \right\| \leq \frac{\left\| \hat{U}_1^\top \mathbf{H} U_0 \right\|}{\lambda_r - \|\mathbf{H}\|} \leq \frac{\left\| \hat{U}_1 \right\| \cdot \left\| \mathbf{H} U_0 \right\|}{\lambda_r - \|\mathbf{H}\|} = \frac{\left\| \mathbf{H} U_0 \right\|}{\lambda_r - \|\mathbf{H}\|}$$

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

$$\text{dist}(\hat{U}_0, U_0) \leq \frac{\|HU_0\|}{\Delta} \leq \frac{\|H\|}{\Delta}$$

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

$$\text{dist}(\hat{U}_0, U_0) \leq \frac{\|H\|}{\underbrace{\lambda_r(M) - \lambda_{r+1}(M)}_{\text{spectral gap}} - \|H\|} \quad (2.5)$$

Back to stochastic block model ...

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

Then the Davis-Kahan sin Θ Theorem yields

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

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

A hammer: matrix Bernstein inequality

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

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

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

Theorem 2.5 (Matrix Bernstein inequality)

For all $\tau \geq 0$,

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

A hammer: matrix Bernstein inequality

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

- **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. $1 - O((d_1 + d_2)^{-10})$

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

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

The Matrix Bernstein inequality yields

Lemma 2.6

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

$$\|\mathbf{A} - \mathbb{E}[\mathbf{A}]\| \lesssim \sqrt{np \log n} \quad (2.8)$$

Statistical accuracy of spectral clustering

Substitute (2.8) into (2.6) to reach

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

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

$$\text{dist}(\hat{\mathbf{u}}, \mathbf{u}) \ll 1 \quad \implies \quad \text{nearly perfect clustering}$$

Statistical accuracy of spectral clustering

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

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

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

$$\sum_{i,j} \mathbb{E}[\mathbf{X}_{i,j}\mathbf{X}_{i,j}^\top] \preceq \sum_{i,j} p\mathbf{e}_i\mathbf{e}_i^\top \preceq np\mathbf{I}$$

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

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

- In addition, $\|\mathbf{X}_{i,j}\| \leq 1 := B$
- Take the matrix Bernstein inequality to conclude that with high prob.,

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

Extension: singular subspaces

Singular value decomposition

Consider two matrices $M, \hat{M} = M + H$ with SVD

$$M = [U_0, U_1] \begin{bmatrix} \Sigma_0 & \mathbf{0} \\ \mathbf{0} & \Sigma_1 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} V_0^\top \\ V_1^\top \end{bmatrix}$$
$$\hat{M} = [\hat{U}_0, \hat{U}_1] \begin{bmatrix} \hat{\Sigma}_0 & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_1 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{V}_0^\top \\ \hat{V}_1^\top \end{bmatrix}$$

where U_0 (resp. \hat{U}_0) and V_0 (resp. \hat{V}_0) represent the top- r singular subspaces of M (resp. \hat{M})

Wedin $\sin \Theta$ Theorem

The Davis-Kahan Theorem generalizes to singular subspace perturbation:

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

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

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

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

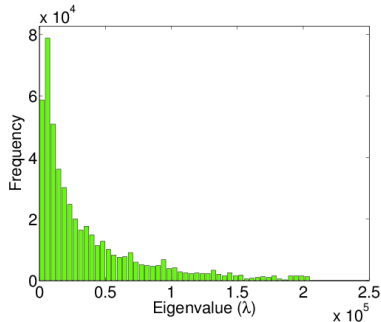
In general, we cannot infer missing ratings

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

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

Example: low-rank matrix completion

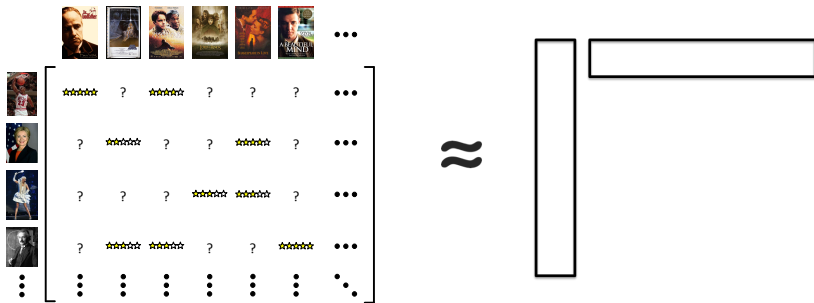
... unless rating matrix has other structure



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 \rightarrow **low-rank** approximation

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

Model for low-rank matrix completion

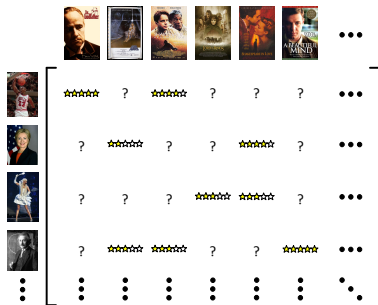


figure credit: Candès

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

Spectral estimate for matrix completion

1. set $\hat{M} \in \mathbb{R}^{n \times n}$ 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:** $\mathbb{E}[\hat{M}] = M$

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

Statistical accuracy of spectral estimate

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

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

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

$$\begin{aligned} \max \{ \text{dist}(\hat{\mathbf{u}}, \mathbf{u}), \text{dist}(\hat{\mathbf{v}}, \mathbf{v}) \} &\leq \frac{\|\hat{\mathbf{M}} - \mathbf{M}\|}{\sigma_1(\mathbf{M}) - \|\hat{\mathbf{M}} - \mathbf{M}\|} \asymp \underbrace{\|\hat{\mathbf{M}} - \mathbf{M}\|}_{\text{controlled by Bernstein}} \\ &\ll 1 \quad (\text{nearly accurate estimates}) \quad (2.9) \end{aligned}$$

Sample complexity

For rank-1 matrix completion,

$$p \gg \frac{\log^3 n}{n} \quad \implies \quad \text{nearly accurate estimates}$$

Sample complexity needed for obtaining reliable spectral estimates is

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

Proof of (2.9)

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

- First,

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

- Next, $\mathbb{E}[\mathbf{X}_{i,j} \mathbf{X}_{i,j}^\top] = \text{Var}(\hat{M}_{i,j}) \mathbf{e}_i \mathbf{e}_i^\top$ and hence

$$\begin{aligned} \mathbb{E}\left[\sum_{i,j} \mathbf{X}_{i,j} \mathbf{X}_{i,j}^\top\right] &\preceq \left\{ \max_{i,j} \text{Var}(\hat{M}_{i,j}) \right\} n \mathbf{I} \preceq \left\{ \frac{n}{p} \max_{i,j} M_{i,j}^2 \right\} \mathbf{I} \\ \implies \left\| \mathbb{E}\left[\sum_{i,j} \mathbf{X}_{i,j} \mathbf{X}_{i,j}^\top\right] \right\| &\leq \frac{n}{p} \max_{i,j} M_{i,j}^2 \lesssim \frac{\log^2 n}{np} \quad (\text{check}) \end{aligned}$$

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

$$v := \max \left\{ \left\| \mathbb{E}\left[\sum_{i,j} \mathbf{X}_{i,j} \mathbf{X}_{i,j}^\top\right] \right\|, \left\| \mathbb{E}\left[\sum_{i,j} \mathbf{X}_{i,j}^\top \mathbf{X}_{i,j}\right] \right\| \right\} \lesssim \frac{\log^2 n}{np}$$

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

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

**Extension: eigen-space for asymmetric
transition matrices**

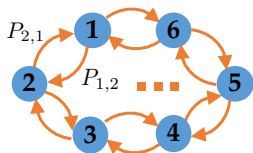
Eigen-decomposition for asymmetric matrices

Eigen-decomposition for asymmetric matrices is much more tricky; for example:

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 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 $\mathbf{P} = [P_{i,j}]_{1 \leq i,j \leq n}$
- stationary distribution $\underbrace{\boldsymbol{\pi} = [\pi_1, \dots, \pi_n]}_{\pi_1 + \dots + \pi_n = 1}$ is 1st eigenvector of \mathbf{P}

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

- $\{X_t\}_{t \geq 0}$ 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 $\|a\|_\pi := \sqrt{\pi_1 a_1^2 + \cdots + \pi_n a_n^2}$

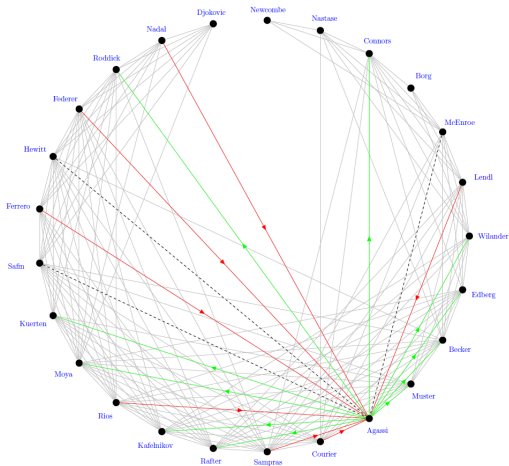
Theorem 2.8 (Chen, Fan, Ma, Wang '17)

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

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

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

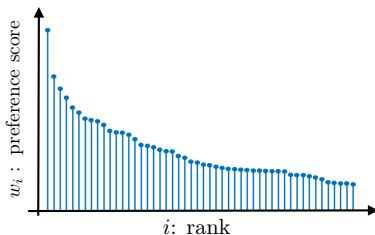
Example: ranking from pairwise comparisons



pairwise comparisons for ranking tennis players

figure credit: Bozókí, Csató, Temesi

Bradley-Terry-Luce (logistic) model



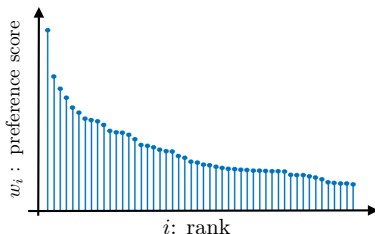
- n items to be ranked
- assign a latent score $\{w_i\}_{1 \leq i \leq n}$ to each item, so that

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

- 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

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

Spectral ranking method

- construct a **probability transition matrix** \hat{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 $\hat{\pi}$ of \hat{P}

— closely related to PageRank!

Rationale behind spectral method

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

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

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

- Thus, the stationary distribution π of P obeys

$$\pi = \frac{1}{\sum_l w_l} \mathbf{w} \quad (\text{reveal true scores})$$

Statistical guarantees for spectral ranking

— Negahban, Oh, Shah '16, Chen, Fan, Ma, Wang '17

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

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

- consequence of Theorem 2.8 and the matrix Bernstein (exercise)

Reference

- [1] "*The rotation of eigenvectors by a perturbation*," C. Davis, W. Kahan, *SIAM Journal on Numerical Analysis*, 1970.
- [2] "*Perturbation bounds in connection with singular value decomposition*," P. Wedin, *BIT Numerical Mathematics*, 1972.
- [3] "*Inference, estimation, and information processing, EE 378B lecture notes*," A. Montanari, Stanford University.
- [4] "*COMS 4772 lecture notes*," D. Hsu, Columbia University.
- [5] "*High-dimensional statistics: a non-asymptotic viewpoint*," M. Wainwright, *Cambridge University Press*, 2019.
- [6] "*Principal component analysis for big data*," J. Fan, Q. Sun, W. Zhou, Z. Zhu, *arXiv:1801.01602*, 2018.

Reference

- [7] "*Community detection and stochastic block models*," E. Abbe, *Foundations and Trends in Communications and Information Theory*, 2018.
- [8] "*Consistency thresholds for the planted bisection model*," E. Mossel, J. Neeman, A. Sly, *ACM Symposium on Theory of Computing*, 2015.
- [9] "*Matrix completion from a few entries*," R. Keshavan, A. Montanari, S. Oh, *IEEE Transactions on Information Theory*, 2010.
- [10] "*The PageRank citation ranking: bringing order to the web*," L. Page, S. Brin, R. Motwani, T. Winograd, 1999.
- [11] "*Rank centrality: ranking from pairwise comparisons*," S. Negahban, S. Oh, D. Shah, *Operations Research*, 2016.
- [12] "*Spectral method and regularized MLE are both optimal for top- K ranking*," Y. Chen, J. Fan, C. Ma, K. Wang, *accepted to Annals of Statistics*, 2017.