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
Community structures are common in many social networks.

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

*figure credit: The Future Buzz*  
*figure credit: S. Papadopoulos*
A simple model: stochastic block model (SBM)

- $n$ 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)

- observe a graph $G$

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

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

- **Goal:** recover community memberships of all nodes, i.e. $\{x_i\}$
Consider the adjacency matrix $A \in \{0, 1\}^{n \times n}$ of $G$:

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

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

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

\[
\mathbb{E}[A] = \begin{bmatrix}
p11^\top & q11^\top \\
qu11^\top & p11^\top
\end{bmatrix} = \underbrace{\frac{p+q}{2}11^\top}_{\text{uninformative bias}} + \underbrace{\frac{p-q}{2}}\begin{bmatrix}1 \\
-1
\end{bmatrix}\begin{bmatrix}1^\top, -1^\top\end{bmatrix}
\]

\[= \mathbf{x} := [x_i]_{1 \leq i \leq n}\]
Spectral clustering

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

1. computing the leading eigenvector \( \hat{u} = [\hat{u}_i]_{1 \leq i \leq n} \) of \( A - \frac{p+q}{2} \mathbf{1} \mathbf{1}^\top \)

2. rounding: output \( \hat{x}_i = \begin{cases} 1, & \text{if } \hat{u}_i > 0 \\ -1, & \text{if } \hat{u}_i < 0 \end{cases} \)
Rationale: recovery is reliable if $A - \mathbb{E}[A]$ is sufficiently small perturbation

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

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

Question: how to quantify the effect of perturbation $A - \mathbb{E}[A]$ on $\hat{u}$?
Distance and angles between two subspaces
Consider 2 symmetric matrices $M, \hat{M} = M + H \in \mathbb{R}^{n \times n}$ with eigen-decompositions

$$M = \sum_{i=1}^{n} \lambda_i u_i u_i^\top$$

and

$$\hat{M} = \sum_{i=1}^{n} \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top$$

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

$$M = [U_0, U_1] \begin{bmatrix} \Lambda_0 & \Lambda_0 \\ \Lambda_1 & \Lambda_1 \end{bmatrix} \begin{bmatrix} U_0^\top \\ U_1^\top \end{bmatrix}$$

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

Here, $U_0 = [u_1, \cdots, u_r], \Lambda_0 = \text{diag} ([\lambda_1, \cdots, \lambda_r]), \cdots$
Setup and notation

\[ M = \begin{bmatrix} u_1 & \cdots & u_r & u_{r+1} & \cdots & u_n \end{bmatrix} \]

\[ \begin{array}{c}
\begin{bmatrix}
\lambda_1 \\
\vdots \\
\lambda_r \\
\Lambda_0
\end{bmatrix} \\
\begin{bmatrix}
\lambda_{r+1} \\
\vdots \\
\lambda_n \\
\Lambda_1
\end{bmatrix}
\end{array} \]

\[ \begin{bmatrix}
u_1^\top \\
u_r^\top \\
u_{r+1}^\top \\
u_n^\top
\end{bmatrix} \begin{bmatrix}U_0^\top \\U_1^\top \end{bmatrix} \]
Setup and notation

• $\|M\|$: spectral norm (largest singular value of $M$)

• $\|M\|_F$: Frobenius norm ($\|M\|_F = \sqrt{\text{tr}(M^\top M)} = \sqrt{\sum_{i,j} M_{i,j}^2}$)
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

∀ 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

$$\text{dist}(X_0, Z_0) := \| X_0 X_0^\top - Z_0 Z_0^\top \|$$  \hspace{1cm} (2.1)

This metric has several equivalent expressions:

**Lemma 2.1**

Suppose $X := [X_0, X_1]$ and $Z := [Z_0, Z_1]$ are orthonormal complement subspaces. Then

$$\text{dist}(X_0, Z_0) = \| X_0^\top Z_1 \| = \| Z_0^\top X_1 \|$$

- sanity check: if $X_0 = Z_0$, then $\text{dist}(X_0, Z_0) = \| X_0^\top Z_1 \| = 0$
Principal angles between two eigen-spaces

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

\[
\theta = \arccos \langle \mathbf{x}_0, \mathbf{z}_0 \rangle
\]

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

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

where $\{\theta_1, \cdots, \theta_r\}$ are called the principal angles between $X_0$ and $Z_0$.
As expected, principal angles and distance are closely related.

**Lemma 2.2**

Suppose $X := [X_0, X_1]$ and $Z := [Z_0, Z_1]$ are orthonormal matrices. Then

$$\|X_0^\top Z_1\| = \|\sin \Theta\| = \max\{|\sin \theta_1|, \cdots, |\sin \theta_r|\}$$

Lemmas 2.1 and 2.2 taken collectively give

$$\text{dist}(X_0, Z_0) = \max\{|\sin \theta_1|, \cdots, |\sin \theta_r|\} \quad (2.2)$$
Proof of Lemma 2.2

\[ \|X_0^\top Z_1\| = \|X_0^\top Z_1 Z_1^\top X_0\|^\frac{1}{2} \]
\[ = I - Z_0 Z_0^\top \]
\[ = \|X_0^\top X_0 - X_0^\top Z_0 Z_0^\top X_0\|^\frac{1}{2} \]
\[ = \|I - U \cos^2 \Theta U^\top\|^\frac{1}{2} \quad \text{(since } X_0^\top Z_0 = U \cos \Theta V^\top\text{)} \]
\[ = \|I - \cos^2 \Theta\|^\frac{1}{2} \]
\[ = \|\sin \Theta\| \]
Proof of Lemma 2.1

We first claim that SVD of $X_1^T Z_0$ can be written as

$$X_1^T Z_0 = \tilde{U} \sin \Theta V^T$$

(2.3)

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

$$Z_0 = [X_0, X_1] \begin{bmatrix} X_0^T \\ X_1^T \end{bmatrix} Z_0 = [X_0, X_1] \begin{bmatrix} U \cos \Theta & V^T \\ \tilde{U} \sin \Theta & V^T \end{bmatrix}$$

$$\implies Z_0 Z_0^T = [X_0, X_1] \begin{bmatrix} U \cos^2 \Theta & U \cos \Theta \sin \Theta & \tilde{U} \cos \Theta \sin \Theta & \tilde{U} \sin^2 \Theta \\ \tilde{U} \cos \Theta \sin \Theta & \tilde{U} \sin^2 \Theta & \tilde{U} \sin \Theta \tilde{U}^T & \tilde{U} \sin \Theta \tilde{U}^T \end{bmatrix} \begin{bmatrix} X_0^T \\ X_1^T \end{bmatrix}$$

As a consequence,

$$X_0 X_0^T - Z_0 Z_0^T$$

$$= [X_0, X_1] \begin{bmatrix} I - U \cos^2 \Theta & U \cos \Theta \sin \Theta & -U \cos \Theta \sin \Theta & -\tilde{U} \sin^2 \Theta \tilde{U}^T \\ -\tilde{U} \cos \Theta \sin \Theta & -\tilde{U} \sin^2 \Theta & -\tilde{U} \sin \Theta \tilde{U}^T & -\tilde{U} \sin \Theta \tilde{U}^T \end{bmatrix} \begin{bmatrix} X_0^T \\ X_1^T \end{bmatrix}$$
Proof of Lemma 2.1 (cont.)

This further gives

$$\|X_0 X_0^T - Z_0 Z_0^T\|$$

$$= \left\| \begin{bmatrix} U & \tilde{U} \end{bmatrix} \begin{bmatrix} \sin^2 \Theta & -\cos \Theta \sin \Theta \\ -\cos \Theta \sin \Theta & -\sin^2 \Theta \end{bmatrix} \begin{bmatrix} U^T \\ \tilde{U}^T \end{bmatrix} \right\|$$

$$= \left\| \begin{bmatrix} \sin^2 \Theta & -\cos \Theta \sin \Theta \\ -\cos \Theta \sin \Theta & -\sin^2 \Theta \end{bmatrix} \right\|$$

$(\| \cdot \| \text{ is rotationally invariant})$

each block is diagonal matrix

$$= \max_{1 \leq i \leq r} \left\| \begin{bmatrix} \sin^2 \theta_i & -\cos \theta_i \sin \theta_i \\ -\cos \theta_i \sin \theta_i & -\sin^2 \theta_i \end{bmatrix} \right\|$$

$$= \max_{1 \leq i \leq r} \left\| \begin{bmatrix} \sin \theta_i & -\cos \theta_i \\ -\cos \theta_i & -\sin \theta_i \end{bmatrix} \right\|$$

$$= \max_{1 \leq i \leq r} |\sin \theta_i| = \| \sin \Theta \|$$
Proof of Lemma 2.1 (cont.)

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

\[ Z_0^\top X_1 X_1^\top Z_0 = Z_0^\top Z_0 - Z_0^\top X_0 X_0^\top Z_0 \]
\[ = I - V \cos^2 \Theta V^\top \]
\[ = V \sin^2 \Theta V^\top \]

and hence right singular space (resp. singular values) of \( X_1^\top Z_0 \) is given by \( V \) (resp. \( \sin \Theta \))
Eigen-space perturbation theory
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 smallest non-zero eigenvalue of $M$ and perturbation eigen-gap between $\lambda_r$ and $\lambda_{r+1}$
Proof of Theorem 2.3

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

$$\left\| \hat{U}_1^T H U_0 \right\| = \left\| \hat{U}_1^T \left( \underbrace{\hat{U} \hat{\Lambda} \hat{U}^T - U \Lambda U^T}_{M+H} \right) U_0 \right\|$$

$$= \left\| \hat{\Lambda}_1 \hat{U}_1^T U_0 - \hat{U}_1^T U_0 \Lambda_0 \right\| \quad \text{(since } U_1^T U_0 = \hat{U}_1^T \hat{U}_0 = 0)$$

$$\geq \left\| \hat{U}_1^T U_0 \Lambda_0 \right\| - \left\| \hat{\Lambda}_1 \hat{U}_1^T U_0 \right\| \quad \text{(triangle inequality)}$$

$$\geq \left\| \hat{U}_1^T U_0 \right\| \lambda_r - \left\| \hat{U}_1^T U_0 \right\| \left\| \hat{\Lambda}_1 \right\| \quad \text{(2.4)}$$

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

$$\left\| \hat{U}_1^T U_0 \right\| \leq \frac{\left\| \hat{U}_1^T H U_0 \right\|}{\lambda_r - \| H \|} \leq \frac{\| U_1 \| \cdot \| H U_0 \|}{\lambda_r - \| H \|} = \frac{\| H U_0 \|}{\lambda_r - \| H \|}$$

This together with Lemma 2.1 completes the proof
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

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

spectral gap
Let \( M = \mathbb{E}[A] - \frac{p + q}{2} \mathbf{1}\mathbf{1}^\top \), \( \hat{M} = A - \frac{p+q}{2} \mathbf{1}\mathbf{1}^\top \) and \( u = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \)

Then the Davis-Kahan sin \( \Theta \) Theorem yields

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

**Question:** how to bound \( \|A - \mathbb{E}[A]\| \)?
A hammer: matrix Bernstein inequality

Consider a sequence of independent random matrices \( \{X_l \in \mathbb{R}^{d_1 \times d_2} \} \)

- \( \mathbb{E}[X_l] = 0 \)
- \( \|X_l\| \leq B \) for each \( l \)
- variance statistic:

\[
v := \max \left\{ \| \mathbb{E} \left[ \sum_l X_l X_l^\top \right] \|, \| \mathbb{E} \left[ \sum_l X_l^\top X_l \right] \| \right\}
\]

**Theorem 2.5 (Matrix Bernstein inequality)**

For all \( \tau \geq 0 \),

\[
\mathbb{P} \left\{ \left\| \sum_l X_l \right\| \geq \tau \right\} \leq (d_1 + d_2) \exp \left( \frac{-\tau^2/2}{v + B\tau/3} \right)
\]
A hammer: matrix Bernstein inequality

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

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

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

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

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

Spectral methods
The Matrix Bernstein inequality yields

**Lemma 2.6**

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

$$\|A - \mathbb{E}[A]\| \lesssim \sqrt{np \log n}$$ (2.8)
Statistical accuracy of spectral clustering

Substitute (2.8) into (2.6) to reach

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

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

\[
\text{dist}(\hat{u}, u) \ll 1 \quad \implies \quad \text{nearly perfect clustering}
\]
Statistical accuracy of spectral clustering

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

• **dense regime:** if \( p \approx q \approx 1 \), 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 \( a, b \approx 1 \), 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 \( A - \mathbb{E}[A] \) as \( \sum_{i,j} X_{i,j} \), where \( X_{i,j} = (A_{i,j} - \mathbb{E}[A_{i,j}])e_i e_j^\top \)

- Since \( \text{Var}(A_{i,j}) \leq p \), one has \( \mathbb{E} [X_{i,j} X_{i,j}^\top] \preceq p e_i e_i^\top \), which gives

\[
\sum_{i,j} \mathbb{E} [X_{i,j} X_{i,j}^\top] \preceq \sum_{i,j} p e_i e_i^\top \preceq np I
\]

Similarly, \( \sum_{i,j} \mathbb{E} [X_{i,j}^\top X_{i,j}] \preceq np I \). As a result,

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

- In addition, \( \|X_{i,j}\| \leq 1 := B \)

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

\[
\|A - \mathbb{E}[A]\| \preceq \sqrt{v \log n} + B \log n \preceq \sqrt{np \log n} \quad (\text{since } p \gtrsim \frac{\log n}{n})
\]
Extension: singular subspaces
Consider two matrices $M, \hat{M} = M + H$ with SVD

$$M = [U_0, U_1] \begin{bmatrix} \Sigma_0 & 0 \\ 0 & \Sigma_1 \end{bmatrix} \begin{bmatrix} V_0^\top \\ V_1^\top \end{bmatrix}$$

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

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

$$\max \left\{ \text{dist} (\hat{U}_0, U_0), \text{dist} (\hat{V}_0, V_0) \right\} \leq \frac{\max \left\{ \|HV_0\|, \|H^TU_0\| \right\}}{\Delta} \quad \text{two-sided interactions}$$

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

\[
\begin{bmatrix}
\checkmark & ? & ? & ? & \checkmark & ? \\
? & ? & \checkmark & \checkmark & ? & ? \\
\checkmark & ? & ? & \checkmark & ? & ? \\
? & ? & \checkmark & ? & ? & \checkmark \\
? & \checkmark & ? & ? & \checkmark & ? \\
? & ? & \checkmark & \checkmark & ? & ? \\
\end{bmatrix}
\]

— 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

• consider a low-rank matrix $M$

• each entry $M_{i,j}$ is observed independently with prob. $p$

• goal: fill in missing entries

*figure credit: Candès*
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 $(\hat{U}, \hat{\Sigma}, \hat{V})$
Let’s analyze a simple case where $M = uv^\top$ with

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

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

$$\max \{\dist(\hat{u}, u), \dist(\hat{v}, v)\} \leq \frac{\|\hat{M} - M\|}{\sigma_1(M) - \|\hat{M} - M\|} \approx \frac{\|\hat{M} - M\|}{\|\hat{M} - M\|}$$

controlled by Bernstein

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

\[ n^2 p \asymp n \log^3 n \]

optimal up to log factor
Proof of (2.9)

Write $\hat{M} - M = \sum_{i,j} X_{i,j}$, where $X_{i,j} = (\hat{M}_{i,j} - M_{i,j})e_i e_j^\top$

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

- Next, $\mathbb{E}[X_{i,j} X_{i,j}^\top] = \text{Var}(\hat{M}_{i,j})e_i e_i^\top$ and hence
  $$\mathbb{E}\left[ \sum_{i,j} X_{i,j} X_{i,j}^\top \right] \preceq \left\{ \max_{i,j} \text{Var}(\hat{M}_{i,j}) \right\} nI \leq \left\{ \frac{n}{p} \max_{i,j} M_{i,j}^2 \right\} I$$

  $$\implies \|\mathbb{E}\left[ \sum_{i,j} X_{i,j} X_{i,j}^\top \right]\| \leq \frac{n}{p} \max_{i,j} M_{i,j}^2 \lesssim \frac{\log^2 n}{np} \quad \text{(check)}$$

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

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

- Take the matrix Bernstein inequality to yield: if $p \gg \log^3 n/n$, then
  $$\|\hat{M} - M\| \lesssim \sqrt{v \log n + B \log n} \ll 1$$
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
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 \( P = [P_{i,j}]_{1 \leq i,j \leq n} \)
- stationary distribution \( \pi = [\pi_1, \cdots, \pi_n] \) is 1st eigenvector of \( P \)
  \[ \pi_1 + \cdots + \pi_n = 1 \]
  \[ \pi P = \pi \]
- \( \{X_t\}_{t \geq 0} \) 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

\[
\| \hat{\pi} - \pi \|_\pi \leq \frac{\| \pi(\hat{P} - P) \|_\pi}{1 - \max \{ \lambda_2(P), -\lambda_n(P) \}} - \| \hat{P} - P \|_\pi
\]

- \( \hat{P} \) does not need to induce 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
  \[ \text{item } i \succ \text{item } j \text{ if } w_i > w_j \]
- 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 $\hat{P}$ 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

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

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

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

- Thus, the stationary distribution \( \pi \) of \( P \) obeys

\[ \pi = \frac{1}{\sum_l w_l} w \quad \text{(reveal true scores)} \]
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.

$$\frac{\| \hat{\pi} - \pi \|_2}{\| \pi \|_2} \lesssim \frac{\| \hat{\pi} - \pi \|_\pi}{\| \pi \|_2} \lesssim \frac{1}{\sqrt{n}} \to 0$$

nearly perfect estimate

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


Reference


