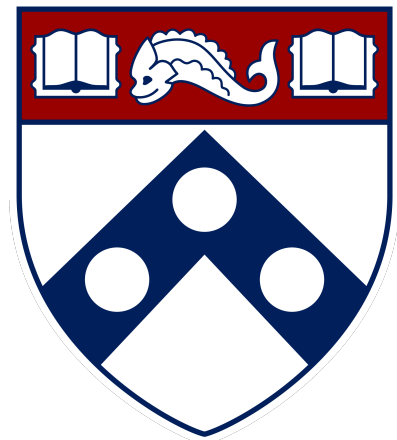
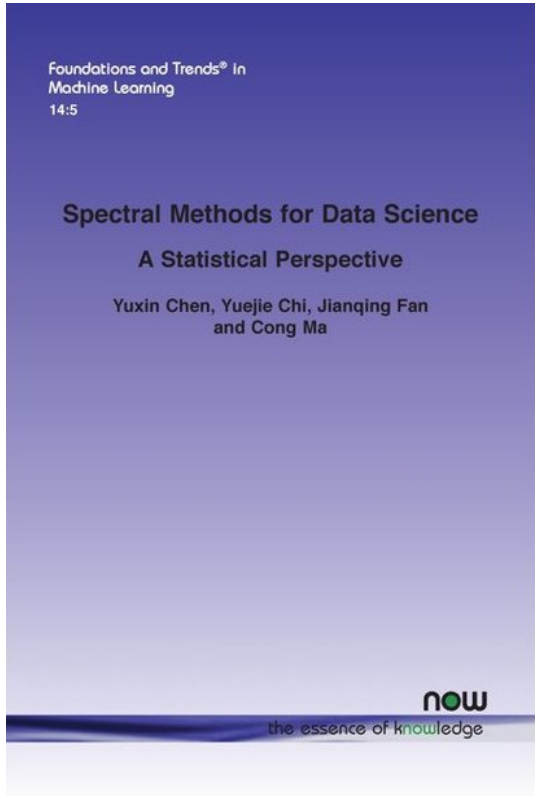
Spectral methods for data science: A statistical perspective



Yuxin Chen, Statistics & Data Science, UPenn

Summer school on theoretical stats., PKU BICMR & Math, 2023



"Spectral methods for data science: a statistical perspective," Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, Foundations and Trends in Machine Learning, 2021

Part 1: Introduction

- Motivating applications
  - community detection
  - matrix/tensor completion
  - ranking
- A general recipe for spectral methods

Motivating application: community detection

Graph clustering / community detection

Community structures are common in many social networks

figure credit: The Future Buzz

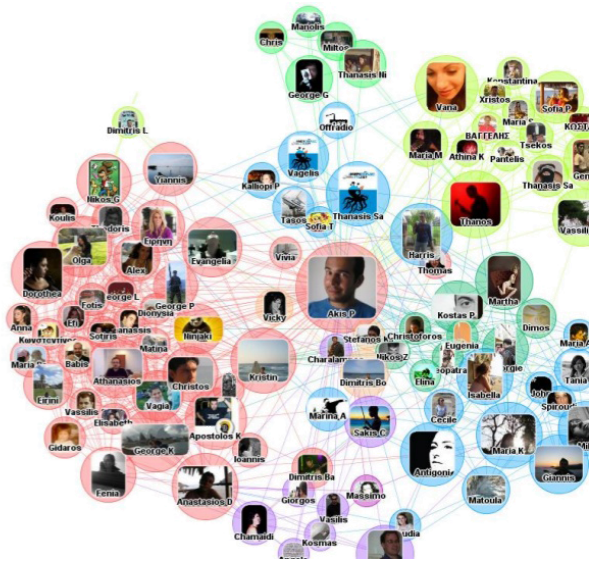
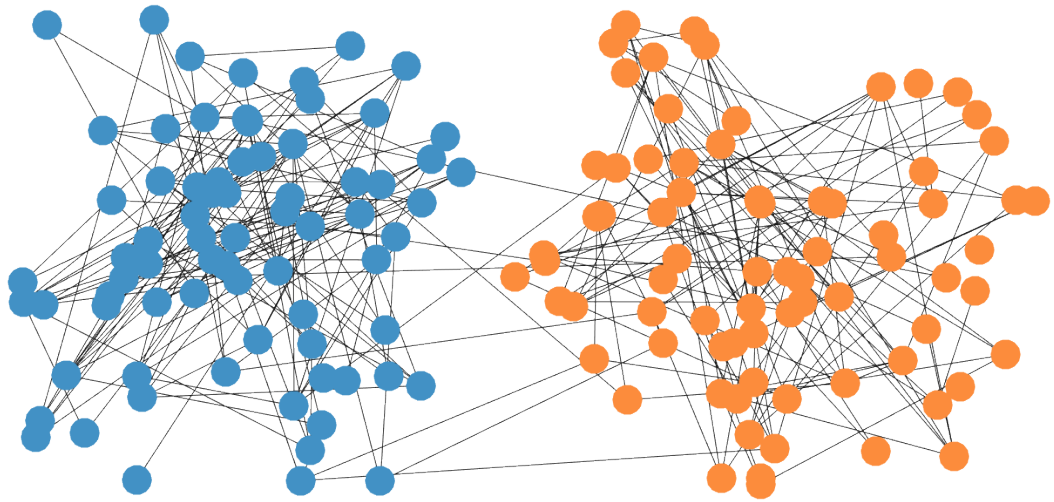


figure credit: S. Papadopoulos

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

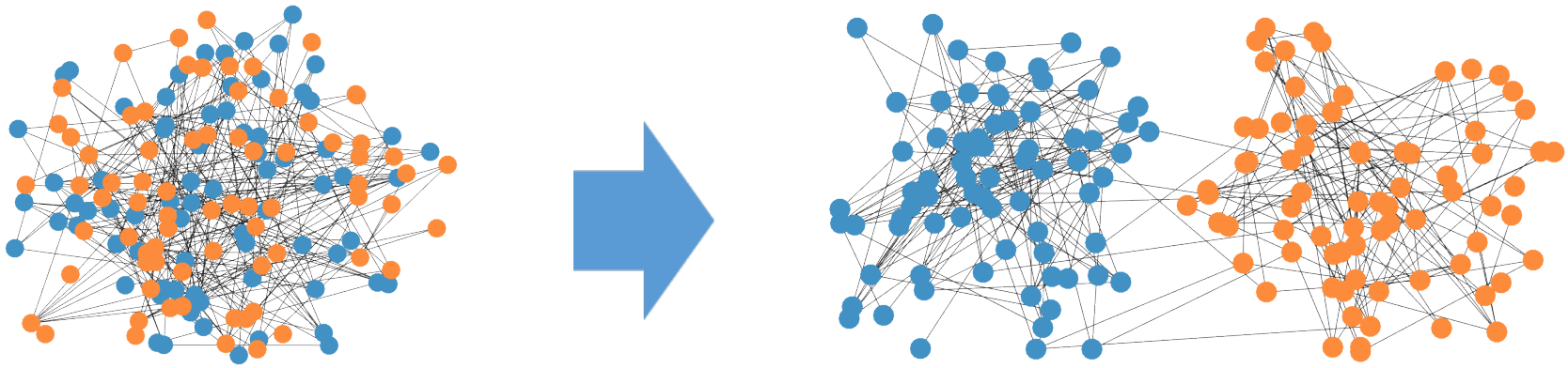
An idealistic model: stochastic block model (SBM)



$x_i = 1$: 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

An idealistic model: stochastic block model (SBM)



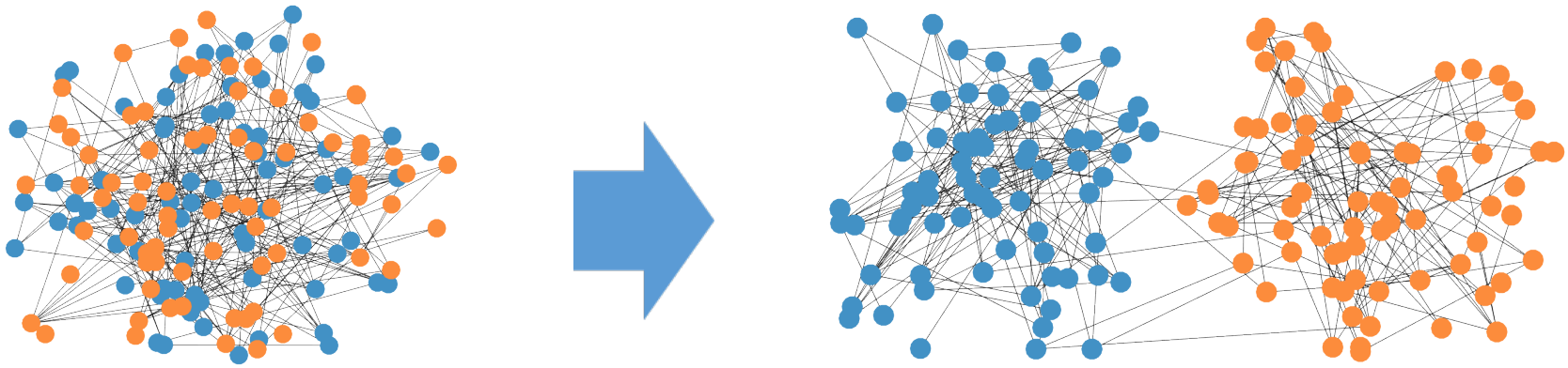
$\mathcal{G}$

- **observation:** a (random) graph $\mathcal{G}$

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

  - $p > q$ (i.e. more within-cluster edges than between-cluster edges)

An idealistic model: stochastic block model (SBM)



$\mathcal{G}$

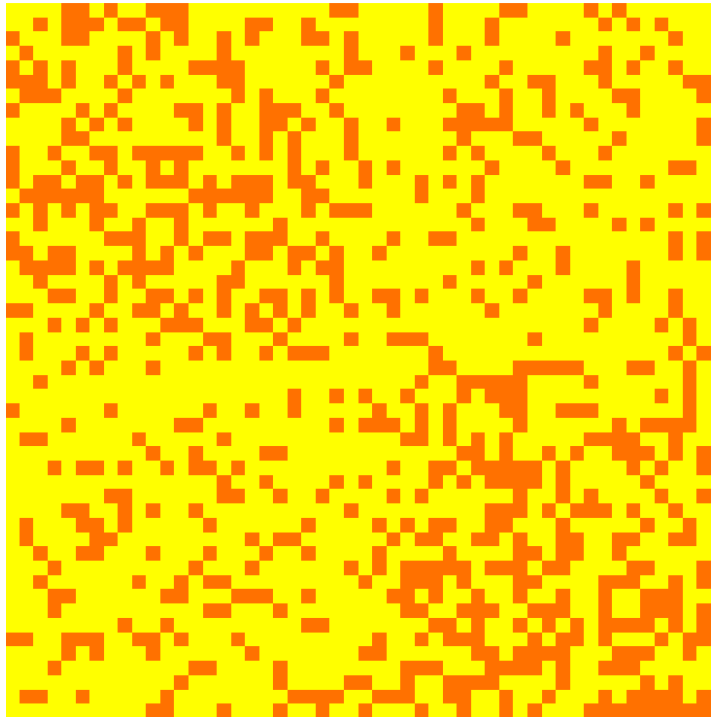
- **observation:** a (random) graph $\mathcal{G}$

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

  - $p > q$ (i.e. more within-cluster edges than between-cluster edges)

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

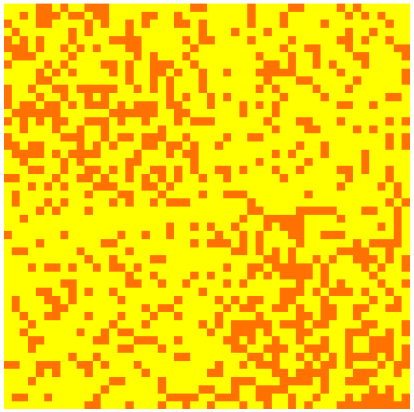
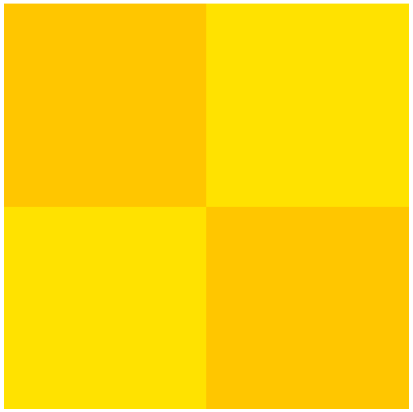
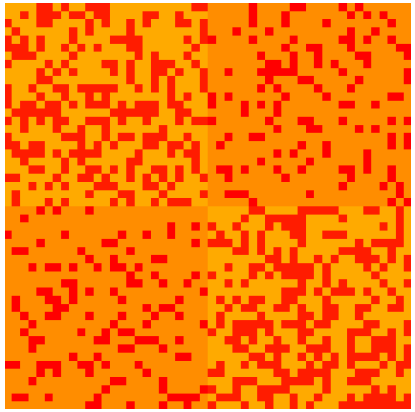
Key structure of adjacency matrix



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

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

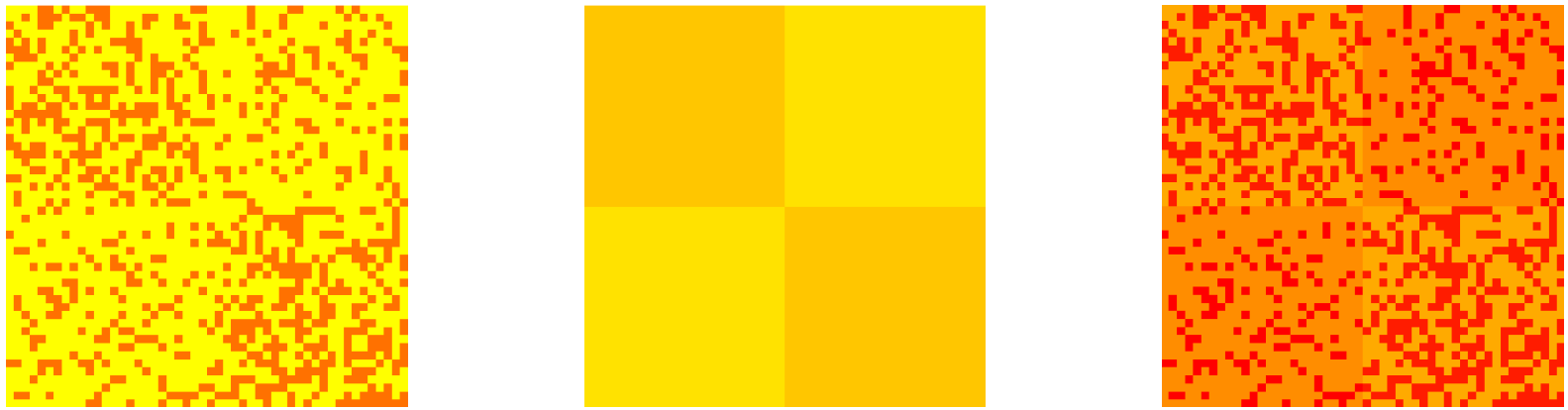
Key structure of adjacency matrix

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

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

Key structure of adjacency matrix

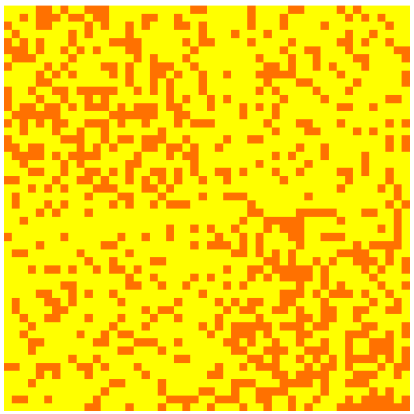
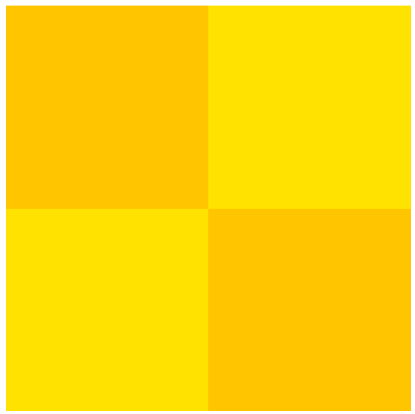
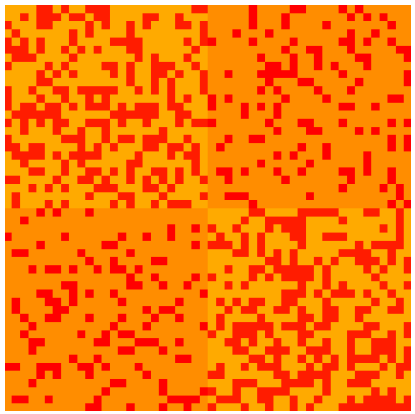


$$\boldsymbol{A} = \underbrace{\mathbb{E}[\boldsymbol{A}]}_{\text{rank 2}} + \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$$

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

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

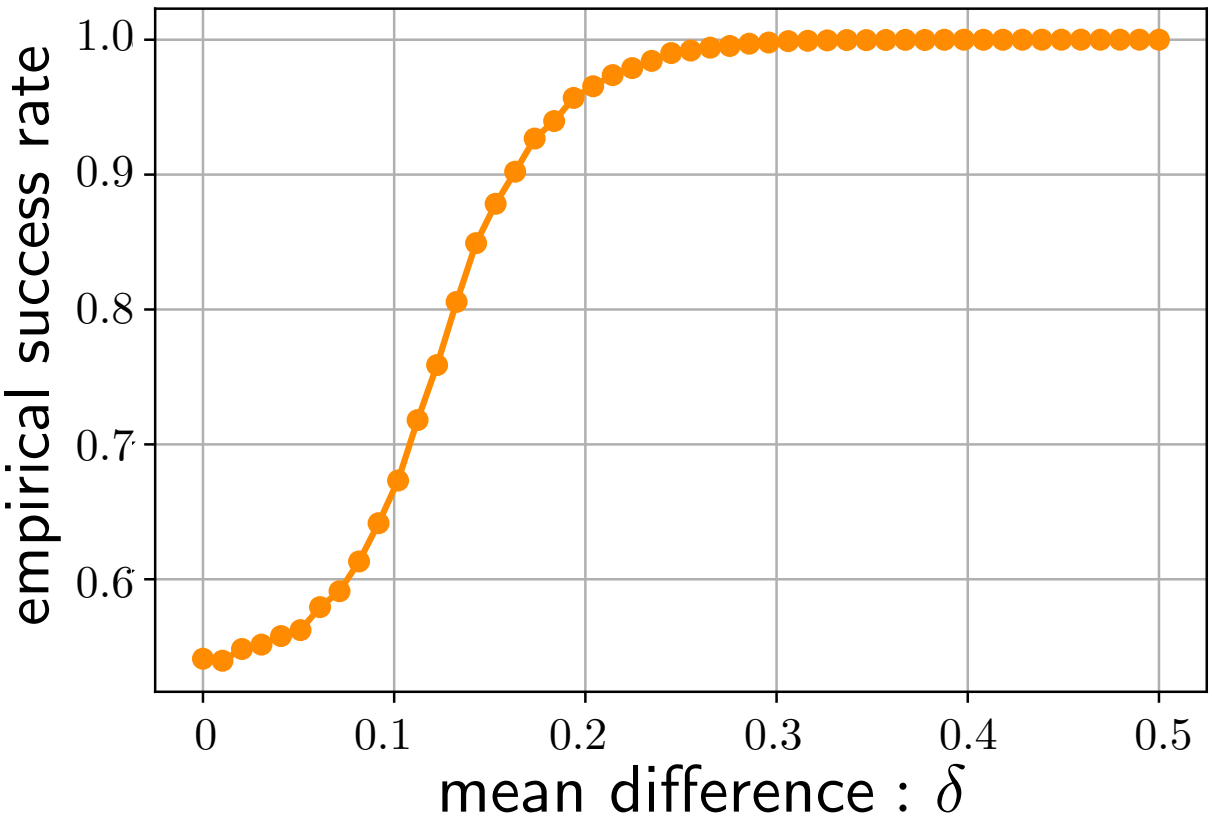
Spectral clustering

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

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

Empirical clustering accuracy



$n = 100$, $p = \frac{1+\delta}{2}$, $q = \frac{1-\delta}{2}$

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

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

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

What we'll demonstrate: effect of perturbation $\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$ on $\boldsymbol{u}$

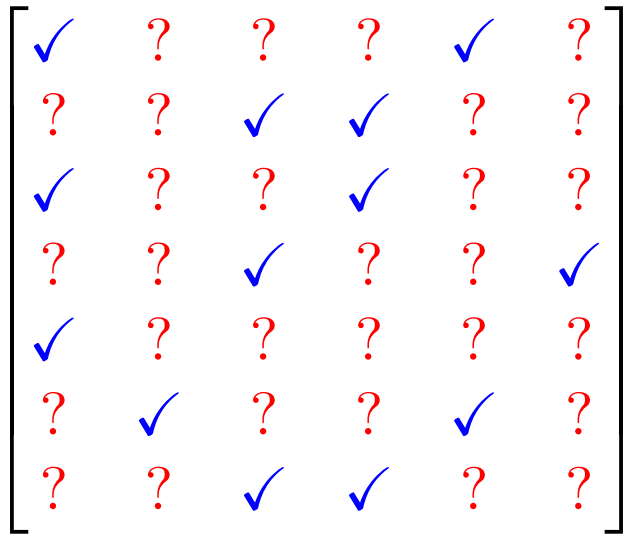
Motivating application: matrix/tensor completion

figure credit: Candes et al.

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

Matrix completion

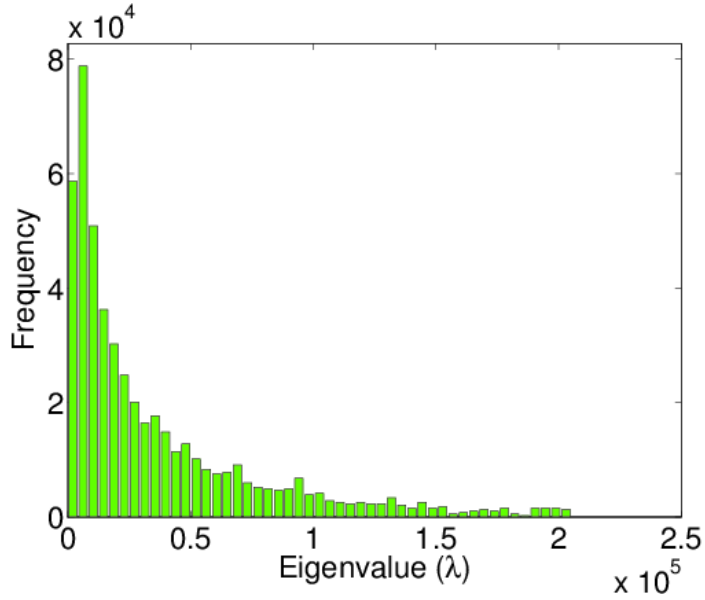
In general, we cannot infer missing ratings



— an underdetermined system (more unknowns than observations)

Matrix completion

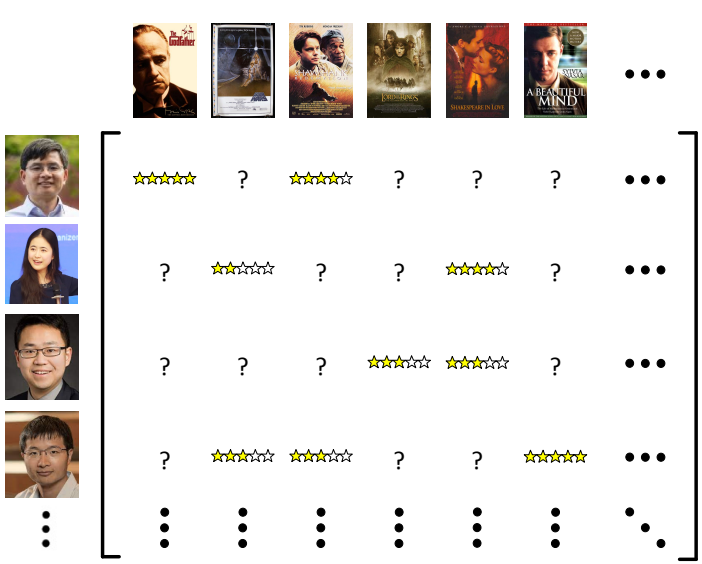
. . . unless rating matrix has other structure

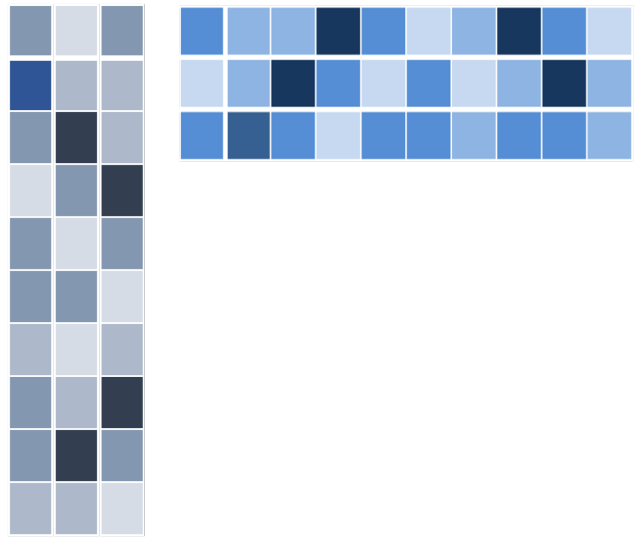


A few factors explain most of the data

Matrix completion

. . . unless rating matrix has other structure

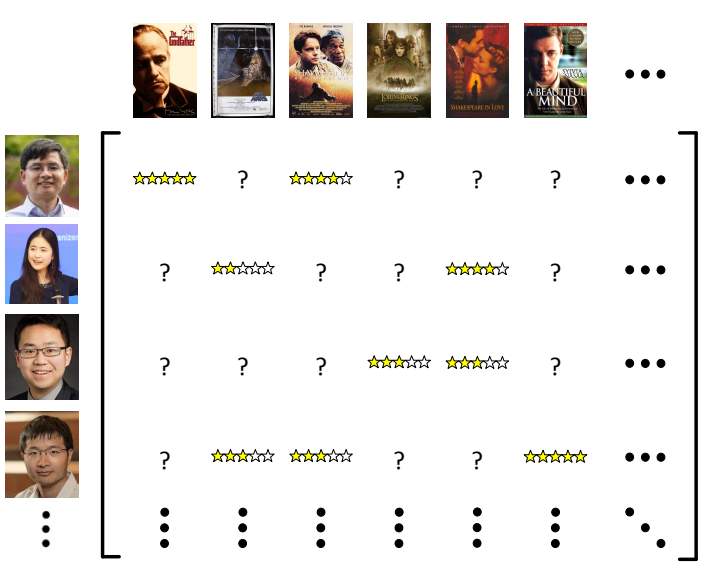


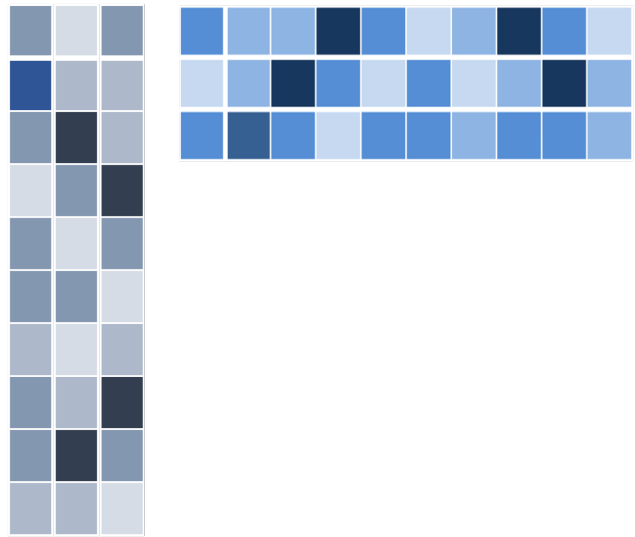


A few factors explain most of the data $\longrightarrow$ low-rank approximation

Matrix completion

. . . unless rating matrix has other structure





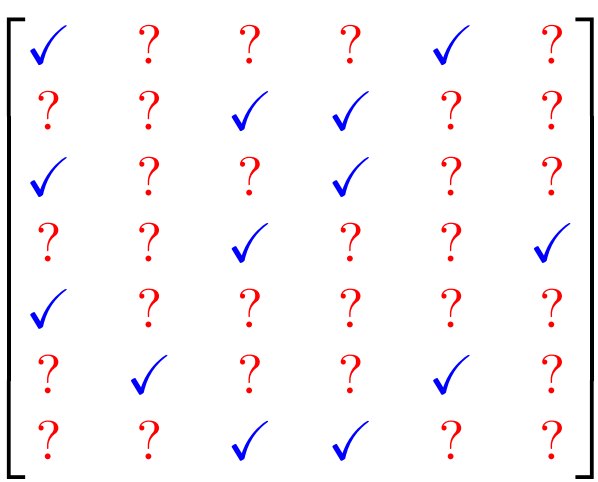
A few factors explain most of the data $\longrightarrow$ low-rank approximation

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

Model: low-rank matrix completion

- ground truth: rank-r matrix $\boldsymbol{M}^\star = \underbrace{\boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}}_{\text{rank-}r\text{ SVD}} \in \mathbb{R}^{n_1 \times n_2}$
- each entry $M^\star_{i,j}$ is observed independently with prob. p
- **goal:** fill in unseen entries of $\boldsymbol{M}^\star$

Model: low-rank matrix completion



- ground truth: rank-r matrix $\boldsymbol{M}^\star = \underbrace{\boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}}_{\text{rank-}r\text{ SVD}} \in \mathbb{R}^{n_1 \times n_2}$
- each entry $M^\star_{i,j}$ is observed independently with prob. p
- **goal:** fill in unseen entries of $\boldsymbol{M}^\star$
- **intermediate step:** estimate $\boldsymbol{U}^\star, \boldsymbol{V}^\star, \boldsymbol{\Sigma}^\star$

Spectral method for matrix completion

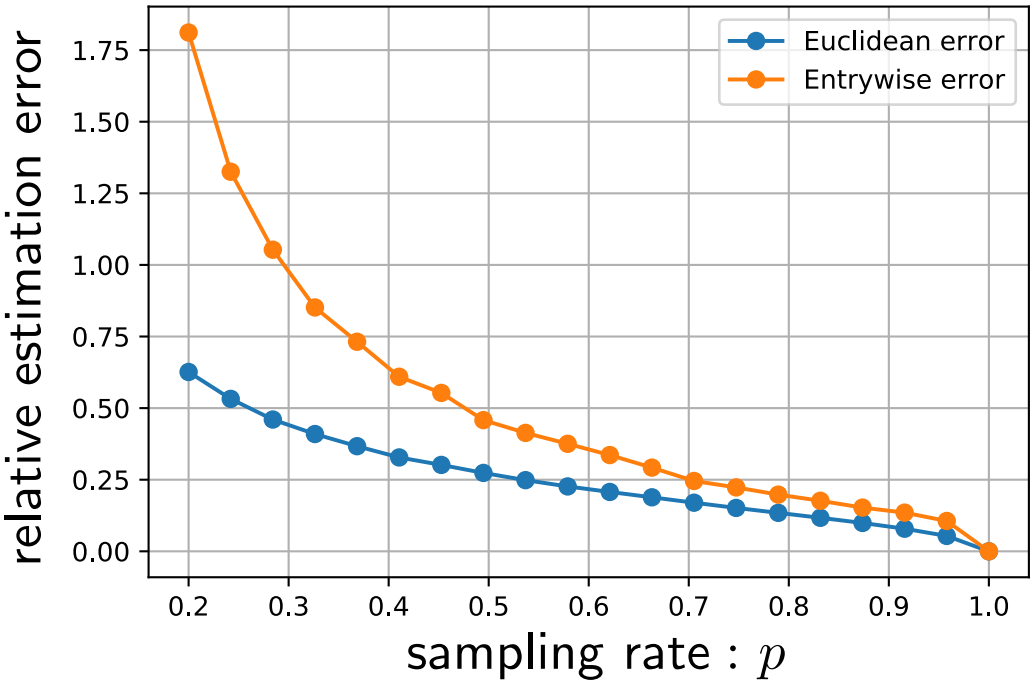
1. construct a rescaled zero-filled matrix $\boldsymbol{M} = [M_{i,j}] \in \mathbb{R}^{n_1 \times n_2}$ as

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

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

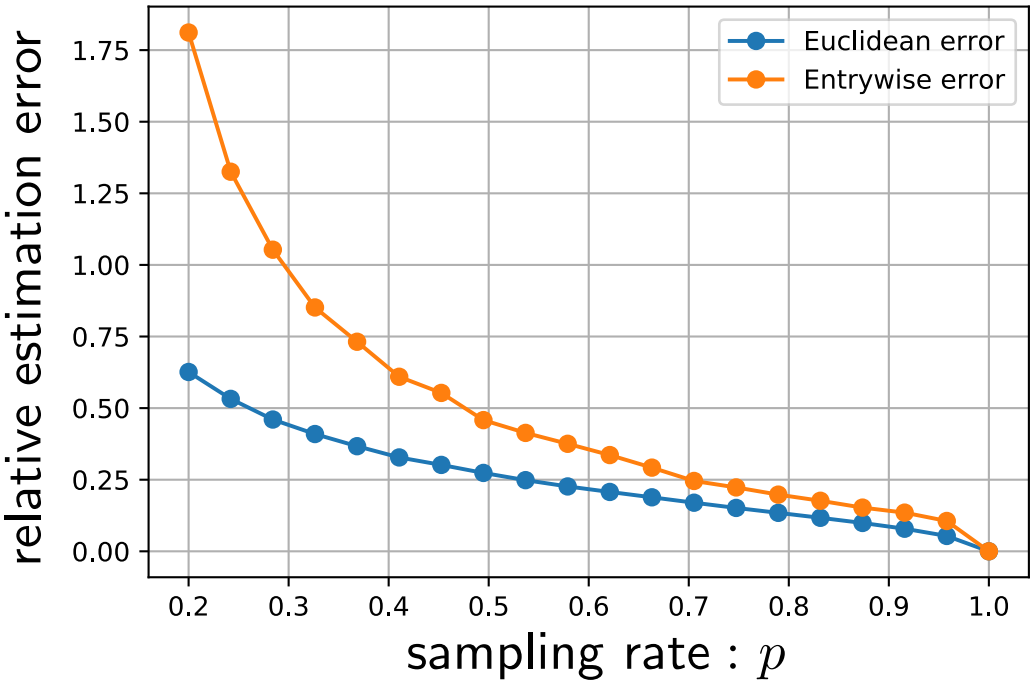
2. compute rank-r SVD $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ of $\boldsymbol{M}$, and return $\widehat{\boldsymbol{M}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

Empirical matrix estimation accuracy



$n = 200,\ r = 5$

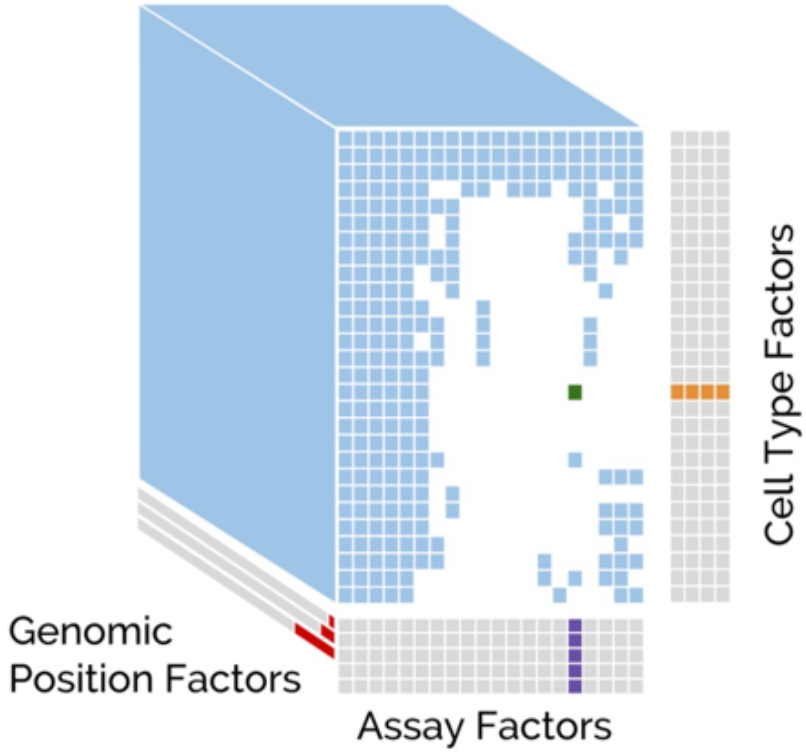
Empirical matrix estimation accuracy



$n = 200,\ r = 5$

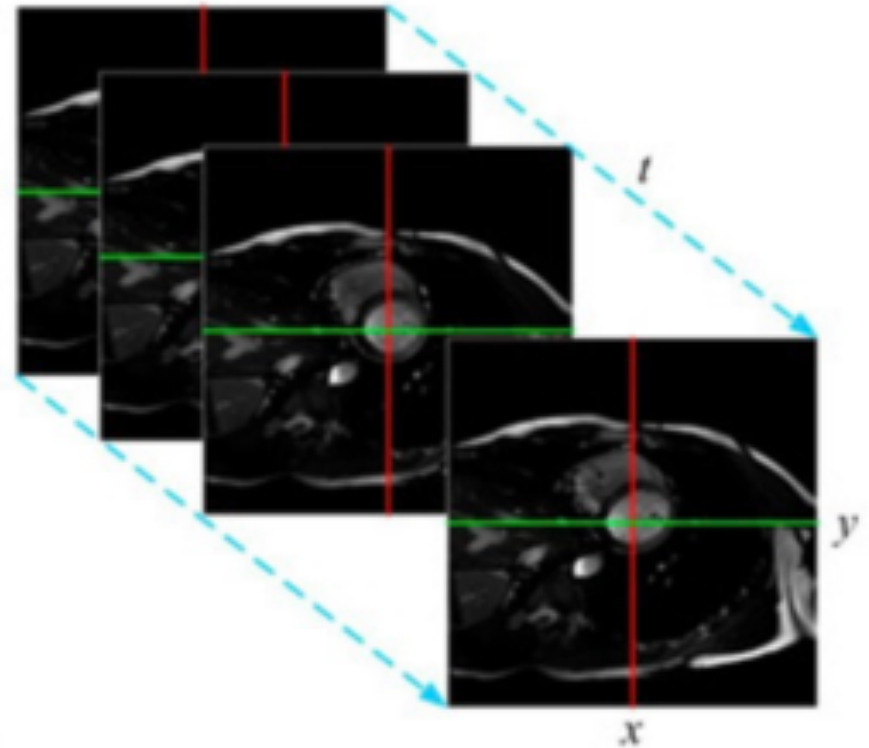
What we'll see: effect of sampling rate p upon estimation accuracy

Extension: tensor data



computational genomics

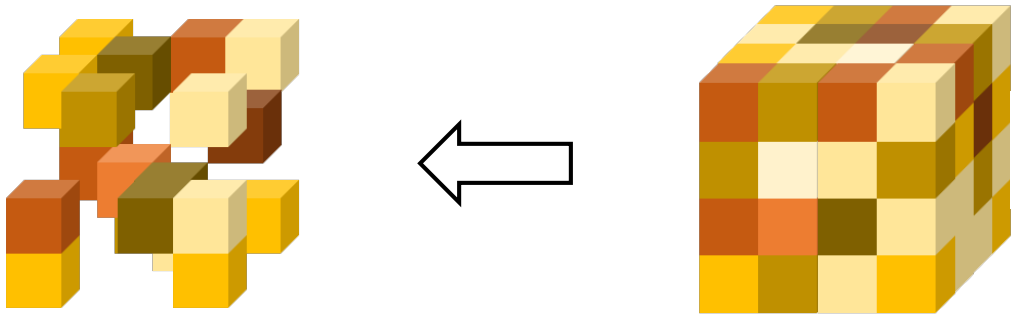
— *fig. credit: Schreiber et al. 19*



dynamic MRI

— *fig. credit: Liu et al. 17*

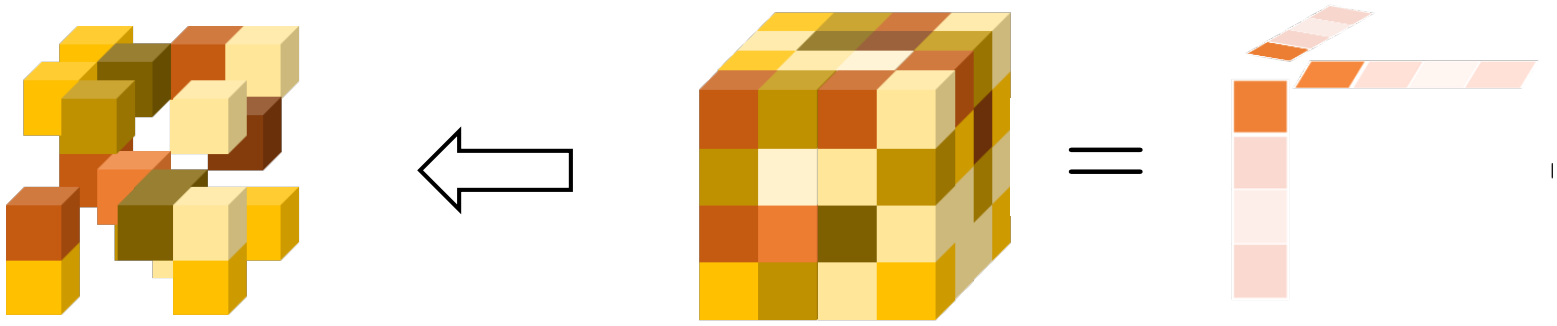
Extension: low-rank tensor completion



observed true tensor $\boldsymbol{T}^\star$

- ground truth: rank-1 tensor $\boldsymbol{T}^\star \in \mathbb{R}^{n \times n \times n}$
- each entry $T^\star_{i,j,k}$ is observed independently with prob. p
- **goal:** fill in unseen entries of $\boldsymbol{T}^\star$

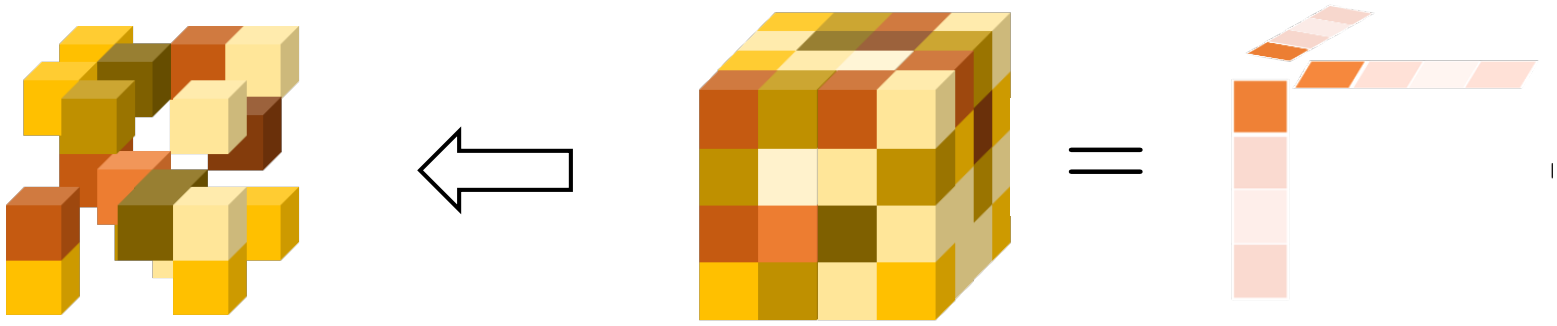
Extension: low-rank tensor completion



observed | true tensor $\boldsymbol{T}^\star$ | rank-1 tensor

- ground truth: rank-1 tensor $\boldsymbol{T}^\star = \boldsymbol{u}^\star \otimes \boldsymbol{u}^\star \otimes \boldsymbol{u}^\star \in \mathbb{R}^{n \times n \times n}$
- each entry $T^\star_{i,j,k}$ is observed independently with prob. p
- **goal:** fill in unseen entries of $\boldsymbol{T}^\star$

Extension: low-rank tensor completion



observed | true tensor $\boldsymbol{T}^\star$ | rank-1 tensor

- ground truth: rank-1 tensor $\boldsymbol{T}^\star = \boldsymbol{u}^\star \otimes \boldsymbol{u}^\star \otimes \boldsymbol{u}^\star \in \mathbb{R}^{n\times n\times n}$
- each entry $T^\star_{i,j,k}$ is observed independently with prob. p
- **goal:** fill in unseen entries of $\boldsymbol{T}^\star$

Can we exploit low-rank tensor structure in prediction?

Spectral method for low-rank tensor completion

1. construct a rescaled zero-filled tensor $\boldsymbol{T} = [T_{i,j,k}] \in \mathbb{R}^{n \times n \times n}$ as

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

   - rescaling ensures $\mathbb{E}[\boldsymbol{T}] = \boldsymbol{T}^{\star}$

2. matricizition: $\boldsymbol{A} = \text{unfold}(\boldsymbol{T})$

Spectral method for low-rank tensor completion

3. compute spectral estimates (after diagonal deletion):

$$\begin{array}{ll} \boldsymbol{u} & \leftarrow \text{leading eigenvector} \\ \lambda & \leftarrow \text{leading eigenvalue} \end{array} \quad \text{of } \mathcal{P}_{\text{off-diag}}\left(\boldsymbol{A}\boldsymbol{A}^\top\right) \text{ (remove diagonal)}$$

4. return $\widehat{\boldsymbol{T}} = \lambda \boldsymbol{u} \otimes \boldsymbol{u} \otimes \boldsymbol{u}$

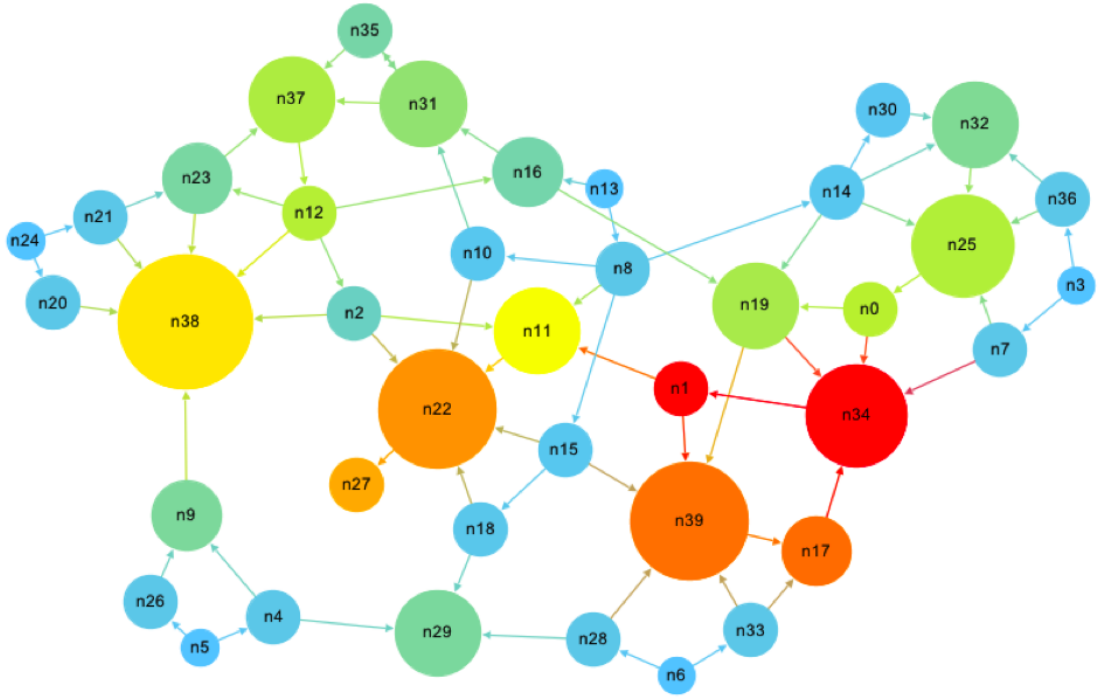
We will explain special treatments for diagonals

Motivating application: ranking

Ranking

A fundamental problem in a wide range of contexts

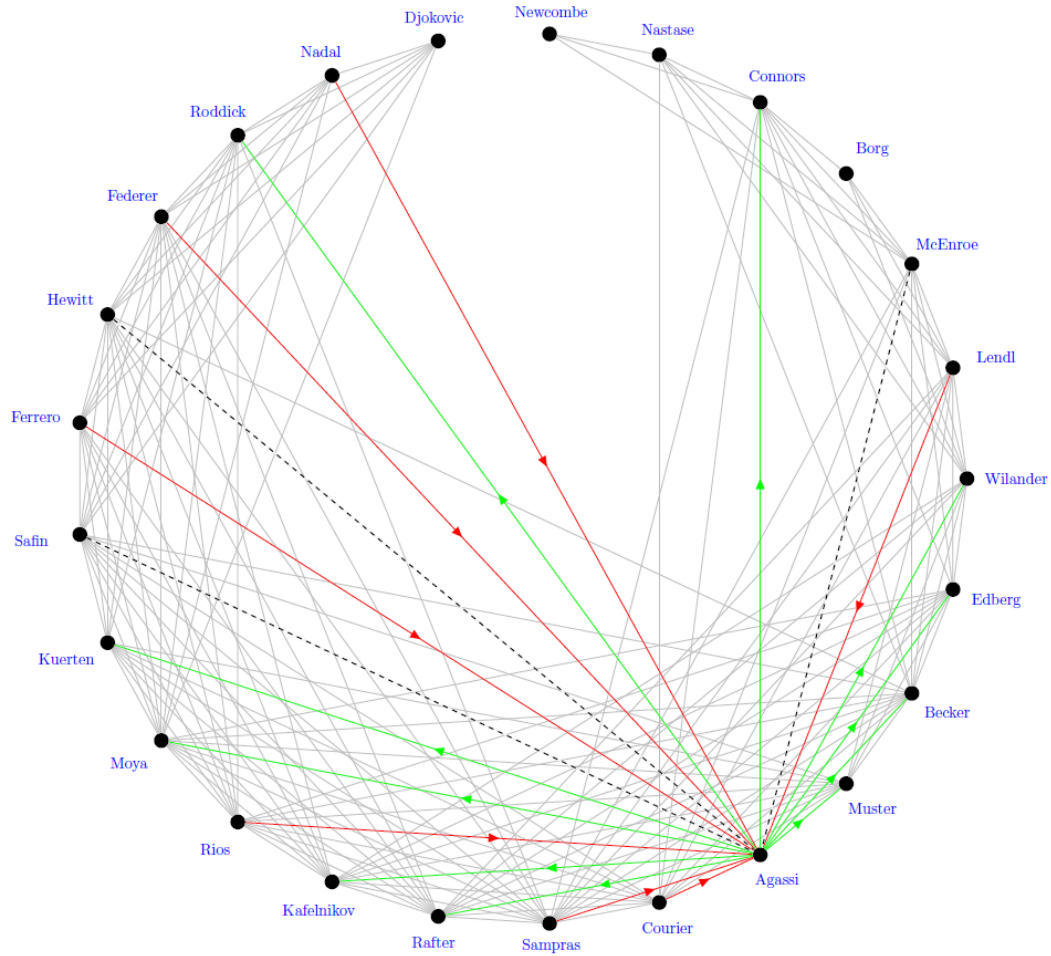
- web search, recommendation systems, admissions, sports competitions, voting, ...



PageRank

figure credit: Dzenan Hamzic

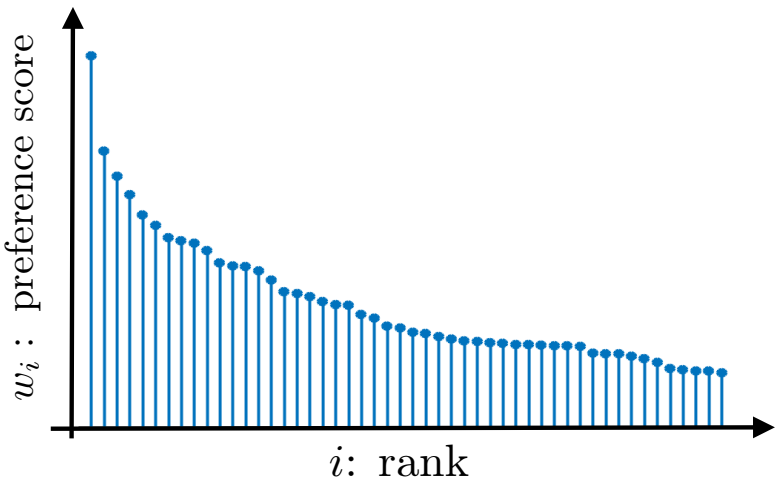
Ranking from pairwise comparisons



pairwise comparisons for ranking tennis players

figure credit: Bozóki, Csató, Temesi

Parametric models

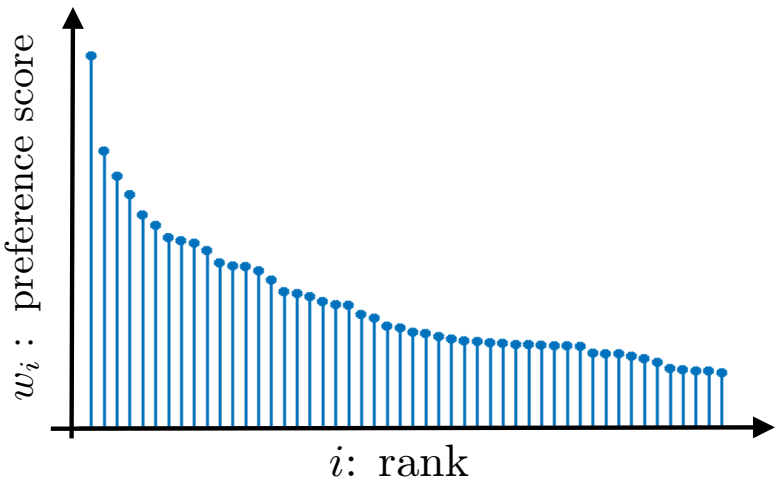


- n items to be ranked
- assign a latent score $\{w_i^\star\}_{1 \le i \le n}$ to each item, so that

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

- rank items <u>in accordance with score estimates</u>

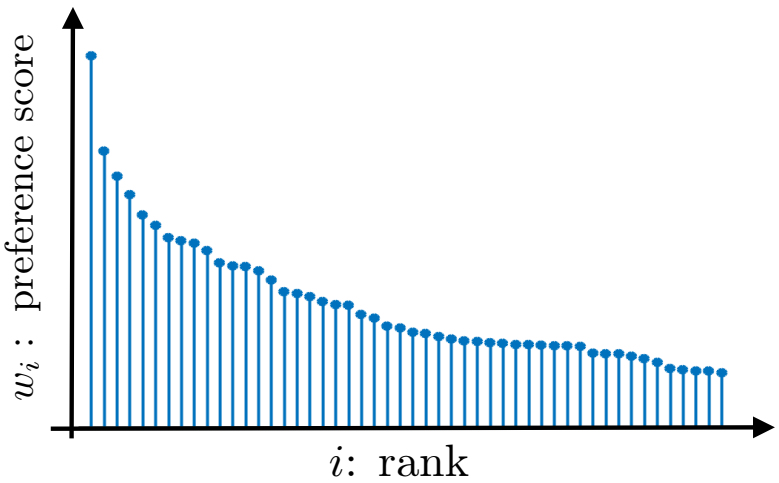
Bradley-Terry-Luce (logistic) model



- 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



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

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

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

Spectral ranking

Key idea: consider a probability transition matrix $\boldsymbol{P}^\star \in \mathbb{R}^{n \times n}$:

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

Spectral ranking

Key idea: consider a probability transition matrix $\boldsymbol{P}^\star \in \mathbb{R}^{n \times n}$:

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

- stationary distribution $\boldsymbol{\pi}^\star$ of $\boldsymbol{P}^\star$: $\boldsymbol{\pi}^\star = \frac{1}{\sum_l w^\star_l} \boldsymbol{w}^\star$

  leading left eigenvector of $\boldsymbol{P}^\star$

  - can be seen from detailed balance property: $w^\star_i P^\star_{i,j} = w^\star_j P^\star_{j,i}$

Spectral ranking

Key idea: consider a probability transition matrix $\boldsymbol{P}^\star \in \mathbb{R}^{n \times n}$:

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

- $\underbrace{\text{stationary distribution } \boldsymbol{\pi}^\star \text{ of } \boldsymbol{P}^\star}_{\text{leading left eigenvector of } \boldsymbol{P}^\star}$: $\boldsymbol{\pi}^\star = \frac{1}{\sum_l w^\star_l} \boldsymbol{w}^\star$
  - can be seen from detailed balance property: $w^\star_i P^\star_{i,j} = w^\star_j P^\star_{j,i}$

True ranks are revealed by leading left eigenvector of $\boldsymbol{P}^\star$

Spectral ranking

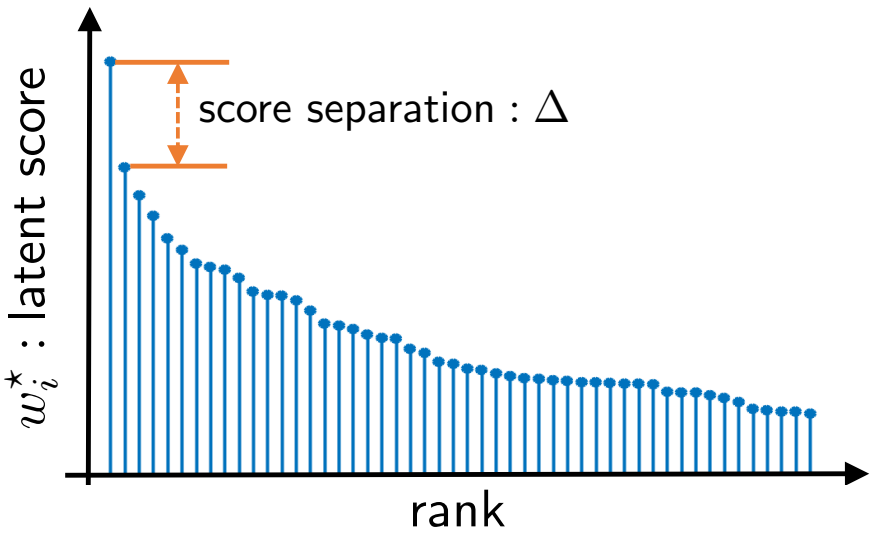
1. construct a surrogate matrix $\boldsymbol{P}$ obeying

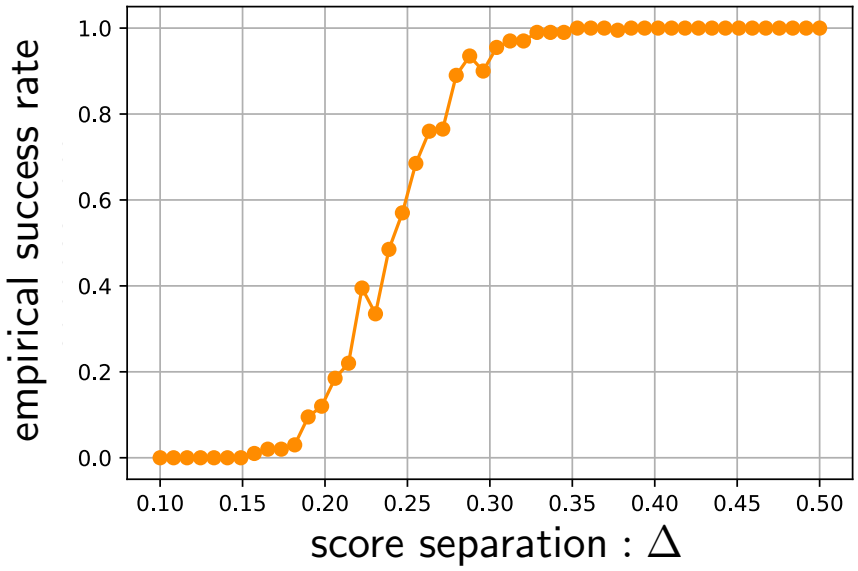
$$P_{i,j} = \begin{cases} \frac{1}{n} y_{i,j}, & \text{if } i \neq j \\ 1 - \sum_{l:l \neq i} P_{i,l}, & \text{if } i = j \end{cases}$$

2. compute leading left eigenvector $\boldsymbol{\pi}$ of $\boldsymbol{P}$ as score estimate
3. rank in accordance with $\boldsymbol{\pi}$

— closely related to PageRank

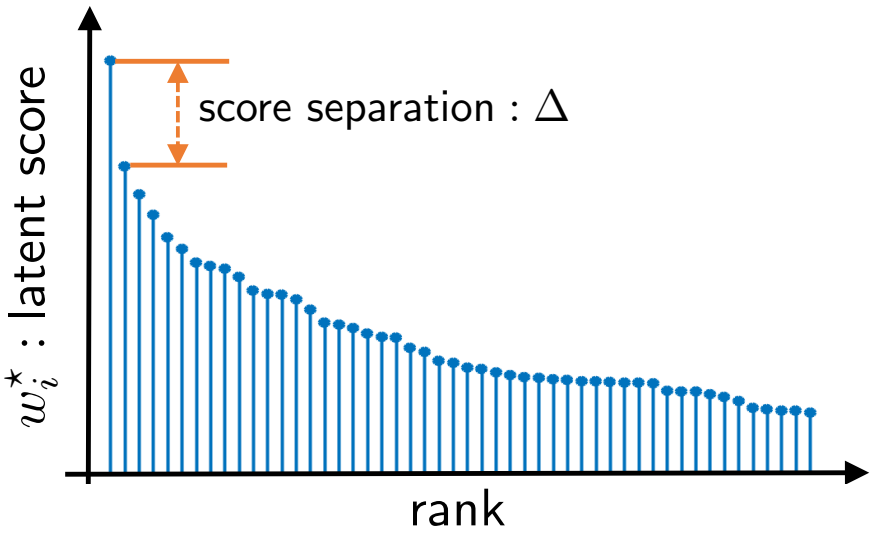
Empirical accuracy in finding top-ranked item

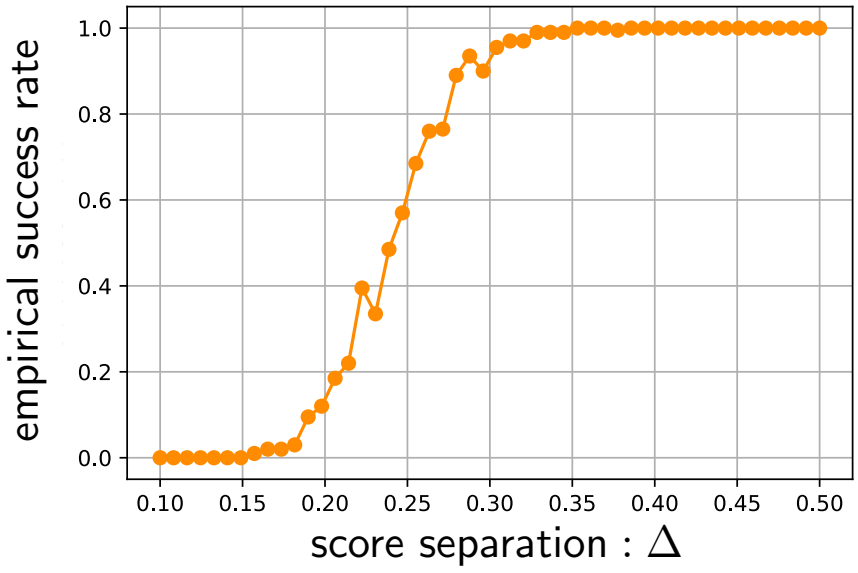




$n = 200$

Empirical accuracy in finding top-ranked item





$n = 200$

What we will demonstrate: efficacy of spectral ranking

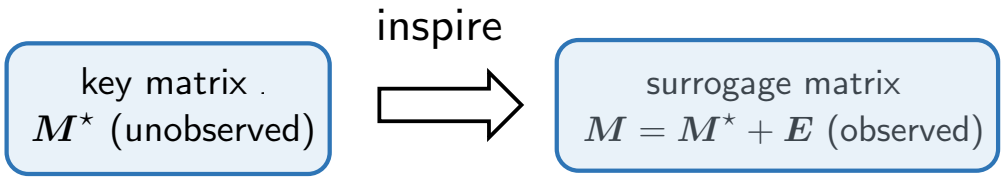
A unified recipe for spectral methods

A unified recipe

key matrix $M^\star$ (unobserved)

1. identify a key matrix $M^\star$ — typically unobserved — whose eigenvectors or singular vectors encode crucial information

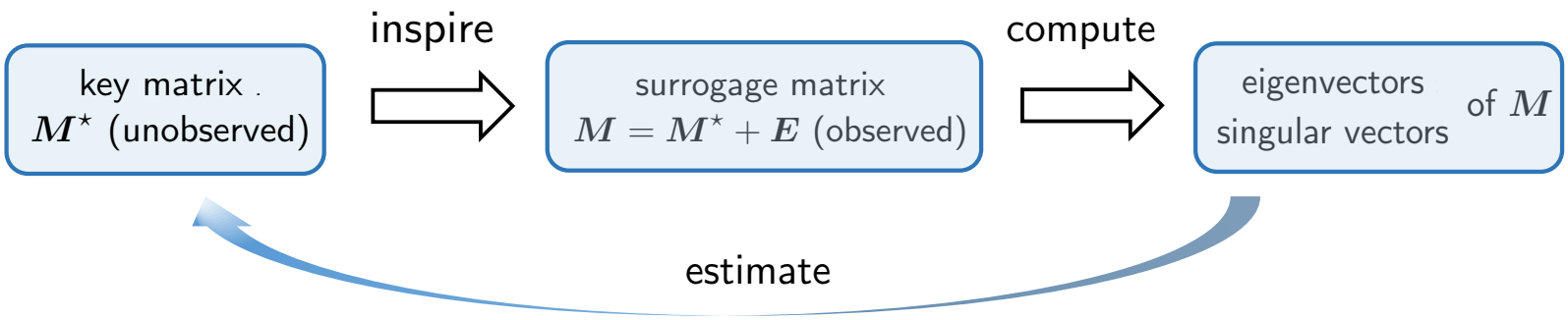
A unified recipe



1. identify a key matrix $M^\star$ — typically unobserved — whose eigenvectors or singular vectors encode crucial information

2. construct a surrogate matrix M of $M^\star$ using data samples

A unified recipe



1. identify a key matrix $M^\star$ — typically unobserved — whose eigenvectors or singular vectors encode crucial information

2. construct a surrogate matrix M of $M^\star$ using data samples

3. compute corresponding eigenvectors or singular vectors of M

Key factors

A few factors that dictate the performance of spectral methods:

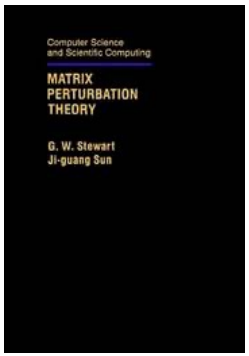
- proximity of $\boldsymbol{M}$ and $\boldsymbol{M}^\star$ (e.g. $\|\boldsymbol{M} - \boldsymbol{M}^\star\|$)
- spectrum (e.g. eigenvalues, singular values) of $\boldsymbol{M}^\star$
- ...

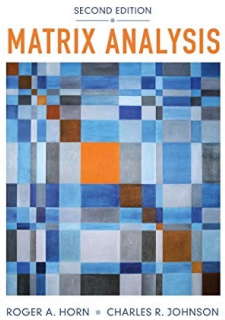
Key factors

A few factors that dictate the performance of spectral methods:

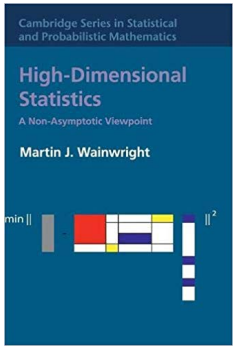
- proximity of $\boldsymbol{M}$ and $\boldsymbol{M}^\star$ (e.g. $\|\boldsymbol{M} - \boldsymbol{M}^\star\|$)
- spectrum (e.g. eigenvalues, singular values) of $\boldsymbol{M}^\star$
- ...

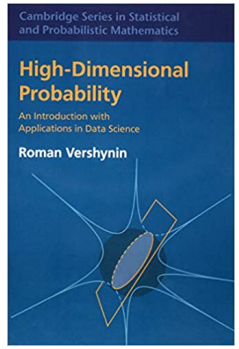
Aim of this tutorial: quantify influences of these factors





matrix perturbation theory





high-dimensional statistics probability

- **algebraic tools:** matrix perturbation theory (Part 2)
- **statistical & probabilistic tools:**
  - matrix concentration bounds (Part 3: ℓ_2 analysis)
  - leave-one-out analysis (Part 4: fine-grained analysis)

Asymptotic notation used in this tutorial

- $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \leq \text{const}$$

- $f(n) \gtrsim g(n)$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \geq \text{const}$$

- $f(n) \asymp g(n)$ means

$$\text{const}_1 \leq \lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \leq \text{const}_2$$

- $f(n) = o(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} = 0$$

Part 2: Matrix perturbation theory

- Eigen-space perturbation theory
  - Distances and angles between two subspaces
  - The Davis-Kahan $\sin\Theta$ theorem
- Singular subspace perturbation theory (Wedin's theorem)
- Eigenvector perturbation for probability transition matrices

Eigen-space perturbation theory

Setup and notation

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

$$\boldsymbol{M}^\star = \sum_{i=1}^{n} \lambda_i^\star \boldsymbol{u}_i^\star \boldsymbol{u}_i^{\star\top} = [\boldsymbol{U}^\star, \boldsymbol{U}_\perp^\star] \begin{bmatrix} \boldsymbol{\Lambda}^\star & \\ & \boldsymbol{\Lambda}_\perp^\star \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{\star\top} \\ \boldsymbol{U}_\perp^{\star\top} \end{bmatrix}$$

$$\boldsymbol{M} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top = [\boldsymbol{U}, \boldsymbol{U}_\perp] \begin{bmatrix} \boldsymbol{\Lambda} & \\ & \boldsymbol{\Lambda}_\perp \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{U}_\perp^\top \end{bmatrix}$$

- eigenvalues: $\lambda_1^\star \geq \cdots \geq \lambda_n^\star$, $\lambda_1 \geq \cdots \geq \lambda_n$
- $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{n \times r}$, $\boldsymbol{\Lambda} = \mathrm{diag}\left([\lambda_1, \cdots, \lambda_r]\right) \in \mathbb{R}^{r \times r}$, $\cdots$

Setup and notation

$$\boldsymbol{M} = \underbrace{\begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_r \end{bmatrix}}_{\boldsymbol{U}} \underbrace{\begin{bmatrix} \boldsymbol{u}_{r+1} & \cdots & \boldsymbol{u}_n \end{bmatrix}}_{=: \boldsymbol{U}_\perp} \cdot \begin{bmatrix} \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}}_{=: \boldsymbol{\Lambda}} & \\ & \underbrace{\begin{bmatrix} \lambda_{r+1} & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}}_{=: \boldsymbol{\Lambda}_\perp} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1^\top \\ \vdots \\ \boldsymbol{u}_r^\top \\ \boldsymbol{u}_{r+1}^\top \\ \vdots \\ \boldsymbol{u}_n^\top \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix}}\right\} =: \boldsymbol{U}^\top \\ \left.\vphantom{\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix}}\right\} =: \boldsymbol{U}_\perp^\top \end{matrix}$$

Setup and notation

- $\mathcal{O}^{r \times r}$: set of all $r \times r$ orthonormal matrices
- $\|\boldsymbol{M}\|$: 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 perturbation matrix $\boldsymbol{E}$ affect "distance" between $\boldsymbol{U}^\star$ and $\boldsymbol{U}$?

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

Eigen-space perturbation theory

Main focus: how does perturbation matrix $\boldsymbol{E}$ affect "distance" between $\boldsymbol{U}^\star$ and $\boldsymbol{U}$?

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

- $\|\boldsymbol{U} - \boldsymbol{U}^\star\|_{\mathrm{F}}$ and $\|\boldsymbol{U} - \boldsymbol{U}^\star\|$ are not appropriate, since they fall short of accounting for $\underbrace{\text{global orthonormal transformation}}$

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

Distances and angles between two subspaces

Two valid choices of distance metrics

Key: taking care of global orthonormal transformation

- **Distance modulo optimal rotation**: adjust for rotation before computing distance:

$$\mathsf{dist}\big(\boldsymbol{U}, \boldsymbol{U}^\star\big) := \min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \big\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\big\| \tag{2.1}$$

Two valid choices of distance metrics

Key: taking care of global orthonormal transformation

- **Distance modulo optimal rotation**: adjust for rotation before computing distance:

$$\mathsf{dist}\big(\boldsymbol{U}, \boldsymbol{U}^\star\big) := \min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \big\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\big\| \tag{2.1}$$

- **Distance using projection matrices**: replace $\boldsymbol{U}$ (resp. $\boldsymbol{U}^\star$) with its associated projection matrix before computing distance:

$$\mathsf{dist}_{\mathsf{p}}\big(\boldsymbol{U}, \boldsymbol{U}^\star\big) := \big\| \underbrace{\boldsymbol{U}\boldsymbol{U}^\top}_{\text{projection onto subspace } \boldsymbol{U}} - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\big\| \tag{2.2}$$

(Near)-equivalence of two distance metrics

$$\mathsf{dist}(\boldsymbol{U}, \boldsymbol{U}^\star) := \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star \right\|$$
$$\mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star) := \left\| \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star \boldsymbol{U}^{\star\top} \right\|$$

Lemma 2.1

Suppose $[\boldsymbol{U}, \boldsymbol{U}_\perp]$, $[\boldsymbol{U}^\star, \boldsymbol{U}_\perp^\star]$ *are square orthonormal matrices. Then*

$$\mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star) \le \mathsf{dist}(\boldsymbol{U}, \boldsymbol{U}^\star) \le \sqrt{2}\, \mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star)$$

- $\mathsf{dist}(\cdot, \cdot)$ and $\mathsf{dist}_{\mathsf{p}}(\cdot, \cdot)$ are orderwise equivalent
- proof: see Section 2.6.3 of Chen et al. 21

An alternative expression for $\mathsf{dist}_{\mathsf{p}}(\cdot, \cdot)$

As it turns out, $\mathsf{dist}_{\mathsf{p}}(\cdot, \cdot)$ has several equivalent expressions:

Lemma 2.2

Recall that $[\boldsymbol{U}, \boldsymbol{U}_\perp]$, $[\boldsymbol{U}^\star, \boldsymbol{U}_\perp^\star]$ *are square orthonormal matrices. Then*

$$\mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star) = \|\boldsymbol{U}^\top \boldsymbol{U}_\perp^\star\| = \|\boldsymbol{U}^{\star\top} \boldsymbol{U}_\perp\|$$

An alternative expression for $\mathsf{dist}_{\mathsf{p}}(\cdot, \cdot)$

As it turns out, $\mathsf{dist}_{\mathsf{p}}(\cdot, \cdot)$ has several equivalent expressions:

Lemma 2.2

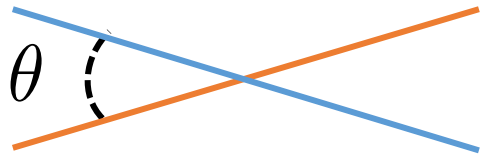
Recall that $[\boldsymbol{U}, \boldsymbol{U}_\perp]$, $[\boldsymbol{U}^\star, \boldsymbol{U}_\perp^\star]$ *are square orthonormal matrices. Then*

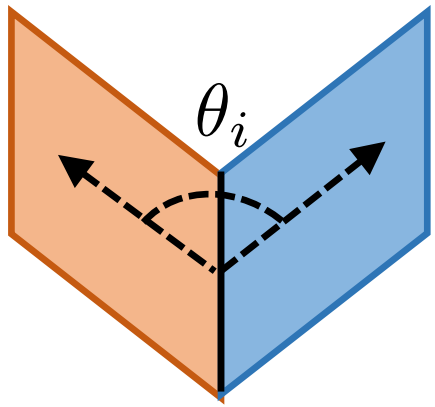
$$\mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star) = \|\boldsymbol{U}^\top \boldsymbol{U}_\perp^\star\| = \|\boldsymbol{U}^{\star\top} \boldsymbol{U}_\perp\|$$

- sanity check: if $\boldsymbol{U} = \boldsymbol{U}^\star$, then $\mathsf{dist}(\boldsymbol{U}, \boldsymbol{U}^\star) = \|\boldsymbol{U}^\top \boldsymbol{U}^\star\| = 0$
- proof: see Slide 2-15

Principal angles between two subspaces

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





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

$$\theta = \arccos\langle \boldsymbol{u}, \boldsymbol{u}^\star \rangle$$

Principal angles between two eigen-spaces

For r-dimensional subspaces, one needs r angles

Specifically, given $\|\boldsymbol{U}^\top \boldsymbol{U}^\star\| \leq 1$, write the SVD of $\boldsymbol{U}^\top \boldsymbol{U}^\star \in \mathbb{R}^{r\times r}$ as

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

- $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{r\times r}$: square orthonormal matrices
- $\{\theta_1, \cdots, \theta_r\}$ are called the principal angles between $\boldsymbol{U}$ and $\boldsymbol{U}^\star$

Relations between principal angles and distance

As expected, principal angles and distances are closely related

Lemma 2.3

Suppose $[\boldsymbol{U}, \boldsymbol{U}_\perp]$, $[\boldsymbol{U}^\star, \boldsymbol{U}_\perp^\star]$ *are square orthonormal matrices. Then*

$$\|\boldsymbol{U}^\top \boldsymbol{U}_\perp^\star\| = \|\sin \boldsymbol{\Theta}\| = \max\{|\sin\theta_1|, \cdots, |\sin\theta_r|\}$$

Lemmas 2.2 and 2.3 taken collectively give

$$\mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star) = \max\{|\sin\theta_1|, \cdots, |\sin\theta_r|\} \tag{2.3}$$

Proof of Lemma 2.3

$$\|\boldsymbol{U}^\top \boldsymbol{U}_\perp^\star\| = \|\boldsymbol{U}^\top \underbrace{\boldsymbol{U}_\perp^\star \boldsymbol{U}_\perp^{\star\top}}_{=\boldsymbol{I}-\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}} \boldsymbol{U}\|^{\frac{1}{2}}$$

$$= \|\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{U}^\top\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{U}\|^{\frac{1}{2}}$$

$$= \|\boldsymbol{I} - \boldsymbol{X}\cos^2\boldsymbol{\Theta}\,\boldsymbol{X}^\top\|^{\frac{1}{2}} \qquad (\text{since } \boldsymbol{U}^\top\boldsymbol{U}^\star = \boldsymbol{X}\cos\boldsymbol{\Theta}\,\boldsymbol{Y}^\top)$$

$$= \|\boldsymbol{I} - \cos^2\boldsymbol{\Theta}\|^{\frac{1}{2}}$$

$$= \|\sin\boldsymbol{\Theta}^2\|^{\frac{1}{2}}$$

$$= \|\sin\boldsymbol{\Theta}\|$$

Proof of Lemma 2.2

We first claim that the SVD of $\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star$ can be written as

$$\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star = \widetilde{\boldsymbol{X}} \sin\boldsymbol{\Theta}\, \boldsymbol{Y}^\top \tag{2.4}$$

for some orthonormal $\widetilde{\boldsymbol{X}}$ (to be proved later). Armed w/ this claim, one has

$$\boldsymbol{U}^\star = [\boldsymbol{U}, \boldsymbol{U}_\perp] \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{U}_\perp^\top \end{bmatrix} \boldsymbol{U}^\star = [\boldsymbol{U}, \boldsymbol{U}_\perp] \begin{bmatrix} \boldsymbol{X} \cos\boldsymbol{\Theta}\, \boldsymbol{Y}^\top \\ \widetilde{\boldsymbol{X}} \sin\boldsymbol{\Theta}\, \boldsymbol{Y}^\top \end{bmatrix}$$

$$\Longrightarrow \boldsymbol{U}^\star \boldsymbol{U}^{\star\top} = [\boldsymbol{U}, \boldsymbol{U}_\perp] \begin{bmatrix} \boldsymbol{X} \cos^2\boldsymbol{\Theta}\, \boldsymbol{X}^\top & \boldsymbol{X} \cos\boldsymbol{\Theta} \sin\boldsymbol{\Theta}\, \widetilde{\boldsymbol{X}}^\top \\ \widetilde{\boldsymbol{X}} \cos\boldsymbol{\Theta} \sin\boldsymbol{\Theta}\, \boldsymbol{X}^\top & \widetilde{\boldsymbol{X}} \sin^2\boldsymbol{\Theta}\, \widetilde{\boldsymbol{X}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{U}_\perp^\top \end{bmatrix}$$

As a consequence,

$$\begin{aligned} &\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star \boldsymbol{U}^{\star\top} \\ &\quad = [\boldsymbol{U}, \boldsymbol{U}_\perp] \begin{bmatrix} \boldsymbol{I} - \boldsymbol{X} \cos^2\boldsymbol{\Theta}\, \boldsymbol{X}^\top & -\boldsymbol{X} \cos\boldsymbol{\Theta} \sin\boldsymbol{\Theta}\, \widetilde{\boldsymbol{X}}^\top \\ -\widetilde{\boldsymbol{X}} \cos\boldsymbol{\Theta} \sin\boldsymbol{\Theta}\, \boldsymbol{X}^\top & -\widetilde{\boldsymbol{X}} \sin^2\boldsymbol{\Theta}\, \widetilde{\boldsymbol{X}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^\top \\ \boldsymbol{U}_\perp^\top \end{bmatrix} \end{aligned}$$

Proof of Lemma 2.2 (cont.)

This further gives

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

Proof of Lemma 2.2 (cont.)

It remains to justify (2.4). To this end, observe that

$$\begin{aligned}\boldsymbol{U}^{\star\top}\boldsymbol{U}_\perp\boldsymbol{U}_\perp^\top\boldsymbol{U}^\star &= \boldsymbol{U}^{\star\top}\boldsymbol{U}^\star - \boldsymbol{U}^{\star\top}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}^\star \\ &= \boldsymbol{I} - \boldsymbol{Y}\cos^2\boldsymbol{\Theta}\,\boldsymbol{Y}^\top \\ &= \boldsymbol{Y}\sin^2\boldsymbol{\Theta}\,\boldsymbol{Y}^\top\end{aligned}$$

and hence the right singular space (resp. singular values) of $\boldsymbol{U}^{\star\top}\boldsymbol{U}_\perp$ is given by $\boldsymbol{Y}$ (resp. $\sin\boldsymbol{\Theta}$). This immediately implies (2.4).

Summary: four (almost) equivalent distance metrics

$$
\begin{aligned}
&1) \quad \left\| \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star \boldsymbol{U}^{\star\top} \right\| \\
&2) \quad \left\| \sin \boldsymbol{\Theta} \right\| \\
&3) \quad \left\| \boldsymbol{U}_\perp^\top \boldsymbol{U}^\star \right\| = \left\| \boldsymbol{U}^\top \boldsymbol{U}_\perp^\star \right\| \\
&4) \quad \min_{\boldsymbol{R} \in \mathcal{O}^{r\times r}} \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star \right\|
\end{aligned}
$$

Summary: four (almost) equivalent distance metrics

$$
\begin{aligned}
&1) \quad \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\right\| \\
&2) \quad \left\|\sin\boldsymbol{\Theta}\right\| \\
&3) \quad \left\|\boldsymbol{U}_\perp^\top\boldsymbol{U}^\star\right\| = \left\|\boldsymbol{U}^\top\boldsymbol{U}_\perp^\star\right\| \\
&4) \quad \min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\right\|
\end{aligned}
$$

Near-equivalence of these metrics continue to hold if $\|\cdot\|$ is replaced by $\|\cdot\|_{\mathrm{F}}$

The Davis-Kahan $\sin\boldsymbol{\Theta}$ *theorem*

Warm-up example ($0 < \epsilon < 1$):

$$\boldsymbol{M}^\star = \begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1-\epsilon \end{bmatrix}, \quad \boldsymbol{E} = \begin{bmatrix} -\epsilon & \epsilon \\ \epsilon & \epsilon \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix}$$

Warm-up example ($0 < \epsilon < 1$):

$$\boldsymbol{M}^\star = \begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1-\epsilon \end{bmatrix}, \quad \boldsymbol{E} = \begin{bmatrix} -\epsilon & \epsilon \\ \epsilon & \epsilon \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix}$$

- leading eigenvectors of $\boldsymbol{M}^\star$ and $\boldsymbol{M}$:

$$\boldsymbol{u}_1^\star = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{u}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \Longrightarrow \quad \left\| \boldsymbol{u}_1 \boldsymbol{u}_1^\top - \boldsymbol{u}_1^\star \boldsymbol{u}_1^{\star\top} \right\|_2 = \frac{1}{\sqrt{2}}$$

  — *eigenvector distance is large regardless of size of ϵ (or size of $\|\boldsymbol{E}\|$)*

Warm-up example ($0 < \epsilon < 1$):

$$\boldsymbol{M}^\star = \begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1-\epsilon \end{bmatrix}, \quad \boldsymbol{E} = \begin{bmatrix} -\epsilon & \epsilon \\ \epsilon & \epsilon \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix}$$

- leading eigenvectors of $\boldsymbol{M}^\star$ and $\boldsymbol{M}$:

$$\boldsymbol{u}_1^\star = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{u}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \Longrightarrow \quad \left\| \boldsymbol{u}_1 \boldsymbol{u}_1^\top - \boldsymbol{u}_1^\star \boldsymbol{u}_1^{\star\top} \right\|_2 = \frac{1}{\sqrt{2}}$$

— eigenvector distance is large regardless of size of ϵ (or size of $\|\boldsymbol{E}\|$)

Diagonsis: eigen-gap $\lambda_1^\star - \lambda_2^\star = 2\epsilon$ also small (proportional to $\|\boldsymbol{E}\|$)

— both perturbation size and eigen-gap might play important roles

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

— recall the setup in Page 2-3

Chandler Davis

William Kahan

Theorem 2.4

Suppose $\boldsymbol{M}^\star \succeq \boldsymbol{0}$ and has rank r. If $\|\boldsymbol{E}\| < (1-1/\sqrt{2})\lambda_r^\star$, then

$$\mathsf{dist}_{\mathsf{p}}\big(\boldsymbol{U},\boldsymbol{U}^\star\big) = \big\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\big\| = \|\sin\boldsymbol{\Theta}\| \le \frac{\sqrt{2}\|\boldsymbol{E}\boldsymbol{U}^\star\|}{\lambda_r^\star} \le \frac{\sqrt{2}\|\boldsymbol{E}\|}{\lambda_r^\star}$$

Interpretations

Suppose $\boldsymbol{M}^\star \succeq \boldsymbol{0}$ and has rank r. If $\|\boldsymbol{E}\| < (1-1/\sqrt{2})\lambda_r^\star$, then

$$\mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star) \leq \frac{\sqrt{2}\|\boldsymbol{E}\boldsymbol{U}^\star\|}{\lambda_r^\star} \leq \frac{\sqrt{2}\|\boldsymbol{E}\|}{\lambda_r^\star}$$

Key factors:

1. eigen-gap: $\lambda_r^\star = \lambda_r^\star - \underbrace{\lambda_{r+1}^\star}_{=0}$
2. perturbation size: $\|\boldsymbol{E}\|$
3. signal-to-noise ratio (SNR): $\frac{\lambda_r^\star}{\|\boldsymbol{E}\|}$

- the bound w/ $\|\boldsymbol{E}\boldsymbol{U}^\star\|$ is sometimes useful (e.g. for ℓ_∞ analysis)

Proof of Theorem 2.4

We intend to control $\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star$ by studying their interactions through $\boldsymbol{E}$:

$$\begin{aligned}
\left\|\boldsymbol{U}_\perp^\top \boldsymbol{E}\boldsymbol{U}^\star\right\| &= \Big\|\boldsymbol{U}_\perp^\top \Big(\underbrace{\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \boldsymbol{U}_\perp \boldsymbol{\Lambda}_\perp \boldsymbol{U}_\perp^\top}_{\boldsymbol{M}^\star + \boldsymbol{E}} - \underbrace{\boldsymbol{U}^\star \boldsymbol{\Lambda}^\star \boldsymbol{U}^{\star\top}}_{\boldsymbol{M}^\star}\Big)\boldsymbol{U}^\star\Big\| \\
&= \left\|\boldsymbol{\Lambda}_\perp \boldsymbol{U}_\perp^\top \boldsymbol{U}^\star - \boldsymbol{U}_\perp^\top \boldsymbol{U}^\star \boldsymbol{\Lambda}^\star\right\| \qquad \text{(since } \boldsymbol{U}_\perp^{\star\top}\boldsymbol{U}^\star = \boldsymbol{U}_\perp^\top \boldsymbol{U} = \boldsymbol{0}\text{)} \\
&\geq \left\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star \boldsymbol{\Lambda}^\star\right\| - \left\|\boldsymbol{\Lambda}_\perp \boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\right\| \qquad \text{(triangle inequality)} \\
&\geq \left\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\right\| \lambda_r^\star - \left\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\right\| \|\boldsymbol{\Lambda}_\perp\| \qquad (2.5)
\end{aligned}$$

Weyl's Theorem gives $\|\boldsymbol{\Lambda}_\perp\| \leq \|\boldsymbol{E}\|$, which combined with (2.5) yields

$$\left\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\right\| \leq \frac{\left\|\boldsymbol{U}_\perp^\top \boldsymbol{E}\boldsymbol{U}^\star\right\|}{\lambda_r^\star - \|\boldsymbol{E}\|} \leq \frac{\|\boldsymbol{U}_\perp\| \cdot \|\boldsymbol{E}\boldsymbol{U}^\star\|}{\lambda_r^\star - \|\boldsymbol{E}\|} = \frac{\|\boldsymbol{E}\boldsymbol{U}^\star\|}{\lambda_r^\star - \|\boldsymbol{E}\|}$$

This together with assumption $\|\boldsymbol{E}\| \leq (1 - \sqrt{2}/2)\lambda_r^\star$ and Lemmas 2.2-2.3 completes the proof

Davis-Kahan's $\sin\Theta$ theorem: general case

— eigenvalues($\boldsymbol{A}$)*: set of eigenvalues of* $\boldsymbol{A}$

Theorem 2.5 (Davis-Kahan's sinΘ theorem: general version)

Assume that

$$\text{eigenvalues}(\boldsymbol{\Lambda}^\star) \subseteq (-\infty, \alpha - \Delta] \cup [\beta + \Delta, \infty); \tag{2.6a}$$

$$\text{eigenvalues}(\boldsymbol{\Lambda}_\perp^\star) \subseteq [\alpha, \beta]. \tag{2.6b}$$

for some eigengap $\Delta > 0$*. Suppose* $\|\boldsymbol{E}\| \le (1 - \sqrt{2}/2)\Delta$*. Then*

$$\mathsf{dist}(\boldsymbol{U}, \boldsymbol{U}^\star) \le \sqrt{2}\,\mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star) = \sqrt{2}\,\|\sin\boldsymbol{\Theta}\| \le \frac{2\|\boldsymbol{E}\boldsymbol{U}^\star\|}{\Delta} \le \frac{2\|\boldsymbol{E}\|}{\Delta}$$

Davis-Kahan's $\sin\Theta$ theorem: general case

— eigenvalues($\boldsymbol{A}$)*: set of eigenvalues of* $\boldsymbol{A}$

Theorem 2.5 (Davis-Kahan's sinΘ theorem: general version)

Assume that

$$\text{eigenvalues}(\boldsymbol{\Lambda}^\star) \subseteq (-\infty, \alpha - \Delta] \cup [\beta + \Delta, \infty); \tag{2.6a}$$

$$\text{eigenvalues}(\boldsymbol{\Lambda}_\perp^\star) \subseteq [\alpha, \beta]. \tag{2.6b}$$

for some eigengap $\Delta > 0$*. Suppose* $\|\boldsymbol{E}\| \le (1 - \sqrt{2}/2)\Delta$*. Then*

$$\mathsf{dist}(\boldsymbol{U}, \boldsymbol{U}^\star) \le \sqrt{2}\,\mathsf{dist}_{\mathsf{p}}(\boldsymbol{U}, \boldsymbol{U}^\star) = \sqrt{2}\,\|\sin\boldsymbol{\Theta}\| \le \frac{2\|\boldsymbol{E}\boldsymbol{U}^\star\|}{\Delta} \le \frac{2\|\boldsymbol{E}\|}{\Delta}$$

- conclusion remains valid if Assumption (2.6) is reversed
- proof: see Section 2.3.4 of Chen et al. '21

Singular subspace perturbation theory

Setup and notation

Consider 2 matrices $\boldsymbol{M}^\star, \boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{E} \in \mathbb{R}^{n_1 \times n_2}$ $(n_1 \le n_2)$ w/ SVD

$$\boldsymbol{M}^\star = \sum_{i=1}^{n_1} \sigma_i^\star \boldsymbol{u}_i^\star \boldsymbol{v}_i^{\star\top} = \begin{bmatrix} \boldsymbol{U}^\star & \boldsymbol{U}_\perp^\star \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}^\star & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_\perp^\star & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^{\star\top} \\ \boldsymbol{V}_\perp^{\star\top} \end{bmatrix}$$

$$\boldsymbol{M} = \sum_{i=1}^{n_1} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{U}_\perp \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_\perp & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^\top \\ \boldsymbol{V}_\perp^\top \end{bmatrix}$$

- $\sigma_1 \ge \cdots \ge \sigma_{n_1}$: singular values of $\boldsymbol{M}$
- $\sigma_1^\star \ge \cdots \ge \sigma_{n_1}^\star$: singular values of $\boldsymbol{M}^\star$
- $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r] \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{\Sigma} = \mathsf{diag}\big(\lambda_1, \ldots, \lambda_r\big) \in \mathbb{R}^{r \times r}$, $\ldots$

Wedin's $\sin\Theta$ theorem

Davis-Kahan's theorem generalizes to singular subspace perturbation:

Theorem 2.6 (Wedin's sinΘ theorem)

If $\|\boldsymbol{E}\| < (1-1/\sqrt{2})\underbrace{(\sigma_r^\star - \sigma_{r+1}^\star)}_{\text{spectral gap}}$, *then one has*

$$\max\left\{\mathsf{dist}\big(\boldsymbol{U},\boldsymbol{U}^\star\big),\mathsf{dist}\big(\boldsymbol{V},\boldsymbol{V}^\star\big)\right\} \le \sqrt{2}\max\left\{\mathsf{dist}_{\mathsf{p}}\big(\boldsymbol{U},\boldsymbol{U}^\star\big),\mathsf{dist}_{\mathsf{p}}\big(\boldsymbol{V},\boldsymbol{V}^\star\big)\right\}$$
$$\le \frac{2\max\left\{\|\boldsymbol{E}^\top\boldsymbol{U}^\star\|,\|\boldsymbol{E}\boldsymbol{V}^\star\|\right\}}{\sigma_r^\star-\sigma_{r+1}^\star} \le \frac{2\|\boldsymbol{E}\|}{\sigma_r^\star-\sigma_{r+1}^\star}$$

- both $\boldsymbol{E}\boldsymbol{V}^\star$ and $\boldsymbol{E}^\top\boldsymbol{U}^\star$ matter

Proof of Theorem 2.6

Similar to proof of Davis-Kahan theorem, we concentrate on $\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star$:

$$\begin{aligned}
\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star &= \boldsymbol{U}_\perp^\top (\boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}) \boldsymbol{V}^\star \boldsymbol{\Sigma}^{\star -1} \\
&= \boldsymbol{U}_\perp^\top \Big(\underbrace{\boldsymbol{M} - \boldsymbol{E}}_{=\boldsymbol{M}^\star} - \boldsymbol{U}_\perp^\star \boldsymbol{\Sigma}_\perp^\star \boldsymbol{V}_\perp^{\star\top} \Big) \boldsymbol{V}^\star \boldsymbol{\Sigma}^{\star -1} \\
&= \boldsymbol{U}_\perp^\top \Big(\underbrace{\cancel{\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top} + \boldsymbol{U}_\perp \boldsymbol{\Sigma}_\perp \boldsymbol{V}_\perp^\top}_{=\boldsymbol{M}} - \boldsymbol{E} - \cancel{\boldsymbol{U}_\perp^\star \boldsymbol{\Sigma}_\perp^\star \boldsymbol{V}_\perp^{\star\top}} \Big) \boldsymbol{V}^\star \boldsymbol{\Sigma}^{\star -1} \\
&= \boldsymbol{\Sigma}_\perp \boldsymbol{V}_\perp^\top \boldsymbol{V}^\star \boldsymbol{\Sigma}^{\star -1} - \boldsymbol{U}_\perp^\top \boldsymbol{E} \boldsymbol{V}^\star \boldsymbol{\Sigma}^{\star -1}
\end{aligned}$$

Applying triangle inequality and Weyls' inequality yields

$$\begin{aligned}
\big\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\big\| &\le \|\boldsymbol{\Sigma}_\perp\| \cdot \big\|\boldsymbol{V}_\perp^\top \boldsymbol{V}^\star\big\| \cdot \|\boldsymbol{\Sigma}^{\star -1}\| + \|\boldsymbol{U}_\perp^\top\| \cdot \big\|\boldsymbol{E}\boldsymbol{V}^\star\big\| \cdot \|\boldsymbol{\Sigma}^{\star -1}\| \\
&= \sigma_{r+1} \cdot \big\|\boldsymbol{V}_\perp^\top \boldsymbol{V}^\star\big\| \cdot \frac{1}{\sigma_r^\star} + \big\|\boldsymbol{E}\boldsymbol{V}^\star\big\| \cdot \frac{1}{\sigma_r^\star} \\
&\le \frac{\sigma_{r+1}^\star + \|\boldsymbol{E}\|}{\sigma_r^\star} \big\|\boldsymbol{V}_\perp^\top \boldsymbol{V}^\star\big\| + \frac{\big\|\boldsymbol{E}\boldsymbol{V}^\star\big\|}{\sigma_r^\star} \qquad (2.7)
\end{aligned}$$

Proof of Theorem 2.6 (cont.)

Repeating the same argument yields

$$\|\boldsymbol{V}_\perp^\top \boldsymbol{V}^\star\| \le \frac{\sigma_{r+1}^\star + \|\boldsymbol{E}\|}{\sigma_r^\star}\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\| + \frac{\|\boldsymbol{E}^\top \boldsymbol{U}^\star\|}{\sigma_r^\star} \tag{2.8}$$

Combine inequalities (2.7) and (2.8) to obtain

$$\begin{aligned}\max\left\{\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\|, \|\boldsymbol{V}_\perp^\top \boldsymbol{V}^\star\|\right\} \le{}& \frac{\max\left\{\|\boldsymbol{E}^\top \boldsymbol{U}^\star\|, \|\boldsymbol{E}\boldsymbol{V}^\star\|\right\}}{\sigma_r^\star} \\ &+ \frac{\sigma_{r+1}^\star + \|\boldsymbol{E}\|}{\sigma_r^\star}\max\left\{\|\boldsymbol{U}_\perp^\top \boldsymbol{U}^\star\|, \|\boldsymbol{V}_\perp^\top \boldsymbol{V}^\star\|\right\}\end{aligned}$$

Rearrange terms and utilize $\|\boldsymbol{E}\| < (1-\sqrt{2}/2)(\sigma_r^\star - \sigma_{r+1}^\star)$ and Lemma 2.1 to arrive at desired result

Eigenvector perturbation for probability transition matrices

Eigen-decomposition for asymmetric matrices

Eigen-decomposition for asymmetric matrices is more tricky:

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

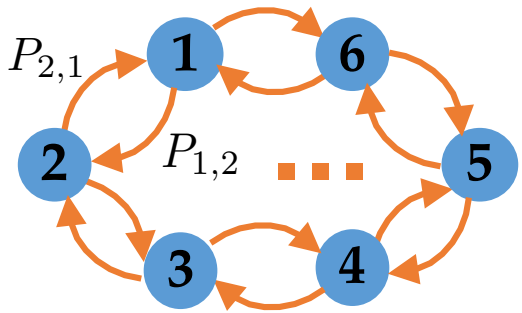
Eigen-decomposition for asymmetric matrices

Eigen-decomposition for asymmetric matrices is more tricky:

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

Let us look at 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 $\boldsymbol{\pi} = [\pi_i]_{1\leq i\leq n}$ obeys

$$\boldsymbol{\pi} \geq \mathbf{0}, \qquad \mathbf{1}^\top \boldsymbol{\pi} = 1, \qquad \text{and} \qquad \boldsymbol{\pi}^\top \boldsymbol{P} = \boldsymbol{\pi}^\top$$

  - leading left eigenvector of $\boldsymbol{P}$ with eigenvalue 1

Reversible Markov chains

Markov chain $\{X_t\}_{t\geq 0}$ with transition matrix $\boldsymbol{P}$ and stationary distribution $\boldsymbol{\pi}$ is said to be reversible if

$$\pi_i P_{i,j} = \pi_j P_{j,i} \quad \text{for all } i, j$$

— detailed balance condition

- If $\boldsymbol{P}$ represents reversible chain, then all eigenvalues of $\boldsymbol{P}$ are real

Setup and notation

- $\boldsymbol{P}^\star$: probability transition matrix of a reversible Markov chain
- $\boldsymbol{P} = \boldsymbol{P}^\star + \boldsymbol{E}$: (perturbed) probability transition matrix
- $\boldsymbol{\pi}^\star$ (resp. $\boldsymbol{\pi}$): leading left eigenvectors of $\boldsymbol{P}^\star$ (resp. $\boldsymbol{P}$)

Question: how does perturbation $\boldsymbol{E}$ affect leading left eigenvector?

Additional notation: for any probability vector $\boldsymbol{\pi} = [\pi_i]_{1 \le i \le n} > \mathbf{0}$:

- vector norm: $\|\boldsymbol{x}\|_{\boldsymbol{\pi}} := \sqrt{\sum_i \pi_i x_i^2}$ with $\boldsymbol{x} = [x_i]_{1 \le i \le n}$
- matrix norm: $\|\boldsymbol{A}\|_{\boldsymbol{\pi}} := \sup_{\|\boldsymbol{x}\|_{\boldsymbol{\pi}}=1} \|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{\pi}}$ with $\boldsymbol{A} = [A_{i,j}]_{1 \le i,j \le n}$

Eigenvector perturbation for transition matrices

Theorem 2.7 (Chen, Fan, Ma, Wang '19)

Suppose $\boldsymbol{P}^\star$ represents a reversible Markov chain, whose stationary distribution vector $\boldsymbol{\pi}^\star$ is strictly positive. Assume

$$\|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star} < (1-1/\sqrt{2})\Big(1-\max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\}\Big)$$

Then one has

$$\|\boldsymbol{\pi}-\boldsymbol{\pi}^\star\|_{\boldsymbol{\pi}^\star} \le \frac{\sqrt{2}\,\|\boldsymbol{\pi}^{\star\top}\boldsymbol{E}\|_{\boldsymbol{\pi}^\star}}{1-\max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\}}$$

- similar to Davis-Kahan theorem
- eigengap: $1-\max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\}$ since $1=\lambda_1(\boldsymbol{P}^\star)$
- perturbation size: $\left\|\boldsymbol{\pi}^{\star\top}\boldsymbol{E}\right\|_{\boldsymbol{\pi}^\star}$

Part 3: Application of ℓ_2 perturbation theory

- Matrix tail bounds
- Community detection
- Matrix completion
- Ranking from pairwise comparisons

Matrix tail bounds

A hammer: matrix Bernstein inequality

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

- $\mathbb{E}[\boldsymbol{X}_l] = \boldsymbol{0}$
- $\|\boldsymbol{X}_l\| \le B$ 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 3.8 (Matrix Bernstein inequality)

For all $\tau \ge 0$,

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

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

$$\left\|\sum_l \boldsymbol{X}_l\right\| \lesssim \sqrt{v\log(d_1+d_2)} + B\log(d_1+d_2) \tag{3.9}$$

Another hammer: spectral norm of random matrices w/ independent entries

Consider a symmetric random matrix $\boldsymbol{X} = [X_{i,j}]_{1 \le i,j \le n}$ with independent entries s.t. $\forall (i,j)$:

- $\mathbb{E}[X_{i,j}] = \mathbf{0}$
- $\mathsf{Var}(X_{i,j}) \le \sigma^2$
- $|X_{i,j}| \le B$

Theorem 3.9 (Bandeira, van Handel '16)

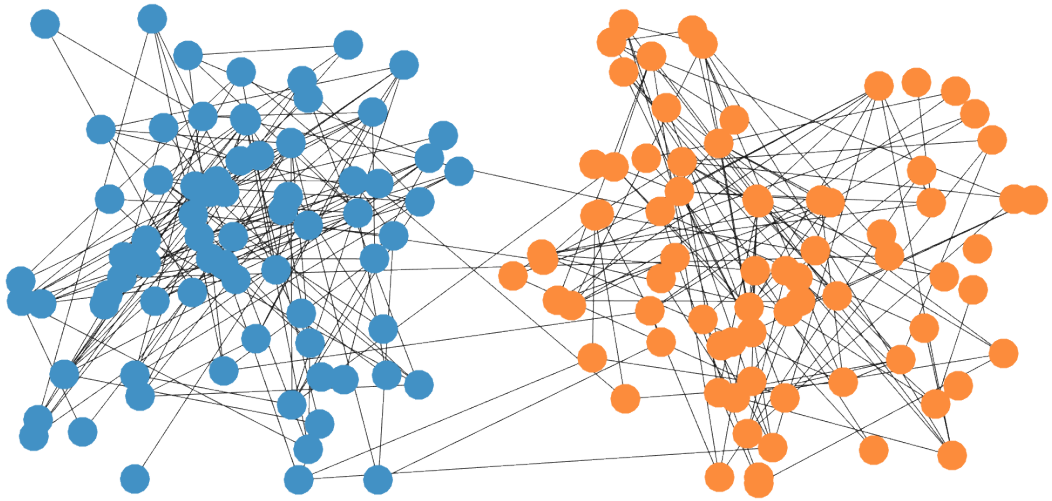
With probability exceeding $1 - O(n^{-10})$,

$$\|\boldsymbol{X}\| \le 4\sigma\sqrt{n} + O\big(B\sqrt{\log n}\big)$$

- often tighter than matrix Bernstein by some log factor

Community detection

Recap: spectral clustering for SBMs

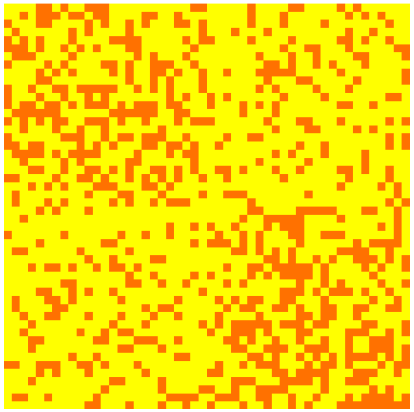
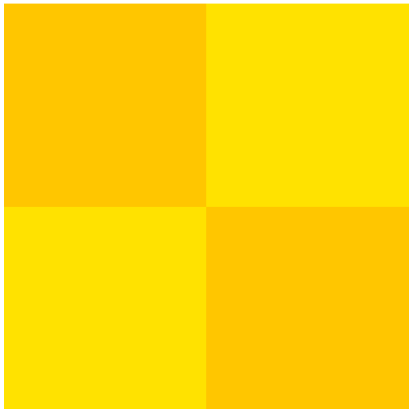
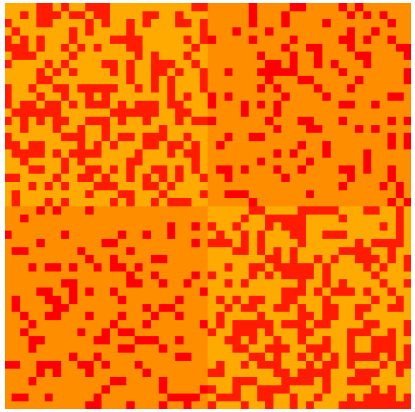


$x_i = 1$: 1^{st} community $\qquad$ $x_i = -1$: 2^{nd} community

- n nodes $\{1, \cdots, n\}$
- 2 communities
- community memberships to recover: $\boldsymbol{x} = [x_i]_{1 \le i \le n} \in \{1, -1\}^n$
- observed: an adjacency matrix $\boldsymbol{A}$ of a random graph s.t.

$$\mathbb{P}(A_{i,j} = 1) = \begin{cases} p, & \text{if } x_i = x_j \\ q, & \text{else} \end{cases}$$

Recap: spectral clustering for SBMs

$$\boldsymbol{A} = \underbrace{\mathbb{E}[\boldsymbol{A}]}_{= \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top + \frac{p-q}{2}\boldsymbol{x}\boldsymbol{x}^\top} + \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$$

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

Analysis via Davis-Kahan's theorem

Let us apply Davis-Kahan to analyze accuracy of spectral clustering:

- take $\boldsymbol{M}^\star = \underbrace{\mathbb{E}[\boldsymbol{A}] - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top}_{=\frac{p-q}{2}\boldsymbol{x}\boldsymbol{x}^\top}$
  - leading eigenvector (resp. value) $\boldsymbol{u}^\star = \frac{1}{\sqrt{n}}\boldsymbol{x}^\star$ (resp. $\lambda^\star = \frac{(p-q)n}{2}$)
- take $\boldsymbol{M} = \boldsymbol{A} - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top$ w/ leading eigenvector $\boldsymbol{u}$
- Theorem 2.4 yields

$$\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \le \frac{2\|\boldsymbol{M} - \boldsymbol{M}^\star\|}{\lambda_1^\star} = \frac{2\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|}{\frac{(p-q)n}{2}} \tag{3.10}$$

Analysis via Davis-Kahan's theorem

Let us apply Davis-Kahan to analyze accuracy of spectral clustering:

- take $\boldsymbol{M}^\star = \underbrace{\mathbb{E}[\boldsymbol{A}] - \frac{p+q}{2}\boldsymbol{1}\boldsymbol{1}^\top}_{=\frac{p-q}{2}\boldsymbol{x}\boldsymbol{x}^\top}$
  - leading eigenvector (resp. value) $\boldsymbol{u}^\star = \frac{1}{\sqrt{n}}\boldsymbol{x}^\star$ (resp. $\lambda^\star = \frac{(p-q)n}{2}$)
- take $\boldsymbol{M} = \boldsymbol{A} - \frac{p+q}{2}\boldsymbol{1}\boldsymbol{1}^\top$ w/ leading eigenvector $\boldsymbol{u}$
- Theorem 2.4 yields

$$\mathrm{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \le \frac{2\|\boldsymbol{M} - \boldsymbol{M}^\star\|}{\lambda_1^\star} = \frac{2\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|}{\frac{(p-q)n}{2}} \tag{3.10}$$

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

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

Lemma 3.10

Consider SBM with $p > q \gtrsim \frac{\log n}{n}$. *Then with prob.* $1 - O(n^{-10})$,

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| \lesssim \sqrt{np} \tag{3.11}$$

proof: note that

- $\mathrm{Var}(A_{i,j}) = \begin{cases} p(1-p) & \text{if } x_i = x_j \\ q(1-q) & \text{else} \end{cases} \leq p =: \sigma^2$
- $A_{i,j} \leq 1 =: B$

Applying Theorem 3.9 and using $p \gtrsim \frac{\log n}{n}$ conclude the proof

Statistical accuracy of spectral clustering

Substitute (3.11) into (3.10) to reach

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

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

Thus, under condition $\frac{p-q}{\sqrt{p}} \gg \frac{1}{\sqrt{n}}$, with high prob. one has

$$\mathrm{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \ll 1 \qquad \Longrightarrow \qquad \text{almost exact clustering}$$

Statistical accuracy of spectral clustering

$$\frac{p-q}{\sqrt{p}} \gg \frac{1}{\sqrt{n}} \quad \Longrightarrow \quad \text{almost exact clustering} \tag{3.12}$$

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

$$p - q \gg \frac{1}{\sqrt{n}}$$

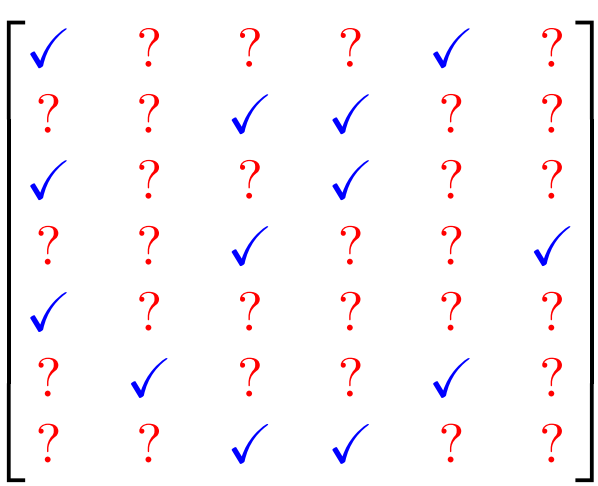
- **"sparse" regime:** if $p = \frac{\alpha \log n}{n}$ and $q = \frac{\beta \log n}{n}$ for $\alpha, \beta \asymp 1$, then

$$\alpha - \beta \gg \frac{1}{\sqrt{\log n}}$$

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

Matrix completion

Recap: spectral method for matrix completion



- ground truth: $\boldsymbol{M}^\star = \boldsymbol{u}^\star \boldsymbol{v}^{\star\top} \in \mathbb{R}^{n\times n}$ (a simple case)

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

- each entry $M^\star_{i,j}$ is observed independently with prob. p
- **goal:** fill in unseen entries of $\boldsymbol{M}^\star$

Recap: spectral method for matrix completion

1. construct a rescaled zero-filled matrix $\boldsymbol{M} = [M_{i,j}] \in \mathbb{R}^{n_1 \times n_2}$ as

$$\forall (i,j): \quad M_{i,j} = \begin{cases} \frac{1}{p} M^{\star}_{i,j}, & \text{if } M^{\star}_{i,j} \text{ is observed} \\ 0, & \text{else} \end{cases}$$

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

2. compute rank-1 SVD $\sigma \boldsymbol{u}\boldsymbol{v}^{\top}$ of $\boldsymbol{M}$, and return $\widehat{\boldsymbol{M}} = \sigma \boldsymbol{u}\boldsymbol{v}^{\top}$

How does sampling rate p affect estimation accuracy?

Statistical accuracy of spectral estimate

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

$$\max\{\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star), \mathsf{dist}(\boldsymbol{v}, \boldsymbol{v}^\star)\} \leq \frac{2\|\boldsymbol{M} - \boldsymbol{M}^\star\|}{\sigma_1^\star} \asymp \|\boldsymbol{M} - \boldsymbol{M}^\star\|$$
$$\lesssim \sqrt{\frac{\log^2 n}{np}} \tag{3.13}$$

where last inequality is a consequence of:

Lemma 3.11

Suppose $p \gg \frac{\log n}{n}$. Then with high prob.,

$$\|\boldsymbol{M} - \boldsymbol{M}^\star\| \lesssim \sqrt{\frac{\log^2 n}{np}} = o(1) \tag{3.14}$$

Sample complexity

For rank-1 matrix completion, (3.13) implies

$$p \gg \frac{\log^2 n}{n} \quad \Longrightarrow \quad \text{nearly accurate estimates of } \boldsymbol{u}^\star \ \& \ \boldsymbol{v}^\star$$

$$\Longrightarrow \quad \text{nearly accurate estimates of } \boldsymbol{M}^\star$$

To yield reliable spectral estimates, it suffices to have sample size

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

Proof of inequality (3.14)

- First, based on Gaussianity, we have

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

- Next,

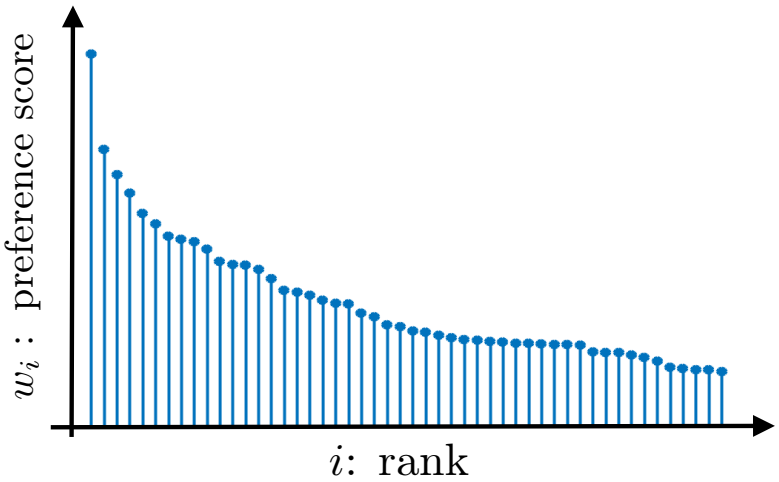
$$\max_{i,j}\mathsf{Var}(M_{i,j}) = \frac{1-p}{p}\max_{i,j}(M^\star_{i,j})^2 \lesssim \frac{\log^2 n}{n^2 p} =: \sigma^2$$

Applying Theorem 3.9 w/ dilation trick $\|\boldsymbol{A}\| = \left\|\begin{bmatrix} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \end{bmatrix}\right\|$ gives

$$\|\boldsymbol{M}-\boldsymbol{M}^\star\| \lesssim \sigma\sqrt{n} + B\sqrt{\log n} \asymp \frac{\log n}{\sqrt{np}} + \frac{\log^{3/2} n}{np} \asymp \frac{\log n}{\sqrt{np}}$$

Ranking from pairwise comparisons

Recap: spectral ranking for BTL model



- n items with latent scores $w_1^\star, \ldots, w_n^\star$
- each pair of items (i, j) is compared independently

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

- estimate $\boldsymbol{w}^\star = [w_i^\star]_{1 \le i \le n}$ (and rank items)

Recap: spectral ranking for BTL model

A key probability transition matrix $\boldsymbol{P}^\star \in \mathbb{R}^{n \times n}$:

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

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

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

2. compute leading left eigenvector $\boldsymbol{\pi}$ of $\boldsymbol{P}$ as score estimate
3. rank in accordance with $\boldsymbol{\pi}$

Can we characterize the accuracy of spectral estimates?

Analysis of spectral ranking

Apply Theorem 2.7 to yield

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

with $\boldsymbol{E} = \boldsymbol{P} - \boldsymbol{P}^\star$, provided that

$$\|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star} \le (1 - 1/\sqrt{2})\Big(1 - \max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\}\Big)$$

Analysis of spectral ranking

Apply Theorem 2.7 to yield

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

with $\boldsymbol{E} = \boldsymbol{P} - \boldsymbol{P}^\star$, provided that

$$\|\boldsymbol{E}\|_{\boldsymbol{\pi}^\star} \le (1 - 1/\sqrt{2})\Big(1 - \max\left\{\lambda_2(\boldsymbol{P}^\star), -\lambda_n(\boldsymbol{P}^\star)\right\}\Big)$$

— need to understand spectral gap and noise size

Analysis of spectral ranking (cont.)

$$\text{condition number:} \quad \kappa := \frac{\max_{1\le i\le n} w_i^\star}{\min_{1\le i\le n} w_i^\star}$$

Lemma 3.12 (spectral gap)

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

- proof is based on comparison between two reversible Markov chains; see Section 3.6.4 of Chen et al. '21

Lemma 3.13 (noise size)

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

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

Analysis of spectral ranking (cont.)

Recall perturbation bound

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

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

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

As a result, one has

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

Analysis of spectral ranking (cont.)

Assuming $\kappa = O(1)$, we arrive at

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

- vanishingly small error as $n \to \infty$
- optimal error up to log factor

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

Proof of Lemma 3.13

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

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

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

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

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

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

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

Proof of Lemma 3.13 (controlling $\|\boldsymbol{E}_{\mathsf{diag}}\|$)

Observe that

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

To invoke Bernstein's inequality, note that

- $|X_j| \le 1 =: B$
- $\sum_{j:j\neq i} \mathbb{E}[X_j^2] = \sum_{j:j\neq i} \mathsf{Var}(y_{i,j}) \le n =: v$

Bernstein's inequality + union bound reveal that: with high prob.

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

Proof of Lemma 3.13 (controlling $\|\boldsymbol{E}_{\text{upper}}\|$)

- $\frac{1}{n}|\boldsymbol{y}_{i,j}| \le \frac{1}{n} =: B$
- $\mathsf{Var}\big(\frac{1}{n} y_{i,j}\big) \le \frac{1}{n^2} =: \sigma^2$

Applying Theorem 3.9 w/ dilation trick $\|\boldsymbol{A}\| = \left\| \begin{bmatrix} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \end{bmatrix} \right\|$ gives

$$\|\boldsymbol{E}_{\text{upper}}\| \lesssim \sigma\sqrt{n} + B\sqrt{\log n} \asymp \frac{1}{\sqrt{n}} + \frac{\sqrt{\log n}}{n} \asymp \frac{1}{\sqrt{n}}$$

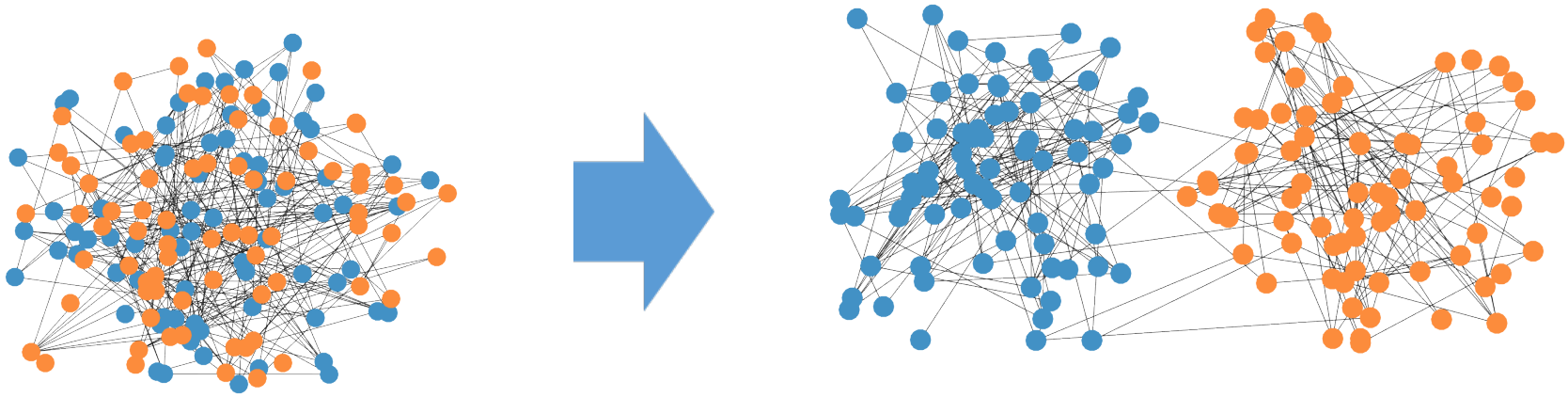
— same bound holds for $\|\boldsymbol{E}_{\text{lower}}\|$

Part 4: ℓ_∞ and $\ell_{2,\infty}$ perturbation theory

- Motivation
  - exact community recovery
  - top-K ranking
- Leave-one-out analysis: an illustrative example
- ℓ_∞ eigenvector perturbation theory (rank-1)
- Application: exact recovery in community detection
- $\ell_{2,\infty}$ eigen-space perturbation theory (rank-r)

Motivation: exact community recovery

Revisiting spectral clustering for SBMs



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

Revisiting spectral clustering for SBMs

It has been shown in (3.12) that: if $p \asymp q \asymp \frac{\log n}{n}$, then

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

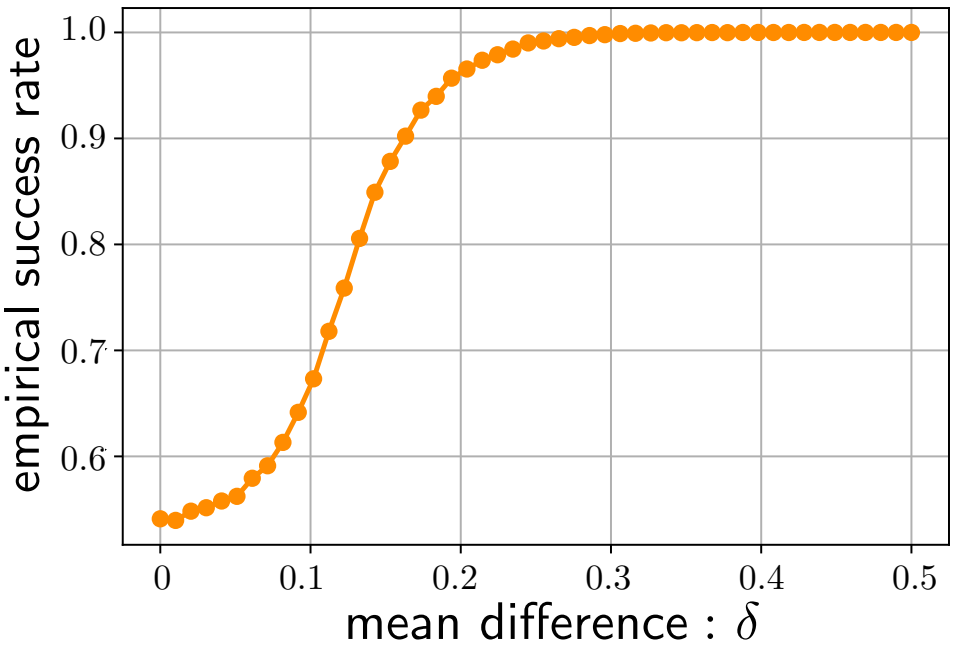
- Almost exact recovery means

$$\min \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \neq x_i^\star\}, \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \neq -x_i^\star\} \right\} = o(1)$$

Exact recovery of all community memberships?

When $\delta := p - q$ increases, exact recovery becomes possible:

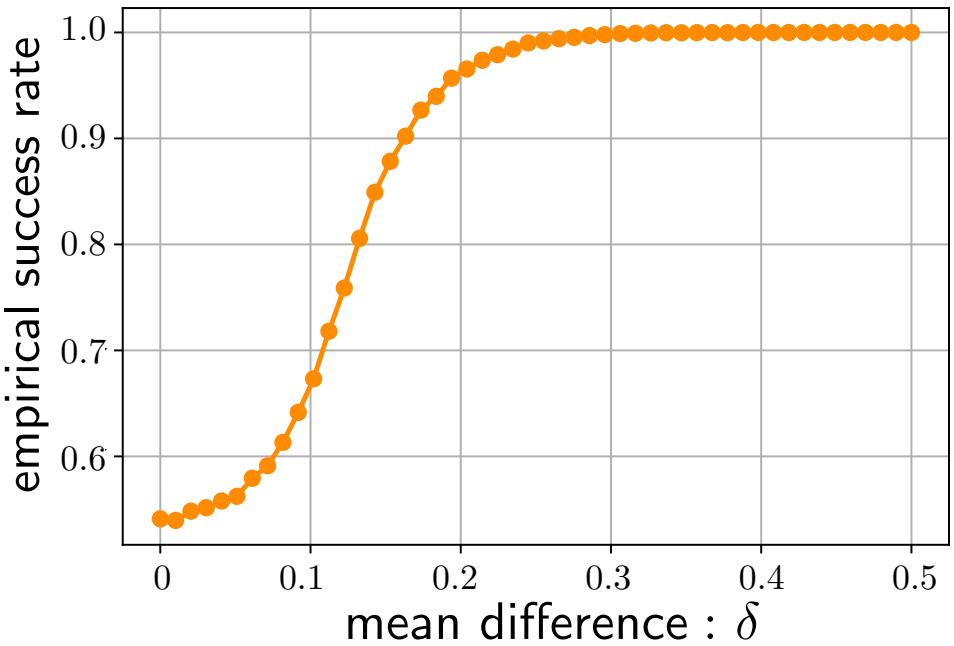
$$\min\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i \neq x_i^\star\}, \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i \neq -x_i^\star\}\right\} = 0$$



Exact recovery of all community memberships?

When $\delta := p - q$ increases, exact recovery becomes possible:

$$\min\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i \neq x_i^\star\}, \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i \neq -x_i^\star\}\right\} = 0$$



ℓ_2 perturbation theory alone falls short of explaining exact recovery

— calls for more fine-grained analysis

Exact recovery $\leftarrow$ ℓ_∞ theory

exact recovery means $u_i u_i^\star > 0$, $\forall i$ (or $u_i u_i^\star < 0$, $\forall i$)

Exact recovery $\leftarrow$ ℓ_∞ theory

exact recovery means $u_i u_i^\star > 0$, $\forall i$ (or $u_i u_i^\star < 0$, $\forall i$)

$\|\boldsymbol{u} - \boldsymbol{u}^\star\|_\infty < 1/\sqrt{n}$ or $\|\boldsymbol{u} + \boldsymbol{u}^\star\|_\infty < 1/\sqrt{n}$

Exact recovery $\leftarrow$ ℓ_∞ theory

exact recovery means $u_i u_i^\star > 0$, $\forall i$ (or $u_i u_i^\star < 0$, $\forall i$)

$\|\boldsymbol{u} - \boldsymbol{u}^\star\|_\infty < 1/\sqrt{n}$ or $\|\boldsymbol{u} + \boldsymbol{u}^\star\|_\infty < 1/\sqrt{n}$

ℓ_∞ eigenvector perturbation theory

Motivation: top-K ranking

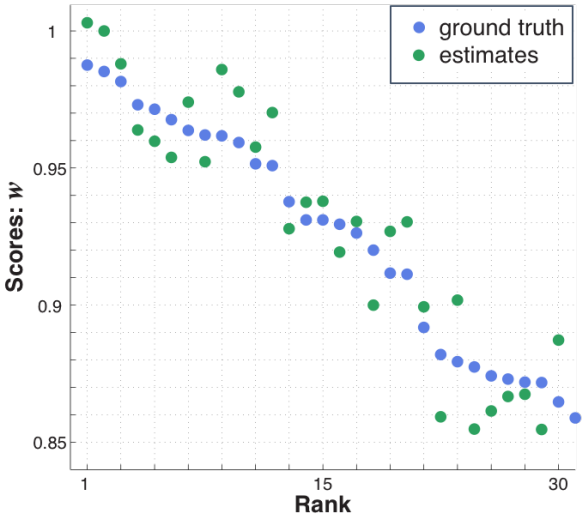
Top-K ranking

Goal: identify the set of top-K ranked items

Typical ranking procedure:

- estimate latent scores
- return top-K items in accordance with score estimates

Top-K ranking for BTL model



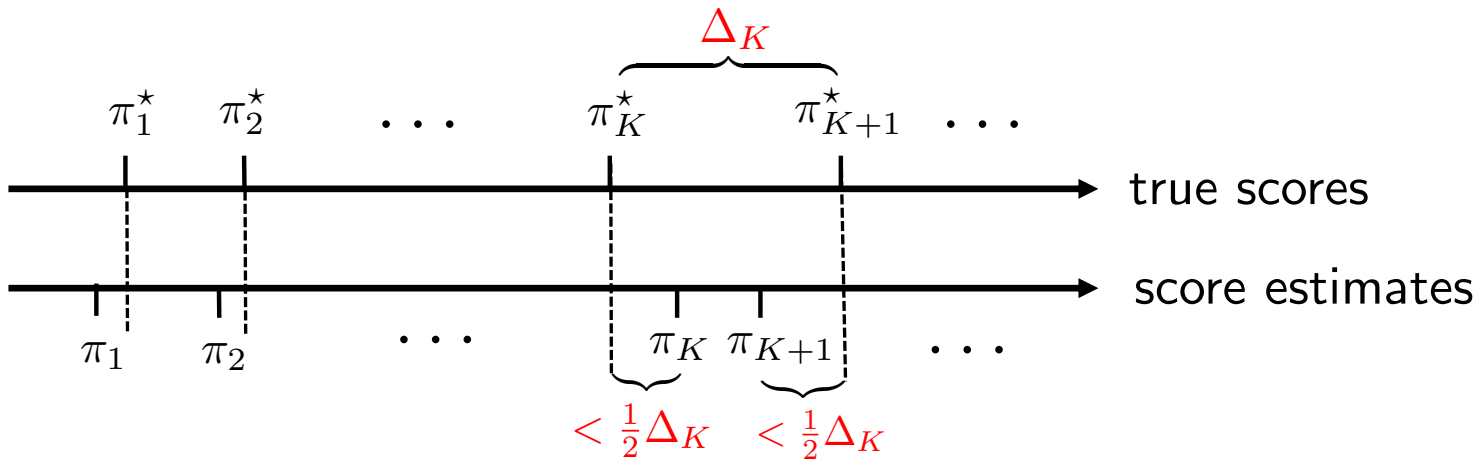
Top 3: $\{1, 2, 3\}$

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

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

2. compute leading left eigenvector $\boldsymbol{\pi}$ of $\boldsymbol{P}$ as score estimate
3. return K items associated with largest score estimate π_i

Controlling entrywise estimation error



exact top-K ranking

$$2\|\boldsymbol{\pi} - \boldsymbol{\pi}^\star\|_\infty < \Delta_K := \pi^\star_{(K)} - \pi^\star_{(K+1)}$$

ℓ_∞ eigenvector perturbation theory

Leave-one-out analysis: an illustrative example

from random matrix theory, stat. physics, etc

Setup and algorithm

- Ground truth: $\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top} \in \mathbb{R}^{n\times n}$, with $\lambda^\star > 0$
  - $\|\boldsymbol{u}^\star\|_2 = 1$, $\|\boldsymbol{u}^\star\|_\infty = \sqrt{\mu/n}$ (μ: incoherence parameter)
- Observation: $\boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{E}$
  - $\boldsymbol{E}$: symmetric, entries in upper triangular part are i.i.d. $\mathcal{N}(0, \sigma^2)$
- Estimate $\boldsymbol{u}^\star$ using leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$

Question: can we characterize entrywise estimation error of $\boldsymbol{u}$, i.e.

$$\mathsf{dist}_\infty(\boldsymbol{u}, \boldsymbol{u}^\star) \coloneqq \min\left\{\|\boldsymbol{u} - \boldsymbol{u}^\star\|_\infty, \|\boldsymbol{u} + \boldsymbol{u}^\star\|_\infty\right\}$$

ℓ_2 guarantees

Davis-Kahan's $\sin\boldsymbol{\Theta}$ theorem together with Theorem 3.9 gives

$$\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \leq \frac{2\|\boldsymbol{E}\|}{\lambda^\star} \leq \frac{10\sigma\sqrt{n}}{\lambda^\star}$$

with high prob., as long as $\sigma\sqrt{n} \leq \frac{1-1/\sqrt{2}}{5}\lambda^\star$

- as an immediate (but very crude) consequence

$$\mathsf{dist}_\infty(\boldsymbol{u}, \boldsymbol{u}^\star) \leq \mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \lesssim \frac{\sigma\sqrt{n}}{\lambda^\star} \tag{4.15}$$

ℓ_∞ guarantees for matrix denoising

Theorem 4.14

Suppose that $\sigma\sqrt{n} \le c_0\lambda^\star$ for some sufficiently small constant $c_0 > 0$. Then with high prob.,

$$\mathsf{dist}_\infty(\boldsymbol{u}, \boldsymbol{u}^\star) \lesssim \frac{\sigma(\sqrt{\log n} + \sqrt{\mu})}{\lambda^\star}$$

- When $\mu \lesssim \log n$ (i.e. energy of $\boldsymbol{u}^\star$ is spread out):

$$\mathsf{dist}_\infty(\boldsymbol{u}, \boldsymbol{u}^\star) \lesssim \frac{\sigma\sqrt{\log n}}{\lambda^\star}$$

- Much sharper (i.e. $\sqrt{n/\log n}$ times better) than (4.15)

Technical hurdle: statistical dependency

Let's take close inspection of l-th entry u_l of $\boldsymbol{u}$:

- Given that $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{M}$, we have

$$\boldsymbol{M}\boldsymbol{u} = \lambda \boldsymbol{u},$$

$$\Longrightarrow \quad u_l = \frac{1}{\lambda}[\boldsymbol{M}]_{l,:}\boldsymbol{u} = \frac{1}{\lambda}[\boldsymbol{M}^\star + \boldsymbol{E}]_{l,:}\boldsymbol{u}$$

- **challenge:** $\boldsymbol{u}$ is statistically dependent on $\boldsymbol{E}$ (in a complicated way)!

How to decouple complicated dependency between $\boldsymbol{u}$ and $\boldsymbol{E}$?

Decomposition w/ the aid of an independent proxy

Suppose we have access to a proxy $\boldsymbol{u}^{(l)}$ independent of $\boldsymbol{E}_{l,:}$, then

$$\underbrace{\boldsymbol{E}_{l,:}\boldsymbol{u}}_{\text{a term of interest}} = \underbrace{\boldsymbol{E}_{l,:}\boldsymbol{u}^{(l)}}_{=:\mathcal{J}_1} + \underbrace{\boldsymbol{E}_{l,:}\big(\boldsymbol{u}-\boldsymbol{u}^{(l)}\big)}_{=:\mathcal{J}_2}$$

- $\mathcal{J}_1$: controllable using independence btw $\boldsymbol{u}^{(l)}$ & $\boldsymbol{E}_{l,:}$
- $\mathcal{J}_2$: small if $\boldsymbol{u}^{(l)} \approx \boldsymbol{u}$

How to construct a useful proxy?

Leave-one-out auxiliary estimates

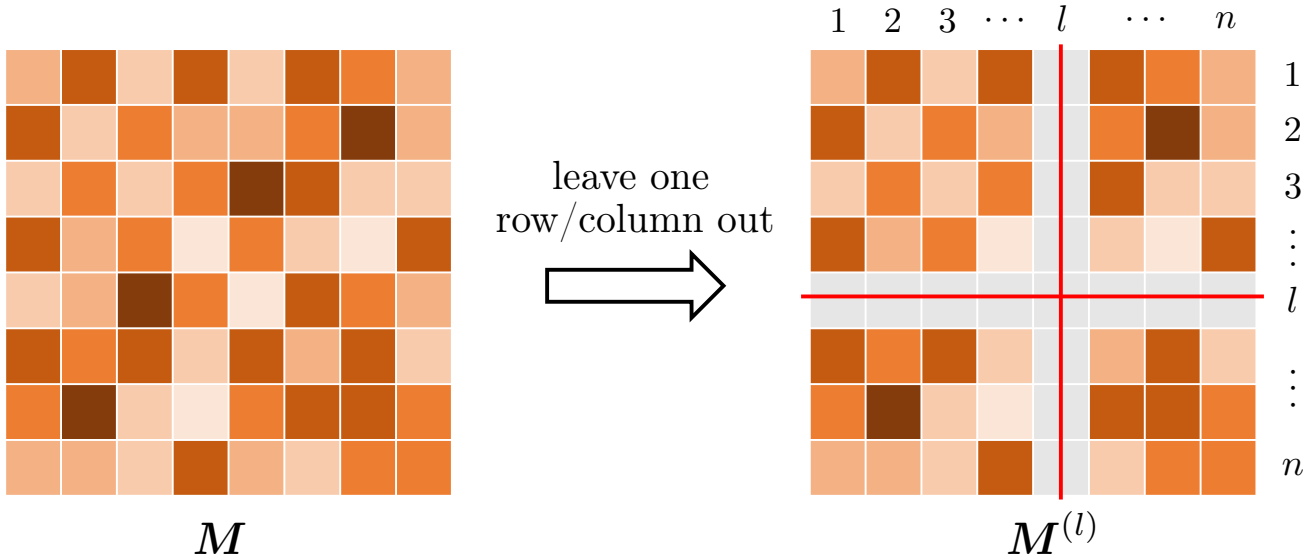
For each $1 \leq l \leq n$, construct an auxiliary matrix $\boldsymbol{M}^{(l)}$

$$\boldsymbol{M}^{(l)} := \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top} + \boldsymbol{E}^{(l)},$$

where the noise matrix $\boldsymbol{E}^{(l)}$ is generated according to

$$E_{i,j}^{(l)} := \begin{cases} E_{i,j}, & \text{if } i \neq l \text{ and } j \neq l \\ 0, & \text{else} \end{cases}$$

(removing l-th row/col)



Leave-one-out auxiliary estimates

For each $1 \leq l \leq n$, construct an auxiliary matrix $\boldsymbol{M}^{(l)}$

$$\boldsymbol{M}^{(l)} := \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top} + \boldsymbol{E}^{(l)},$$

where the noise matrix $\boldsymbol{E}^{(l)}$ is generated according to

$$E_{i,j}^{(l)} := \begin{cases} E_{i,j}, & \text{if } i \neq l \text{ and } j \neq l \\ 0, & \text{else} \end{cases} \qquad \text{(removing } l\text{-th row/col)}$$

leave-one-out estimates: $\boldsymbol{u}^{(l)}$ (resp. $\lambda^{(l)}$) is leading eigenvalue (resp. eigenvector) of $\boldsymbol{M}^{(l)}$

- **key property:** $\boldsymbol{u}^{(l)}$ is independent of $\boldsymbol{E}_{l,:}$

Intuition

WLOG, suppose $\boldsymbol{u}^\top \boldsymbol{u}^\star > 0$ and $\boldsymbol{u}^{(l)\top} \boldsymbol{u}^\star > 0$ ...

- **proximity of $\boldsymbol{u}^{(l)}$ and $\boldsymbol{u}$:** since $\boldsymbol{u}^{(l)}$ is obtained by dropping only a tiny fraction of data, we expect $\boldsymbol{u} \approx \boldsymbol{u}^{(l)}$
- **proximity of $u_l^{(l)}$ and $u_l^\star$:** by construction,

$$\begin{aligned} u_l^{(l)} &= \frac{1}{\lambda^{(l)}} \boldsymbol{M}_{l,\cdot}^{(l)} \boldsymbol{u}^{(l)} = \frac{1}{\lambda^{(l)}} \boldsymbol{M}_{l,\cdot}^{\star} \boldsymbol{u}^{(l)} = \frac{\lambda^\star}{\lambda^{(l)}} u_l^\star \boldsymbol{u}^{\star\top} \boldsymbol{u}^{(l)} \\ &\approx u_l^\star \end{aligned}$$

Proof of Theorem 4.14

What we have learned from ℓ_2 analysis

$$\|\boldsymbol{E}\| \le 5\sigma\sqrt{n} \qquad \|\boldsymbol{E}^{(l)}\| \le \|\boldsymbol{E}\| \le 5\sigma\sqrt{n}$$

$$\mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \le \frac{10\sigma\sqrt{n}}{\lambda^\star} \qquad \mathsf{dist}(\boldsymbol{u}^{(l)}, \boldsymbol{u}^\star) \le \frac{10\sigma\sqrt{n}}{\lambda^\star}$$

$$|\lambda - \lambda^\star| \le 5\sigma\sqrt{n} \qquad |\lambda^{(l)} - \lambda^\star| \le 5\sigma\sqrt{n}$$

$$\max_{j:j\ge 2} |\lambda_j(\boldsymbol{M})| \le 5\sigma\sqrt{n} \qquad \max_{j:j\ge 2} |\lambda_j(\boldsymbol{M}^{(l)})| \le 5\sigma\sqrt{n}$$

Addressing ambiguity

WLOG, assume

$$\|\boldsymbol{u} - \boldsymbol{u}^\star\|_2 = \mathsf{dist}(\boldsymbol{u}, \boldsymbol{u}^\star),$$
$$\left\|\boldsymbol{u}^{(l)} - \boldsymbol{u}^\star\right\|_2 = \mathsf{dist}(\boldsymbol{u}^{(l)}, \boldsymbol{u}^\star), \quad 1 \le l \le n$$

A useful byproduct: if $c_0\sigma\sqrt{n} < \lambda^\star$ for some small constant $c_0 > 0$, then one necessarily has (exercise)

$$\left\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\right\|_2 = \mathsf{dist}\big(\boldsymbol{u}, \boldsymbol{u}^{(l)}\big), \qquad 1 \le l \le n$$

Key steps

To bound $u_l - u_l^\star$ $(1 \leq l \leq n)$, we see from triangle inequality that

$$|u_l - u_l^\star| \leq |u_l^{(l)} - u_l^\star| + \|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_\infty \leq |u_l^{(l)} - u_l^\star| + \|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_2$$

Key steps

To bound $u_l - u_l^\star$ $(1 \le l \le n)$, we see from triangle inequality that

$$|u_l - u_l^\star| \le |u_l^{(l)} - u_l^\star| + \|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_\infty \le |u_l^{(l)} - u_l^\star| + \|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_2$$

- control $\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_2$ (Davis-Kahan)
- control $|u_l^{(l)} - u_l^\star|$ ($u_l^{(l)}$ is independent from $\boldsymbol{E}_{l,:}$)

Bounding proximity $\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_2$

Key: view $\boldsymbol{M}$ as perturbation of $\boldsymbol{M}^{(l)}$, apply Davis-Kahan

$$\left\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\right\|_2 \leq \frac{2\|\left(\boldsymbol{M} - \boldsymbol{M}^{(l)}\right)\boldsymbol{u}^{(l)}\|_2}{\lambda^{(l)} - \max\limits_{j\geq 2}\left|\lambda_j\left(\boldsymbol{M}^{(l)}\right)\right|} \leq \frac{4\|\left(\boldsymbol{M} - \boldsymbol{M}^{(l)}\right)\boldsymbol{u}^{(l)}\|_2}{\lambda^\star}$$

as long as

$$\|\boldsymbol{M} - \boldsymbol{M}^{(l)}\| \leq (1 - 1/\sqrt{2})\Big(\lambda^{(l)} - \max_{j\geq 2}\left|\lambda_j\left(\boldsymbol{M}^{(l)}\right)\right|\Big),$$

$$\lambda^{(l)} - \max_{j\geq 2}\left|\lambda_j\left(\boldsymbol{M}^{(l)}\right)\right| \geq \lambda^\star/2$$

Bounding $\|(\boldsymbol{M} - \boldsymbol{M}^{(l)})\boldsymbol{u}^{(l)}\|_2$

By design,

$$(\boldsymbol{M} - \boldsymbol{M}^{(l)})\boldsymbol{u}^{(l)} = \boldsymbol{e}_l \boldsymbol{E}_{l,\cdot}\boldsymbol{u}^{(l)} + u_l^{(l)}(\boldsymbol{E}_{\cdot,l} - E_{l,l}\boldsymbol{e}_l),$$

which together with triangle inequality yields

$$\begin{aligned}\|(\boldsymbol{M} - \boldsymbol{M}^{(l)})\boldsymbol{u}^{(l)}\|_2 &\leq \underbrace{|\boldsymbol{E}_{l,\cdot}\boldsymbol{u}^{(l)}|}_{\boldsymbol{E}_{l,\cdot} \text{ and } \boldsymbol{u}^{(l)} \text{ are independent}} + \|\boldsymbol{E}_{\cdot,l}\|_2 \cdot |u_l^{(l)}| \\ &\leq 5\sigma\sqrt{\log n} + \|\boldsymbol{E}_{\cdot,l}\|_2(|u_l| + \|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_\infty) \\ &\leq 5\sigma\sqrt{\log n} + 5\sigma\sqrt{n}\|\boldsymbol{u}\|_\infty + 5\sigma\sqrt{n}\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_2\end{aligned}$$

Bounding $\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\|_2$ (cont.)

Combining previous bounds, we arrive at

$$\begin{aligned}\left\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\right\|_2 &\leq \frac{20\sigma\sqrt{\log n} + 20\sigma\sqrt{n}\|\boldsymbol{u}\|_\infty + 20\sigma\sqrt{n}\left\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\right\|_2}{\lambda^\star} \\ &\leq \frac{20\sigma\sqrt{\log n} + 20\sigma\sqrt{n}\|\boldsymbol{u}\|_\infty}{\lambda^\star} + \frac{1}{2}\left\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\right\|_2\end{aligned}$$

provided that $40\sigma\sqrt{n} \leq \lambda^\star$

Rearranging terms and taking union bound give: with high prob.

$$\left\|\boldsymbol{u} - \boldsymbol{u}^{(l)}\right\|_2 \leq \frac{40\sigma\sqrt{\log n} + 40\sigma\sqrt{n}\|\boldsymbol{u}\|_\infty}{\lambda^\star} \qquad 1 \leq l \leq n$$

Analyzing leave-one-out iterates

Recall that

$$u_l^{(l)} = \frac{1}{\lambda^{(l)}} \boldsymbol{M}_{l,\cdot}^{(l)} \boldsymbol{u}^{(l)} = \frac{1}{\lambda^{(l)}} \boldsymbol{M}_{l,\cdot}^{\star} \boldsymbol{u}^{(l)} = \frac{\lambda^\star}{\lambda^{(l)}} u_l^\star \boldsymbol{u}^{\star\top} \boldsymbol{u}^{(l)}$$

This implies

$$\begin{aligned} u_l^{(l)} - u_l^\star &= u_l^\star \Big(\frac{\lambda^\star}{\lambda^{(l)}} \boldsymbol{u}^{\star\top} \boldsymbol{u}^{(l)} - \boldsymbol{u}^{\star\top} \boldsymbol{u}^\star \Big) \\ &= u_l^\star \Big(\frac{\lambda^\star - \lambda^{(l)}}{\lambda^{(l)}} \boldsymbol{u}^{\star\top} \boldsymbol{u}^{(l)} \Big) + u_l^\star \boldsymbol{u}^{\star\top} \big(\boldsymbol{u}^{(l)} - \boldsymbol{u}^\star \big) \end{aligned}$$

Analyzing leave-one-out iterates (cont.)

Triangle inequality gives

$$\begin{aligned}
\left|u_l^{(l)} - u_l^\star\right| &\le \left|u_l^\star\right| \cdot \frac{\left|\lambda^\star - \lambda^{(l)}\right|}{\left|\lambda^{(l)}\right|} \cdot \|\boldsymbol{u}^\star\|_2 \cdot \|\boldsymbol{u}^{(l)}\|_2 \\
&\quad + \left|u_l^\star\right| \cdot \|\boldsymbol{u}^\star\|_2 \cdot \left\|\boldsymbol{u}^{(l)} - \boldsymbol{u}^\star\right\|_2 \\
&\le \left|u_l^\star\right| \cdot \frac{10\sigma\sqrt{n}}{\lambda^\star} + \left|u_l^\star\right| \cdot \frac{10\sigma\sqrt{n}}{\lambda^\star} \\
&\le \frac{20\sigma\sqrt{n}}{\lambda^\star}\left\|\boldsymbol{u}^\star\right\|_\infty
\end{aligned}$$

Putting all pieces together

Now we come to conclude that

$$\begin{aligned}\|\boldsymbol{u}-\boldsymbol{u}^\star\|_\infty &= \max_l |u_l - u_l^\star| \le \max_l \left\{ |u_l^{(l)} - u_l^\star| + \|\boldsymbol{u}-\boldsymbol{u}^{(l)}\|_2 \right\} \\ &\le \frac{20\sigma\sqrt{n}}{\lambda^\star}\|\boldsymbol{u}^\star\|_\infty + \frac{40\sigma\sqrt{\log n}+40\sigma\sqrt{n}\|\boldsymbol{u}\|_\infty}{\lambda^\star}\end{aligned}$$

One more triangle inequality gives

$$\|\boldsymbol{u}-\boldsymbol{u}^\star\|_\infty \le \frac{40\sigma\sqrt{\log n}+60\sigma\sqrt{n}\,\|\boldsymbol{u}^\star\|_\infty}{\lambda^\star} + \frac{1}{2}\|\boldsymbol{u}-\boldsymbol{u}^\star\|_\infty$$

provided that $80\sigma\sqrt{n} \le \lambda^\star$. Rearranging terms yields

$$\|\boldsymbol{u}-\boldsymbol{u}^\star\|_\infty \le \frac{80\sigma\sqrt{\log n}+120\sigma\sqrt{n}\,\|\boldsymbol{u}^\star\|_\infty}{\lambda^\star} \underbrace{= \frac{80\sigma\sqrt{\log n}+120\sigma\sqrt{\mu}}{\lambda^\star}}_{\text{from definition of } \mu}$$

ℓ_∞ *eigenvector perturbation theory (rank-1)*

Setup and algorithm

- Ground truth: $\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top} \in \mathbb{R}^{n\times n}$, with $\lambda^\star > 0$
  - $\|\boldsymbol{u}^\star\|_2 = 1$, $\|\boldsymbol{u}^\star\|_\infty = \sqrt{\mu/n}$ (μ: incoherence parameter)
- Observation: $\boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{E}$ with symmetric $\boldsymbol{E}$
- Estimate $\boldsymbol{u}^\star$ using leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$

Question: can we accommodate more general noise distributions beyond Gaussian?

Noise assumptions

Entries in lower triangular part of $\boldsymbol{E} = [E_{i,j}]_{1 \le i,j \le n}$ are independently generated obeying

$$\mathbb{E}[E_{i,j}] = 0, \quad \mathbb{E}[E_{i,j}^2] \le \sigma^2, \quad |E_{i,j}| \le B, \quad \text{for all } i \ge j$$

Further, assume that

$$c_{\mathsf{b}} \coloneqq \frac{B}{\sigma \sqrt{n/(\mu \log n)}} = O(1)$$

- in general, B is allowed to be significantly larger than σ

ℓ_∞ perturbation theory

Theorem 4.15

With high prob, there exists $z \in \{1, -1\}$ *such that*

$$\|z\boldsymbol{u} - \boldsymbol{u}^\star\|_\infty \lesssim \frac{\sigma\sqrt{\mu} + \sigma\sqrt{\log n}}{\lambda^\star} \tag{4.18a}$$

$$\left\|z\boldsymbol{u} - \frac{1}{\lambda^\star}\boldsymbol{M}\boldsymbol{u}^\star\right\|_\infty \lesssim \frac{\sigma\sqrt{\mu}}{\lambda^\star} + \frac{\sigma^2\sqrt{n\log n} + \sigma B\sqrt{\mu\log^3 n}}{(\lambda^\star)^2} \tag{4.18b}$$

provided $\sigma\sqrt{n\log n} \le c_\sigma\lambda^\star$ *for some small enough constant* $c_\sigma > 0$

- estimation error is delocalized (recall that $\mathrm{dist}(\boldsymbol{u}, \boldsymbol{u}^\star) \lesssim \sigma\sqrt{n}$)

First-order expansion

Theorem 4.15 reveals tightness of first-order approximation

$$\boldsymbol{u} = \frac{\boldsymbol{M}\boldsymbol{u}}{\lambda} \approx \frac{\boldsymbol{M}\boldsymbol{u}^\star}{\lambda^\star} \approx \frac{\boldsymbol{M}^\star\boldsymbol{u}^\star}{\lambda^\star} = \boldsymbol{u}^\star$$

- (4.18b) often leads to tighter approximation than (4.18a)
- important in certain applications such as SBM

Application: exact recovery in community detection

Exact recovery using spectral methods

Consider the case where

$$p = \frac{\alpha \log n}{n}, \qquad q = \frac{\beta \log n}{n}$$

Theorem 4.16

Fix any constant $\varepsilon > 0$. Suppose $\alpha > \beta > 0$ are large enough[a] and

$$\left(\sqrt{\alpha} - \sqrt{\beta}\right)^2 \geq 2\left(1 + \varepsilon\right)$$

With probability $1 - o(1)$, spectral method achieves exact recovery

[a] this assumption can be removed

Optimality of spectral method

Lower bound: if

$$\left(\sqrt{\alpha} - \sqrt{\beta}\right)^2 \leq 2\left(1 - \varepsilon\right) \tag{4.19}$$

for any constant $\varepsilon > 0$, then no method can achieve exact recovery

- taking this w/ Theorem 4.16 reveals information-theoretic optimality of spectral method

What is the operational meaning of $\left(\sqrt{\alpha} - \sqrt{\beta}\right)^2$ or $\left(\sqrt{p} - \sqrt{q}\right)^2$?

Squared Hellinger distance

Definition 4.17

Consider two distributions P and Q over a finite alphabet $\mathcal{Y}$. The squared Hellinger distance $\mathsf{H}^2(P,Q)$ between P and Q is

$$\mathsf{H}^2(P,Q) := \frac{1}{2}\sum_{y\in\mathcal{Y}}\left(\sqrt{P(y)}-\sqrt{Q(y)}\right)^2$$

- squared Hellinger distance between $\mathsf{Bern}(p)$ and $\mathsf{Bern}(q)$:

$$\begin{aligned}\mathsf{H}^2\big(\mathsf{Bern}(p),\mathsf{Bern}(q)\big) &:= \frac{1}{2}\big(\sqrt{p}-\sqrt{q}\big)^2+\frac{1}{2}\big(\sqrt{1-p}-\sqrt{1-q}\big)^2\\ &= \big(1+o(1)\big)\frac{1}{2}\big(\sqrt{p}-\sqrt{q}\big)^2\end{aligned}$$

when $p=o(1)$ and $q=o(1)$

Optimality of spectral method (cont.)

Theorem 4.16 and lower bound (4.19) reveal sharp phase transition:

$$\text{spectral method works} \quad \text{if } \mathsf{H}^2\big(\mathsf{Bern}(p), \mathsf{Bern}(q)\big) \geq (1+\varepsilon)\frac{\log n}{n}$$

$$\text{no algorithm works} \quad \text{if } \mathsf{H}^2\big(\mathsf{Bern}(p), \mathsf{Bern}(q)\big) \leq (1-\varepsilon)\frac{\log n}{n}$$

for arbitrarily small constant $\varepsilon > 0$

Fine-grained analysis of spectral clustering

WLOG, assume $x_1^\star = \ldots = x_{n/2}^\star = 1$ and $x_{n/2+1}^\star = \ldots = x_n^\star = -1$, and recall that

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

These imply

$$\lambda^\star = \tfrac{n(p-q)}{2}, \qquad \mu = 1,$$
$$B = 1, \qquad \sigma^2 \le \max\{p, q\} = p$$

Applying ℓ_∞ perturbation theory

ℓ_∞ perturbation bound (4.18b) yields: for some constant $C > 0$,

$$\begin{aligned}\|z\lambda^\star \boldsymbol{u} - \boldsymbol{M}\boldsymbol{u}^\star\|_\infty &\lesssim \sigma + \frac{\sigma^2\sqrt{n\log n}}{\lambda^\star} + \frac{\sigma B\,\log^{3/2} n}{\lambda^\star} \\ &\le C\Big(\sqrt{p} + \frac{p\sqrt{\log n}}{\sqrt{n}(p-q)} + \frac{\sqrt{p}\log^{3/2} n}{n(p-q)}\Big) =: \Delta \end{aligned} \tag{4.20}$$

It boils down to characterizing entrywise behavior of $\boldsymbol{M}\boldsymbol{u}^\star$
— what happens at phase transition point $(\sqrt{\alpha} - \sqrt{\beta})^2 = 2$?

Bounding entries in $\boldsymbol{M}\boldsymbol{u}^\star$

Lemma 4.18

Suppose that $\left(\sqrt{\alpha}-\sqrt{\beta}\right)^2 \geq 2\left(1+\varepsilon\right)$ *for some constant* $\varepsilon > 0$. *Then with prob.* $1 - o(1)$,

$$\boldsymbol{M}_{l,\cdot}\boldsymbol{u}^\star \geq \frac{\eta \log n}{\sqrt{n}} \text{ for all } l \leq \frac{n}{2}, \quad \boldsymbol{M}_{l,\cdot}\boldsymbol{u}^\star \leq -\frac{\eta \log n}{\sqrt{n}} \text{ for all } l > \frac{n}{2}$$

where $\eta > 0$ *obeys* $(\sqrt{\alpha}-\sqrt{\beta})^2 - \eta \log(\alpha/\beta) > 2$

key message: entries of $\boldsymbol{M}\boldsymbol{u}^\star$ are bounded away from 0 with correct signs if $(\sqrt{\alpha}-\sqrt{\beta})^2 > 2$

Completing the picture

Combine Lemma 4.18 with (4.20) leads to a claim: if

$$\frac{\eta \log n}{\sqrt{n}} > \Delta \tag{4.21}$$

then it follows that

$$z u_l u_l^\star > 0 \quad \text{for all } 1 \leq l \leq n \quad \Longrightarrow \quad \text{exact recovery}$$

Proof of relation (4.21)

Lemma 4.18 and (4.20) tell us that: it suffices to show

$$\frac{\eta \log n}{\sqrt{n}} \geq C\Big(\sqrt{p} + \frac{p\sqrt{\log n}}{\sqrt{n}(p-q)} + \frac{\sqrt{p}\log^{3/2} n}{n(p-q)}\Big)$$

- 1st term: $\sqrt{p} \asymp \sqrt{\frac{\log n}{n}} \ll \frac{\eta \log n}{\sqrt{n}}$
- 2nd term: $\frac{p\sqrt{\log n}}{\sqrt{n}(p-q)} \asymp \sqrt{\frac{\log n}{n}} \ll \frac{\eta \log n}{\sqrt{n}}$
- 3rd term: divide discussion into two cases $\alpha/\beta \leq 2$, and $\alpha/\beta \geq 2$

Comparing two sets of Bernoulli r.v.s

Lemma 4.19

Suppose $\alpha > \beta$, $\{W_i\}_{1 \le i \le n/2}$ *are i.i.d.* Bern$(\frac{\alpha \log n}{n})$, *and* $\{Z_i\}_{1 \le i \le n/2}$ *are i.i.d.* Bern$(\frac{\beta \log n}{n})$, *which are independent of* W_i. *For any* $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \le t \log n\right) \le n^{-(\sqrt{\alpha}-\sqrt{\beta})^2/2 + t\log(\alpha/\beta)/2}$$

Comparing two sets of Bernoulli r.v.s

Lemma 4.19

Suppose $\alpha > \beta$, $\{W_i\}_{1 \le i \le n/2}$ *are i.i.d.* Bern$(\frac{\alpha \log n}{n})$, *and* $\{Z_i\}_{1 \le i \le n/2}$ *are i.i.d.* Bern$(\frac{\beta \log n}{n})$, *which are independent of* W_i. *For any* $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \le t \log n\right) \le n^{-(\sqrt{\alpha}-\sqrt{\beta})^2/2 + t\log(\alpha/\beta)/2}$$

- roughly speaking,

$$\mathbb{P}\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \le 0\right) \le n^{-(\sqrt{\alpha}-\sqrt{\beta})^2/2}$$

Comparing two sets of Bernoulli r.v.s

Lemma 4.19

Suppose $\alpha > \beta$, $\{W_i\}_{1\le i\le n/2}$ *are i.i.d.* Bern$(\frac{\alpha \log n}{n})$, *and* $\{Z_i\}_{1\le i\le n/2}$ *are i.i.d.* Bern$(\frac{\beta \log n}{n})$, *which are independent of* W_i. *For any* $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \le t \log n\right) \le n^{-(\sqrt{\alpha}-\sqrt{\beta})^2/2 + t\log(\alpha/\beta)/2}$$

- $(\sqrt{\alpha}-\sqrt{\beta})^2 > 2$ guarantees $\mathbb{P}\big(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \le 0\big) < n^{-1}$
  - probability of error $o(n^{-1})$ is crucial, since in $\boldsymbol{M}\boldsymbol{u}^\star$ we have n independent groups of $\{W_i\}$ and $\{Z_i\}$ (need union bound)

Proof of Lemma 4.18

Note that $\boldsymbol{M}\boldsymbol{u}^\star = (\boldsymbol{A} - \frac{p+q}{2}\mathbf{1}\mathbf{1}^\top)\boldsymbol{u}^\star = \boldsymbol{A}\boldsymbol{u}^\star$. Hence

$$\boldsymbol{M}_{1,:}\boldsymbol{u}^\star = \boldsymbol{A}_{1,:}\boldsymbol{u}^\star = \frac{1}{\sqrt{n}}\left(\sum_{j=1}^{n/2} A_{1,j} - \sum_{j=n/2+1}^{n} A_{1,j}\right)$$

Apply Lemma 4.19 to obtain with probability at least $1 - n^{-(\sqrt{a}-\sqrt{b})^2/2+\eta\log(a/b)/2} = 1 - o(n^{-1})$

$$\boldsymbol{M}_{1,:}\boldsymbol{u}^\star \geq \frac{\eta \log n}{\sqrt{n}}$$

Invoke union bound to complete proof

Proof of Lemma 4.19

We apply the Laplace transform method: for any $\lambda < 0$

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \leq t \log n\right) \\
&\quad = \mathbb{P}\left(\exp\left(\lambda\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i\right)\right) \geq \exp\left(\lambda t \log n\right)\right) \\
&\quad \leq \frac{\mathbb{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i\right)\right)\right]}{\exp\left(\lambda t \log n\right)}
\end{aligned}$$

By independence, one has

$$\mathbb{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i\right)\right)\right] = \prod_{i=1}^{n/2} \mathbb{E}\left[\exp\left(\lambda W_i\right)\right] \mathbb{E}\left[\exp\left(-\lambda Z_i\right)\right]$$

Proof of Lemma 4.19 (cont.)

By definition and using $1 + x \leq e^x$, one has

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda W_i\right)\right] &= \frac{\alpha \log n}{n} \exp\left(\lambda\right) + \left(1 - \frac{\alpha \log n}{n}\right) \\ &\leq \exp\left(\frac{\alpha \log n}{n} \exp\left(\lambda\right) - \frac{\alpha \log n}{n}\right)\end{aligned}$$

Similarly, for Z_i one has

$$\mathbb{E}\left[\exp\left(-\lambda W_i\right)\right] \leq \exp\left(\frac{\beta \log n}{n} \exp\left(-\lambda\right) - \frac{\beta \log n}{n}\right)$$

Combine these two to see that

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda W_i\right)\right] \mathbb{E}\left[\exp\left(-\lambda Z_i\right)\right] \\ &\quad \leq \exp\left(\frac{\log n}{n}\left(\alpha \exp\left(\lambda\right) + \beta \exp\left(-\lambda\right) - \alpha - \beta\right)\right)\end{aligned}$$

Proof of Lemma 4.19 (cont.)

Combine previous two pages to see

$$\log \mathbb{P}\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \leq t \log n\right)$$
$$\leq -\lambda t \log n + \frac{n}{2}\frac{\log n}{n}\left(\alpha \exp(\lambda) + \beta \exp(-\lambda) - \alpha - \beta\right)$$

Set $\lambda = -\log(\alpha/\beta)/2$ to obtain

$$\alpha \exp(\lambda) + \beta \exp(-\lambda) - \alpha - \beta = \alpha\sqrt{\frac{\beta}{\alpha}} + \beta\sqrt{\frac{\alpha}{\beta}} - \alpha - \beta = -\left(\sqrt{\alpha} - \sqrt{\beta}\right)^2$$

thus concluding the proof

$\ell_{2,\infty}$ *eigen-space perturbation theory (rank-*r*)*

Setup and algorithm

- Ground truth: $\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \in \mathbb{R}^{n_1 \times n_2}$, with singular values $\sigma_1^\star \geq \sigma_2^\star \geq \cdots \geq \sigma_r^\star > 0$ (assume $n_1 \leq n_2$)
- Observation: $\boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{E}$
- Convenient notation:

$$\kappa := \frac{\sigma_1^\star}{\sigma_r^\star}, \qquad n := n_1 + n_2$$

- Estimate $\boldsymbol{U}^\star$ (resp. $\boldsymbol{V}^\star$) using rank-r leading left (resp. right) singular subspace $\boldsymbol{U}$ (resp. $\boldsymbol{V}$) of $\boldsymbol{M}$

Question: can we characterize entrywise estimation error of $\boldsymbol{U}$, i.e.

$$\mathsf{dist}_{2,\infty}(\boldsymbol{U}, \boldsymbol{U}^\star) := \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\|_{2,\infty}$$

Assumptions

Noise assumptions: entries of $\boldsymbol{E} = [E_{i,j}]_{1 \le i \le n_1, 1 \le j \le n_2}$ are independent obeying

$$\mathbb{E}[E_{i,j}] = 0, \quad \mathbb{E}[E_{i,j}^2] \le \sigma^2, \quad |E_{i,j}| \le B, \quad \text{for all } i, j$$

and it is assumed that

$$c_{\mathsf{b}} \coloneqq \frac{B}{\sigma \sqrt{n_1/(\mu \log n)}} = O(1)$$

Assumptions

Noise assumptions: entries of $\boldsymbol{E} = [E_{i,j}]_{1 \le i \le n_1, 1 \le j \le n_2}$ are independent obeying

$$\mathbb{E}[E_{i,j}] = 0, \quad \mathbb{E}[E_{i,j}^2] \le \sigma^2, \quad |E_{i,j}| \le B, \quad \text{for all } i, j$$

and it is assumed that

$$c_{\mathsf{b}} := \frac{B}{\sigma \sqrt{n_1/(\mu \log n)}} = O(1)$$

Incoherence parameter of orthonormal matrix $\boldsymbol{U}^\star \in \mathbb{R}^{n \times r}$ is

$$\mu(\boldsymbol{U}^\star) := \frac{n \|\boldsymbol{U}^\star\|_{2,\infty}^2}{r}$$

and for $\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$ we define $\mu := \max\{\mu(\boldsymbol{U}^\star), \mu(\boldsymbol{V}^\star)\}$

$\ell_{2,\infty}$ distance between U and $U^\star$

Need to take into account rotational ambiguity

— which rotation matrix to use?

$\ell_{2,\infty}$ distance between U and $U^\star$

Need to take into account rotational ambiguity

— which rotation matrix to use?

Definition 4.20

For any square matrix $\boldsymbol{Z}$ with SVD $\boldsymbol{Z} = \boldsymbol{U}_Z \boldsymbol{\Sigma}_Z \boldsymbol{V}_Z^\top$, define

$$\mathsf{sgn}(\boldsymbol{Z}) := \boldsymbol{U}_Z \boldsymbol{V}_Z^\top \tag{4.22}$$

to be matrix sign function of $\boldsymbol{Z}$ (solution to Procrustes problem)

Let us employ $\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star)$ and look at

$$\|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star) - \boldsymbol{U}^\star\|_{2,\infty}$$

$\ell_{2,\infty}$ eigen-space perturbation theory

Theorem 4.21

With probability at least $1 - O(n^{-5})$*, one has*

$$\max\left\{\|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top\boldsymbol{U}^\star) - \boldsymbol{U}^\star\|_{2,\infty},\ \|\boldsymbol{V}\mathsf{sgn}(\boldsymbol{V}^\top\boldsymbol{V}^\star) - \boldsymbol{V}^\star\|_{2,\infty}\right\}$$
$$\lesssim \frac{\sigma\sqrt{r}\big(\kappa\sqrt{\frac{n_2}{n_1}\mu} + \sqrt{\log n}\big)}{\sigma_r^\star}$$

provided that $\sigma\sqrt{n\log n} \le c_1\sigma_r^\star$ *for some small constant* $c_1 > 0$

Entrywise matrix reconstruction error

Recall $\boldsymbol{M} = \underbrace{\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top}_{\text{rank } r \text{ approx.}} + \boldsymbol{U}_\perp \boldsymbol{\Sigma}_\perp \boldsymbol{V}_\perp^\top$

Corollary 4.22

In addition, if $\sigma\kappa\sqrt{n\log n} \le c_2\sigma_r^\star$ *for some small enough constant* $c_2 > 0$, *then the following holds with probability at least* $1 - O(n^{-5})$*:*

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^\star\|_\infty \lesssim \sigma\kappa^2\mu r\sqrt{\frac{(n_2/n_1)\log n}{n_1}}$$

De-localization of estimation error

For simplicity, let us consider the case where $\mu, \kappa, n_2/n_1 = O(1)$. Davis-Kahan theorem results in the following ℓ_2 estimation guarantees

$$\mathsf{dist}_{\mathrm{F}}(\boldsymbol{U}, \boldsymbol{U}^\star) \le \sqrt{r}\, \mathsf{dist}(\boldsymbol{U}, \boldsymbol{U}^\star) \lesssim \frac{\sigma\sqrt{nr}}{\sigma_r^\star}$$

In comparison, the $\ell_{2,\infty}$ bound derived in Theorem 4.21 simplifies to

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\right\|_{2,\infty} \le \left\|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top\boldsymbol{U}^\star) - \boldsymbol{U}^\star\right\|_{2,\infty} \lesssim \frac{\sigma\sqrt{r\log n}}{\sigma_r^\star}$$

De-localization of estimation error (cont.)

For the matrix reconstruction error, one has

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^\star\| \le 2\|\boldsymbol{M} - \boldsymbol{M}^\star\| \lesssim \sigma\sqrt{n},$$

which implies $\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{nr}$

In comparison, one has

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^\star\|_\infty \lesssim \sigma r\sqrt{\frac{\log n}{n}}$$

Concluding remarks

Concluding remarks: spectral methods

- A powerful family that permeates data science applications
  - community detection
  - matrix/tensor completion
  - ranking
  - phase retrieval
  - (joint) graph matching
  - robust PCA
  - clustering in mixture models
  - ...
- Simple yet efficient; sometimes optimal (in some weak sense)
- Commonly used to initialize nonconvex optimization algorithms

— *see overview article Chi, Lu, Chen '19*

Concluding remarks: leave-one-out analysis

A power fine-grained analysis framework that proves effective for problems far beyond spectral methods

- robust M-estimation (e.g. El Karoui et al. '13, El Karoui '18)
- generalized power methods (e.g. Zhong, Boumal '18)
- likelihood ratio test in logistic regression (e.g. Sur et al. '19)
- nonconvex optimization (e.g. Ma et al. '20, Chen et al. '19, Cai et al. '22)
- convex relaxation (e.g. Chen et al. '19, '20, '21)
- reinforcement learning (e.g. Agarwal et al. '19, Pananjady et al. '20, Li et al. '23)
- ...

References: general

- "*Spectral methods for data science: A statistical perspective*," Y. Chen, Y. Chi, J. Fan, C. Ma, *Foundations and Trends in Machine Learning*, 2021.
- "*Inference, estimation, and information processing, EE 378B lecture notes*," A. Montanari, Stanford University.
- "*COMS 4772 lecture notes*," D. Hsu, Columbia University.
- "*Spectral algorithms*," R. Kannan, S. Vempala, *Foundations and Trends in Theoretical Computer Science*, 2009.
- "*Principal component analysis for big data*," J. Fan, Q. Sun, W. Zhou, Z. Zhu, *arXiv:1801.01602*, 2018.
- "*Nonconvex optimization meets low-rank matrix factorization: An overview*," Y. Chi, Y. M. Lu, Y. Chen, *IEEE Transactions on Signal Processing*, 2019.

References: matrix perturbation theory

- "*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.
- "*Matrix perturbation theory*," G. W. Stewart, J. Sun, 1990.
- "*A useful variant of the Davis-Kahan theorem for statisticians*," Y. Yu, T. Wang, R. J. Samworth, *Biometrika*, vol. 102, no. 2, pp. 315-323, 2015.
- "*Spectral method and regularized MLE are both optimal for top-K ranking*," Y. Chen, J. Fan, C. Ma, K. Wang, *Annals of Statistics*, 2019.

References: matrix tail bounds

- "*An introduction to matrix concentration inequalities*," J. Tropp, *Foundations and Trends in Machine Learning*, 2015.
- "*Sharp nonasymptotic bounds on the norm of random matrices with independent entries*," A. S. Bandeira, R. van Handel, *Annals of Probability*, 2016.
- "*High-dimensional probability: An introduction with applications in data science*," R. Vershynin, 2018.
- "*High-dimensional statistics: a non-asymptotic viewpoint*," M. Wainwright, *Cambridge University Press*, 2019.

References: applications

- "*Community detection and stochastic block models*," E. Abbe, *Foundations and Trends in Communications and Information Theory*, 2018.
- "*Consistency thresholds for the planted bisection model*," E. Mossel, J. Neeman, A. Sly, *ACM Symposium on Theory of Computing*, 2015.
- "*Exact recovery in the stochastic block model*," E. Abbe, A. S. Bandeira, G. Hall, *IEEE Transactions on information theory*, vol. 62, no. 1, pp. 471-487, 2015.
- "*Exact matrix completion via convex optimization*," E. Candes, B. Recht, *Foundations of Computational Mathematics*, 2019.
- "*Matrix completion from a few entries*," R. Keshavan, A. Montanari, S. Oh, *IEEE Transactions on Information Theory*, 2010.
- "*Low-rank matrix completion using alternating minimization*," P. Jain, P. Netrapalli, S. Sanghavi, *Symposium on Theory of computing*, 2013.

References: applications

- "*Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion*," C. Ma, K. Wang, Y. Chi, Y. Chen, *Foundations of Computational Mathematics*, 2020.
- "*Inference and uncertainty quantification for noisy matrix completion*," Y. Chen, J. Fan, C. Ma, Y. Yan, *Proceedings of the National Academy of Sciences*, vol. 116, no. 46, pp. 22931-22937, 2019.
- "*Spectral algorithms for tensor completion*," A. Montanari, N. Sun, *Communications on Pure and Applied Mathematics*, vol. 71, no. 11, pp. 2381-2425, 2018.
- "*Nonconvex low-rank tensor completion from noisy data*," C. Cai, G. Li, H. V. Poor, Y. Chen, *Operations Research*, vol. 70, no. 2, 2022.
- "*The PageRank citation ranking: bringing order to the web*," L. Page, S. Brin, R. Motwani, T. Winograd, 1999.

References: applications

- "*Rank centrality: ranking from pairwise comparisons*," S. Negahban, S. Oh, D. Shah, *Operations Research*, 2017.
- "*Heteroskedastic PCA: Algorithm, optimality, and applications*," A. Zhang, T. Cai, Y. Wu, *Annals of Statistics*, 2022.
- "*Normal approximation and confidence region of singular subspaces*," D. Xia, *Electronic Journal of Statistics*, vol. 15, no. 2, pp. 3798-3851.
- "*Phase retrieval via Wirtinger flow: Theory and algorithms*," E. J. Candes, X. Li, M. Soltanolkotabi, *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985-2007, 2015.
- "*Solving random quadratic systems of equations is nearly as easy as solving linear systems*," Y. Chen, E. J. Candes, *Communications on pure and applied mathematics*, vol. 70, no. 5, pp. 822-883, 2017.

References: applications

- "*Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics*," T. Cai, A. Zhang, *Annals of Statistics*, vol. 46, no. 1, pp. 60-89, 2018.
- "*Optimality of spectral clustering in the Gaussian mixture model*," M. Loffler, A. Y. Zhang, H. H. Zhou, *Annals of Statistics*, vol. 49, no. 5, pp. 2506-2530, 2021.
- "*Angular synchronization by eigenvectors and semidefinite programming*," A. Singer, *Applied and computational harmonic analysis*, vol. 30, no. 1, pp. 20-36, 2011.

References: ℓ_∞ and $\ell_{2,\infty}$ theory

- "*Near-optimal bounds for phase synchronization*," Y. Zhong, N. Boumal, *SIAM Journal on Optimization*, 2018.
- "*Entrywise eigenvector analysis of random matrices with low expected rank*," E. Abbe, J. Fan, K. Wang, Y. Zhong, *Annals of Statistics*, 2020.
- "*Spectral method and regularized MLE are both optimal for top-K ranking*," Y. Chen, J. Fan, C. Ma, K. Wang, *Annals of Statistics*, 2019.
- "*The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics*," J. Cape, M. Tang, C. E. Priebe, *Annals of Statistics*, vol. 47, no. 5, pp. 2405-2439, 2019.
- "*Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion*," C. Ma, K. Wang, Y. Chi, Y. Chen, *Foundations of Computational Mathematics*, 2020.

References: ℓ_∞ and $\ell_{2,\infty}$ theory

- "*Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees*," C. Cai, G. Li, Y. Chi, H. V. Poor, Y. Chen, vol. 49, no. 2, 2021.
- "*Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices*," L. Lei, arXiv:1909.04798, 2019.
- "*SIMPLE: Statistical inference on membership profiles in large networks*," J. Fan, Y. Fan, X. Han, J. Lv, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 2, pp. 630-653, 2022.
- "*Entrywise Estimation of Singular Vectors of Low-Rank Matrices with Heteroskedasticity and Dependence*," J. Agterberg, Z. Lubberts, C. Priebe, *IEEE Transactions on Information Theory*, vol. 68, no. 7, pp. 4618-4650, 2022.

References: ℓ_∞ and $\ell_{2,\infty}$ theory

- "*Partial recovery for top-K ranking: optimality of MLE and suboptimality of the spectral method*," P. Chen, C. Gao, and A. Y. Zhang, *Annals of Statistics*, vol. 50, no. 3, pp. 1618-1652, 2022.
- "*Uncertainty quantification in the Bradley-Terry-Luce model*," C. Gao, Y. Shen, A. Y. Zhang, *Information and Inference: A Journal of the IMA*, vol. 12, no. 2, pp. 1073-1140, 2023.
- "*Inference for heteroskedastic PCA with missing data*," Y. Yan, Y. Chen, J. Fan, arXiv:2107.12365, 2021.

References: leave-one-out analysis (other methods)

- "*On robust regression with high-dimensional predictors*," N. El Karoui, D. Bean, P. J. Bickel, C. Lim, B. Yu, *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14557-14562, 2013.
- "*On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators*," N. El Karoui, *Probability Theory and Related Fields*, vol. 170, pp. 95-175, 2018.
- "*Near-optimal bounds for phase synchronization*," Y. Zhong, N. Boumal, *SIAM Journal on Optimization*, 2018.
- "*The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square*," P. Sur, Y. Chen, E. Candes, *Probability theory and related fields*, vol. 175, pp. 487-558, 2019.

References: leave-one-out analysis (other methods)

- "*Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion*," C. Ma, K. Wang, Y. Chi, Y. Chen, *Foundations of Computational Mathematics*, 2020.
- "*Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval*," Y. Chen, Y. Chi, J. Fan, C. Ma, *Mathematical Programming*, vol. 176, pp. 5-37, 2019.
- "*Nonconvex low-rank tensor completion from noisy data*," C. Cai, G. Li, H. V. Poor, Y. Chen, *Operations Research*, vol. 70, no. 2, 2022.
- "*Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization*," J. Chen, D. Liu, X. Li, *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5806-5841, 2020.

References: leave-one-out analysis (other methods)

- "*Inference and uncertainty quantification for noisy matrix completion*," Y. Chen, J. Fan, C. Ma, Y. Yan, *Proceedings of the National Academy of Sciences*, vol. 116, no. 46, pp. 22931-22937, 2019.
- "*Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization*," Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, *SIAM journal on optimization*, vol. 30, no. 4, pp. 3098-3121.
- "*Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data*," Y. Chen, J. Fan, C. Ma, Y. Yan, *Annals of Statistics*, vol. 49, no. 5, pp. 2948-2971, 2021.
- "*Model-based reinforcement learning with a generative model is minimax optimal*," A. Agarwal, S. Kakade, L. Yang, *Conference on Learning Theory*, 2020.

References: leave-one-out analysis (other methods)

- "*Breaking the sample size barrier in model-based reinforcement learning with a generative model*," G. Li, Y. Wei, Y. Chi, Y. Chen, *Operations Research*, 2023+.
- "*Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning*," A. Pananjady, M. Wainwright, *IEEE Transactions on Information Theory*, vol. 67, no. 1, pp. 566-585, 2020.