Inference and Uncertainty Quantification for Low-Rank Models

Yuxin Chen

ECE, Princeton University

A ubiquitous low-complexity model

Composition C by Piet Mondrian

reconstructing low-rank structure
from imperfect measurements

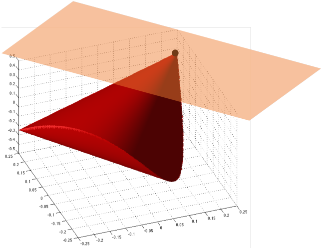
A ubiquitous low-complexity model

Composition C by Piet Mondrian

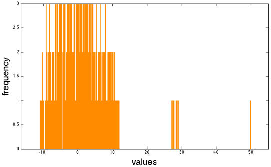
reconstructing low-rank structure
from imperfect measurements

- matrix completion
- solving quadratic equations
- blind deconvolution
- tensor completion
- localization
- PCA / factor models
- community recovery
- joint shape mapping
- linear neural networks
- ...

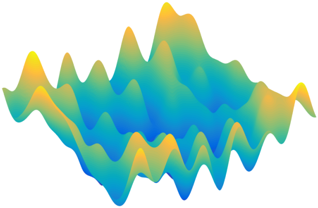
Various estimation schemes have been proposed



convex relaxation



spectral methods



nonconvex optimization

Various estimation schemes have been proposed

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 67, NO. 20, OCTOBER 15, 2019 5239

Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview

Yuejie Chi, Yue M. Lu, and Yuxin Chen

(Overview Article)

Foundations and Trends® in Machine Learning

Spectral Methods for Data Science: A Statistical Perspective

Suggested Citation: Yuxin Chen, Yuejie Chi, Jianqing Fan and Cong Ma (2020), "Spectral Methods for Data Science: A Statistical Perspective", Foundations and Trends® in Machine Learning: Vol. XX, No. XX, pp 1–243. DOI: XXX.

Yuxin Chen
Princeton University
yuxin.chen@princeton.edu

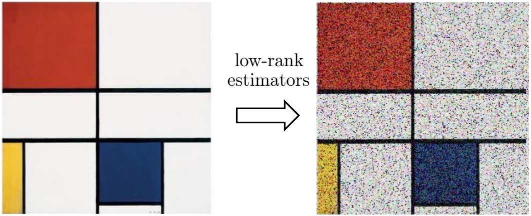
Yuejie Chi
Carnegie Mellon University
yuejiechi@cmu.edu

Jianqing Fan
Princeton University
jqfan@princeton.edu

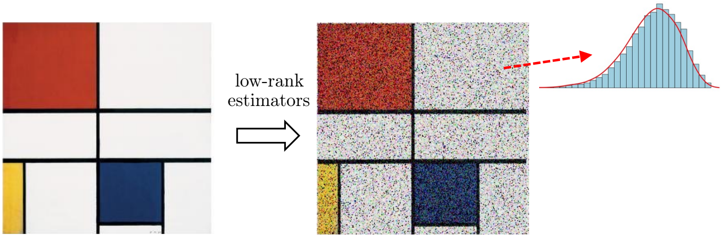
Cong Ma
University of Chicago
congm@uchicago.edu

now
the essence of knowledge
Boston — Delft

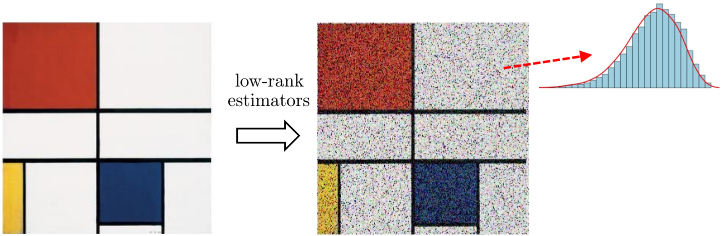
One step further: reasoning about uncertainty?



One step further: reasoning about uncertainty?



One step further: reasoning about uncertainty?



How to assess uncertainty, or “confidence”, of obtained low-rank estimates due to imperfect data acquisition?

- noise
- missing data
- $\cdots$

Inference in high dimensional regression

organized by
Peter Buehlmann, Andrea Montanari, and Jonathan Taylor

American Institute of Mathematics

The open problems discussion was also very productive, and led to formulating a selection of special topics addressed in the working groups. These were

⋮

(3) *Confidence intervals for matrix completion.* In matrix completion, the data analyst is given a large data matrix with a number of missing entries. In many interesting applications (e.g. to collaborative filtering) it is indeed the case that the vast majority of entries is missing. In order to fill the missing entries, the assumption is made that the underlying –unknown– matrix has a low-rank structure.

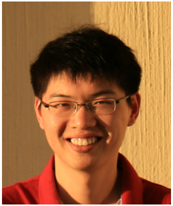
Substantial work has been devoted to methods for computing point estimates of the missing entries. In applications, it would be very interesting to compute confidence intervals as well. This requires developing distributional characterizations of standard matrix completion methods.

This talk: two recent examples

1. Inference for noisy matrix completion

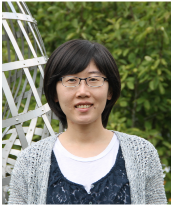
2. Inference for heteroskedastic PCA with missing data

Part 1: Inference for noisy matrix completion

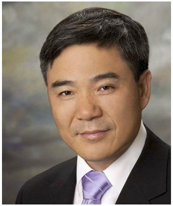


Cong Ma
U Chicago Stats

Yuling Yan
Princeton ORFE



Yuejie Chi
CMU ECE



Jianqing Fan
Princeton ORFE

Noisy low-rank matrix completion

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

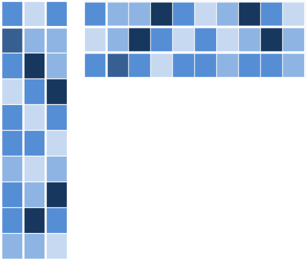
figure credit: E. J. Candès

Given partial samples of a <u>low-rank</u> matrix $M^\star$, fill in missing entries

Noisy low-rank matrix completion

observations: $\quad M_{i,j} = M^\star_{i,j} + \text{noise}, \quad (i,j) \in \Omega$

goal: $\quad$ estimate $\boldsymbol{M}^\star$



unknown rank-r matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$

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

sampling set Ω

Noisy low-rank matrix completion

observations: $M_{i,j} = M^\star_{i,j} + \text{noise}, \quad (i,j) \in \Omega$

goal: estimate $\boldsymbol{M}^\star$

A widely used convex relaxation method:

$$\underset{\boldsymbol{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \underbrace{\sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2}_{\text{squared loss}} + \lambda \|\boldsymbol{Z}\|_*$$

Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z}} \underbrace{\sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2}_{\text{empirical loss}} + \lambda \|\boldsymbol{Z}\|_*$$

- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm

Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z}} \underbrace{\sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2}_{\text{empirical loss}} + \lambda \|\boldsymbol{Z}\|_*$$

- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm
- highly challenging to pin down distributions of obtained estimates

Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z}} \underbrace{\sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2}_{\text{empirical loss}} + \lambda \|\boldsymbol{Z}\|_*$$

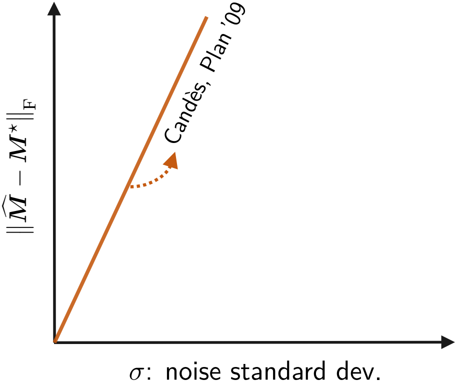
- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm
- highly challenging to pin down distributions of obtained estimates
- existing estimation error bounds are highly sub-optimal
  $\longrightarrow$ overly wide confidence intervals

Step 1: sharpening estimation guarantees

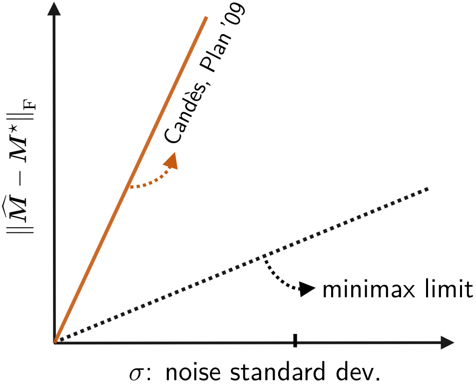
Prior statistical guarantees for convex relaxation

- **random sampling:** each $(i,j) \in \Omega$ with prob. p
- **random noise:** i.i.d. Gaussian noise with mean zero and variance σ^2
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: rank $r = O(1)$, incoherent, well-conditioned ...

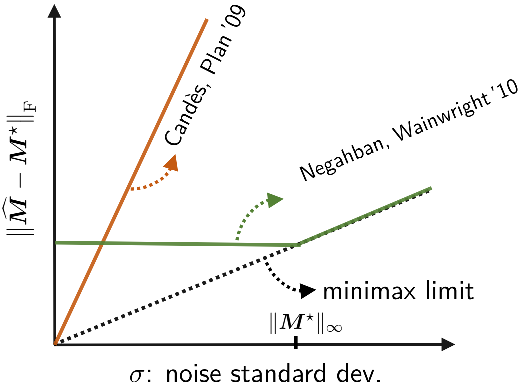
Candès, Plan '09	$\sigma n^{1.5}$



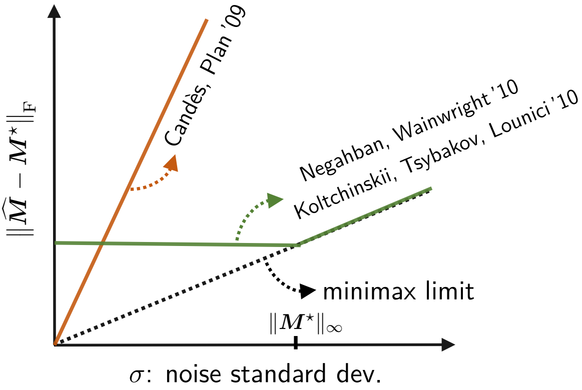
minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$

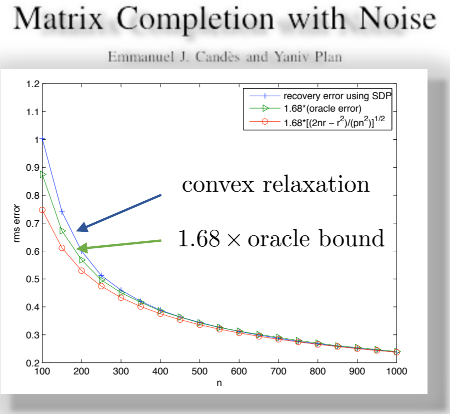


minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \|\boldsymbol{M}^{\star}\|_{\infty}\}\sqrt{n/p}$



minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$
Koltchinskii, Tsybakov, Lounici '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$





Existing theory for convex relaxation does not match practice . . .

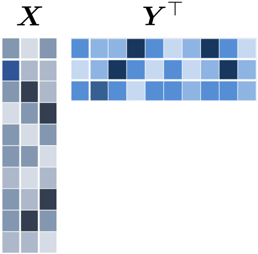
dual certification (golfing scheme)

dual certification (golfing scheme)

nonconvex optimization

A detour: nonconvex optimization

Burer–Monteiro: represent $\boldsymbol{Z}$ by $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\underbrace{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}}_{\text{low-rank factors}}$



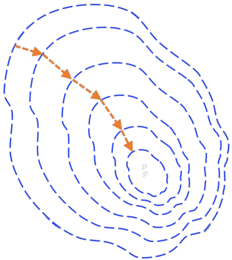
$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \underbrace{\sum_{(i,j)\in\Omega}\left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{i,j} - M_{i,j}\right]^2}_{\text{squared loss}} + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

A detour: nonconvex optimization

- Burer, Monteiro '03
- Rennie, Srebro '05
- Keshavan, Montanari, Oh '09 '10
- Jain, Netrapalli, Sanghavi '12
- Hardt '13
- Sun, Luo '14
- Chen, Wainwright '15
- Tu, Boczar, Simchowitz, Soltanolkotabi, Recht '15
- Zhao, Wang, Liu '15
- Zheng, Lafferty '16
- Yi, Park, Chen, Caramanis '16
- Ge, Lee, Ma '16
- Ge, Jin, Zheng '17
- Ma, Wang, Chi, Chen '17
- Chen, Li '18
- Chen, Liu, Li '19
- ...

A detour: nonconvex optimization

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \sum_{(i,j)\in\Omega}\left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{i,j} - M_{i,j}\right]^2 + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

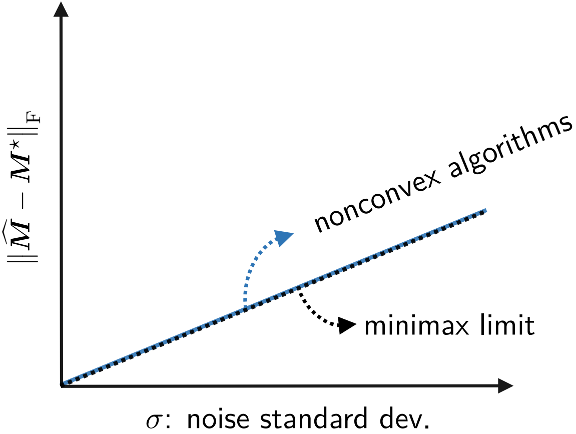


- **suitable initialization:** $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$
- **gradient descent:** for $t = 0, 1, \ldots$

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$
$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta_t \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$

A detour: nonconvex optimization

minimax limit	$\sigma\sqrt{n/p}$
nonconvex algorithms	$\sigma\sqrt{n/p}$ (optimal!)



A motivating experiment

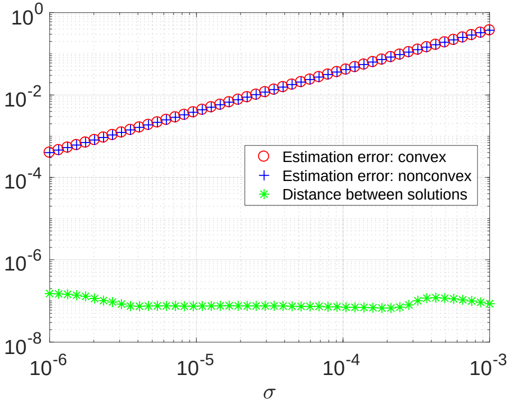
convex: $$\underset{\boldsymbol{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \sum_{(i,j) \in \Omega} \left(Z_{i,j} - M_{i,j} \right)^2 + \lambda \|\boldsymbol{Z}\|_*$$

nonconvex: $$\underset{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n \times r}}{\text{minimize}} \sum_{(i,j) \in \Omega} \left[\left(\boldsymbol{X} \boldsymbol{Y}^\top \right)_{i,j} - M_{i,j} \right]^2 + \underbrace{\frac{\lambda}{2} \|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2} \|\boldsymbol{Y}\|_{\mathrm{F}}^2}_{\text{reg}(\boldsymbol{X}, \boldsymbol{Y})}$$

— $$\|\boldsymbol{Z}\|_* = \min_{\boldsymbol{Z} = \boldsymbol{X} \boldsymbol{Y}^\top} \frac{1}{2} \|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{1}{2} \|\boldsymbol{Y}\|_{\mathrm{F}}^2$$

A motivating experiment

$$n = 1000, \; r = 5, \; p = 0.2, \; \lambda = 5\sigma\sqrt{np}$$



Convex and nonconvex solutions are exceedingly close!

convex  nonconvex

$$\Downarrow$$

$$\text{stability}\left(\boxed{\text{convex}}\right) \approx \text{stability}\left(\boxed{\text{nonconvex}}\right)$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

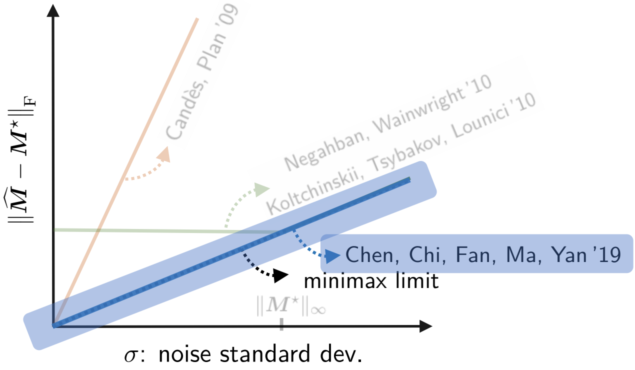
$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Theorem 1 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

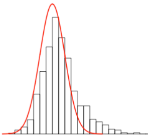
1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*
2. $$\left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$

$$\|M^{\mathsf{cvx}} - M^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}} : \quad \text{minimax optimal when } r = O(1)$$



Step 2: from estimation to inference ...

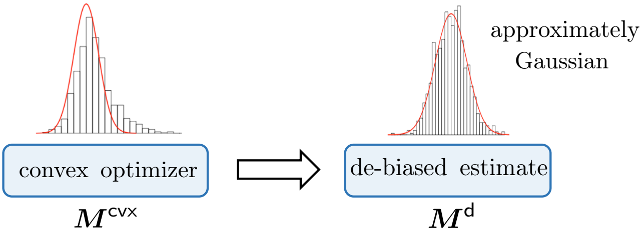
— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



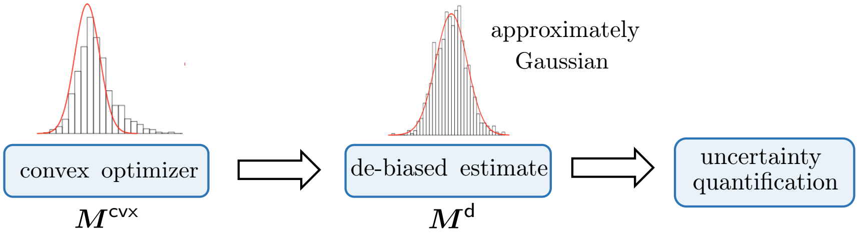
convex optimizer

M^{cvx}

— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



De-biasing convex estimate

$$\boldsymbol{M}^{\mathsf{cvx}} \quad \overset{\text{de-biasing}}{\longrightarrow} \quad \underbrace{\boldsymbol{M}^{\mathsf{cvx}} + \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\mathsf{cvx}})}_{\text{(nearly) unbiased estimate of } \boldsymbol{M}^{\star}}$$

De-biasing convex estimate

$$\boldsymbol{M}^{\mathsf{cvx}} \quad \xrightarrow{\text{de-biasing}} \quad \underbrace{\boldsymbol{M}^{\mathsf{cvx}} + \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\mathsf{cvx}})}_{\text{(nearly) unbiased estimate of } \boldsymbol{M}^{\star}}$$

- **issue:** high-rank after de-biasing; statistical accuracy suffers

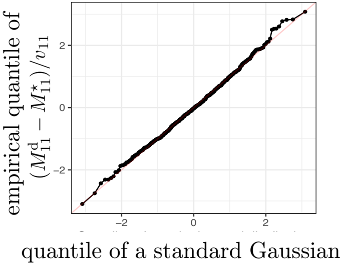
De-biasing convex estimate

$$M^{\mathsf{cvx}} \xrightarrow{\text{de-biasing}} \underbrace{\mathsf{proj}_{\mathsf{rank}\text{-}r}\Big(M^{\mathsf{cvx}} + \tfrac{1}{p}\mathcal{P}_{\Omega}(M^{\star} + E - M^{\mathsf{cvx}}) \Big)}_{\text{1 iteration of singular value projection (Jain, Meka, Dhillon '10)}} =: M^{\mathrm{d}}$$

- **issue:** high-rank after de-biasing; statistical accuracy suffers
- **solution:** low-rank projection

Distributional theory

- random sampling
- i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2)$
- ground truth: low-rank, incoherent, well-conditioned



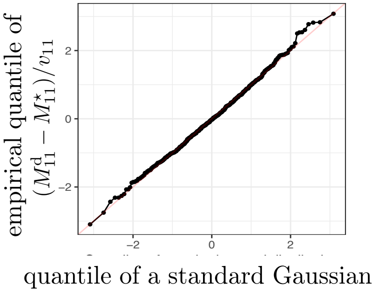
Theorem 2 (Chen, Fan, Ma, Yan '19)

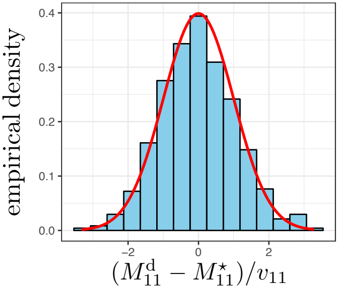
Consider any (i,j) *s.t.* $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ *is not too small. Then*

$$M^{\mathsf{d}}_{ij} - M^{\star}_{ij} \sim \mathcal{N}(\mathbf{0}, \text{Cramer-Rao}) + \textit{negligible term}$$

— *asymptotically optimal!*

Numerical experiments





$$n = 1000,\ p = 0.2,\ r = 5,\ \|\boldsymbol{M}^{\star}\| = 1,\ \kappa = 1,\ \sigma = 10^{-3}$$

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

Theorem 3 (Chen, Fan, Ma, Yan '19)

$$\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}}^2 = \underbrace{\frac{(2+o(1))nr\sigma^2}{p}}_{\text{Cramer-Rao lower bound}} \qquad \textit{with high prob.}$$

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

Theorem 3 (Chen, Fan, Ma, Yan '19)

$$\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^\star\|_{\mathrm{F}}^2 = \underbrace{\frac{(2+o(1))nr\sigma^2}{p}}_{\textit{Cramer-Rao lower bound}} \qquad \textit{with high prob.}$$

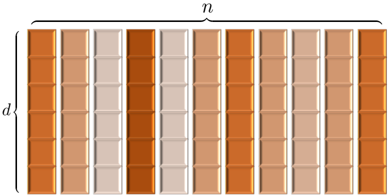
- precise characterization of estimation accuracy
- achieves full statistical efficiency (including pre-constant)

Part 2: Inference for heteroskedastic PCA with missing data

Yuling Yan
Princeton ORFE

Jianqing Fan
Princeton ORFE

Principal component analysis



$$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$$

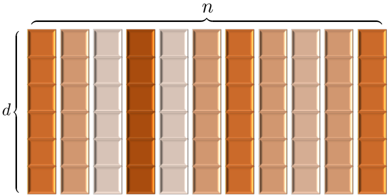
- Ground-truth data

$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}, \qquad \boldsymbol{x}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}^\star)$$

where $\boldsymbol{S}^\star = \boldsymbol{U}^\star \boldsymbol{\Lambda}^\star \boldsymbol{U}^{\star\top} \in \mathbb{R}^{d\times d}$

Principal component analysis

span$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subseteq \boldsymbol{U}^\star$ (r-dimensional)



$$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$$

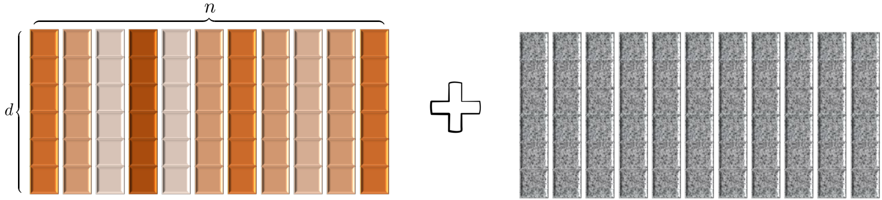
- Ground-truth data

$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}, \qquad \boldsymbol{x}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}^\star)$$

where $\boldsymbol{S}^\star = \boldsymbol{U}^\star \boldsymbol{\Lambda}^\star \boldsymbol{U}^{\star\top} \in \mathbb{R}^{d \times d}$ has rank r

Principal component analysis

span$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subseteq \boldsymbol{U}^\star$ (r-dimensional)



$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$ noise matrix: $\boldsymbol{E}$

- Ground-truth data

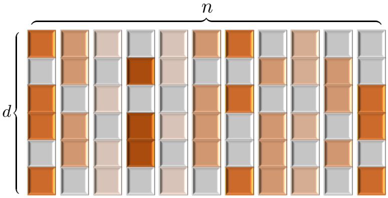
$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}, \qquad \boldsymbol{x}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}^\star)$$

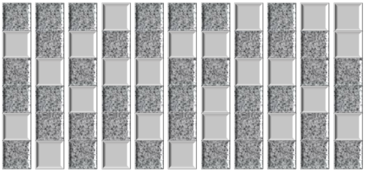
where $\boldsymbol{S}^\star = \boldsymbol{U}^\star \boldsymbol{\Lambda}^\star \boldsymbol{U}^{\star\top} \in \mathbb{R}^{d \times d}$ has rank r

- Noisy observations: $\boldsymbol{X} + \boldsymbol{E}$ (a.k.a. spiked covariance model)

Principal component analysis

span$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subseteq \boldsymbol{U}^\star$ (r-dimensional)





$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$

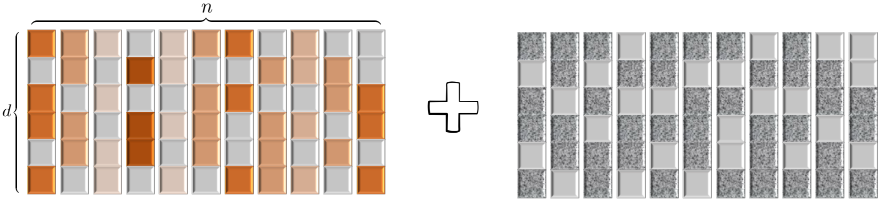
noise matrix: $\boldsymbol{E}$

- Incomplete observations $\longrightarrow$ sampling set Ω:

$$Y_{i,j} = \begin{cases} X^\star_{i,j} + E_{i,j}, & (i,j) \in \Omega \\ 0, & \text{else} \end{cases} \qquad \text{or} \qquad \boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{E})$$

Principal component analysis

span$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subseteq \boldsymbol{U}^\star$ (r-dimensional)



$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$

noise matrix: $\boldsymbol{E}$

- **Goal:**
  - Construct confidence regions for principal subspace $\boldsymbol{U}^\star$
  - Construct entrywise confidence intervals for covariance matrix $\boldsymbol{S}^\star$

What we consider here ...

- **Heteroskedastic noise**: $\{E_{i,j}\}$ are ind. sub-Gaussian obeying

$$\mathbb{E}[E_{i,j}] = 0, \quad \mathbb{E}[E_{i,j}^2] = \omega_i^{\star 2} \in [\omega_{\min}^2, \omega_{\max}^2], \qquad \underbrace{\|E_{i,j}\|_{\psi_2}}_{\text{sub-Gaussian norm}} = O(\omega_i^\star)$$

  - noise variance $\{\omega_i^{\star 2}\}$: unknown, location-varying

What we consider here ...

- **Heteroskedastic noise**: $\{E_{i,j}\}$ are ind. sub-Gaussian obeying

$$\mathbb{E}[E_{i,j}] = 0, \quad \mathbb{E}[E_{i,j}^2] = \omega_i^{\star 2} \in [\omega_{\min}^2, \omega_{\max}^2], \qquad \underbrace{\|E_{i,j}\|_{\psi_2}}_{\text{sub-Gaussian norm}} = O(\omega_i^\star)$$

  - noise variance $\{\omega_i^{\star 2}\}$: unknown, location-varying

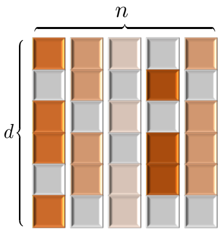
- **Random sampling**: $(i, j) \in \Omega$ independently with prob. p

What we consider here ...

Our focus: estimating/inferring column subspace when $\underbrace{n \gg d}_{\text{more challenging regime}}$

What we consider here . . .

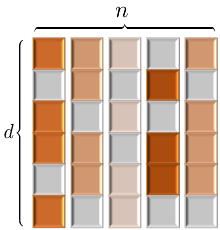
Our focus: estimating/inferring column subspace when $\underbrace{n \gg d}_{\text{more challenging regime}}$



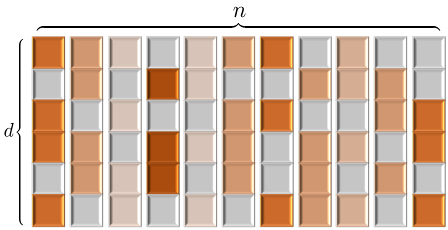
$n \lesssim d$: solvable via matrix completion methods

What we consider here ...

Our focus: estimating/inferring column subspace when $\underbrace{n \gg d}_{\text{more challenging regime}}$



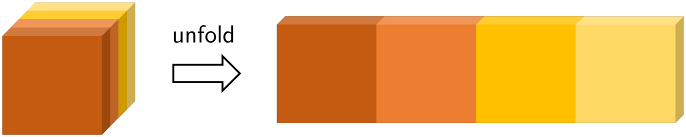
$n \lesssim d$: solvable via matrix completion methods



$n \gg d$: sometimes it's only feasible to estimate col-space instead of whole matrix

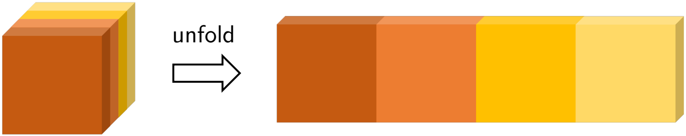
Applications beyond PCA

- Tensor completion

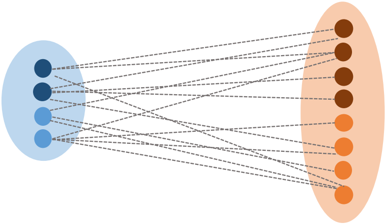


Applications beyond PCA

- Tensor completion



- One-sided community recovery in bipartite random graphs



A natural SVD-based algorithm

- **Compute:** rank-r SVD $\boldsymbol{U\Sigma V}^\top$ of $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{E})$
- **Output:** $\boldsymbol{U} \longrightarrow$ estimate of $\boldsymbol{U}^\star$

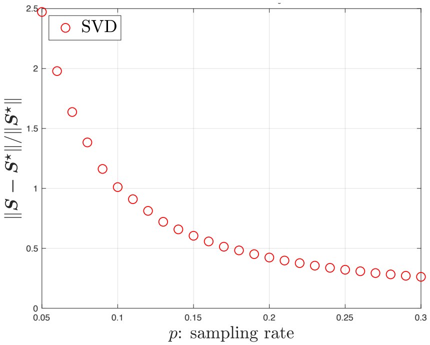
A natural SVD-based algorithm

- **Compute:** rank-r SVD $\boldsymbol{U\Sigma V}^\top$ of $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{E})$
- **Output:** $\boldsymbol{U} \;\longrightarrow\;$ estimate of $\boldsymbol{U}^\star$

Rationale: when $\mathbb{E}[\boldsymbol{E}] = \boldsymbol{0}$ and Ω is randomly sampled, we have

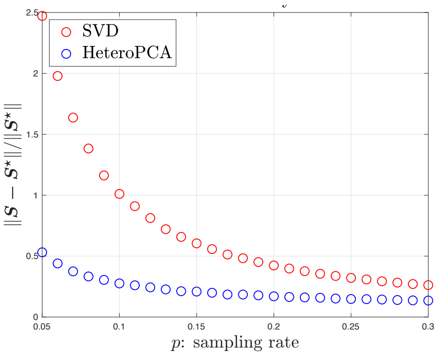
$$\text{col-space}(\,\mathbb{E}[\boldsymbol{Y}]\,) = \text{col-space}(\boldsymbol{X}) = \boldsymbol{U}^\star$$

Numerical suboptimality of SVD-based approach



$$n = 2000, \quad d = 100, \quad r = 3, \quad \omega_1^\star, \cdots, \omega_d^\star \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[0.025, 0.1]$$

Numerical suboptimality of SVD-based approach



$$n = 2000, \quad d = 100, \quad r = 3, \quad \omega_1^\star, \cdots, \omega_d^\star \overset{\text{i.i.d.}}{\sim} \text{Unif}[0.025, 0.1]$$

Plain SVD is suboptimal in the presence of missing data if $n \gg d$

Diagonal entries need special treatment

$$\text{col-space}(\boldsymbol{Y}) = \text{eig-space}(\boldsymbol{Y}\boldsymbol{Y}^\top) = \text{eig-space}\Big(\mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{E})\mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{E})^\top\Big)$$

Diagonal entries need special treatment

$$\text{col-space}(\boldsymbol{Y}) = \text{eig-space}(\boldsymbol{Y}\boldsymbol{Y}^\top) = \text{eig-space}\Big(\mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{E})\mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{E})^\top\Big)$$

Large bias in diagonal entries:

$$\frac{1}{p^2}\mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\top] = \underbrace{\boldsymbol{X}\boldsymbol{X}^\top}_{\checkmark} + \underbrace{\left(\frac{1}{p}-1\right)\mathcal{P}_{\mathsf{diag}}(\boldsymbol{X}\boldsymbol{X}^\top) + \frac{n}{p}\mathsf{diag}\Big\{[\omega_i^{\star 2}]\Big\}}_{\text{potentially large diagonal matrix!}}$$

Diagonal entries need special treatment

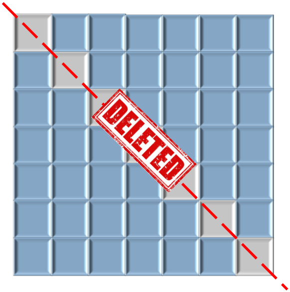
$$\text{col-space}(\boldsymbol{Y}) = \text{eig-space}(\boldsymbol{Y}\boldsymbol{Y}^\top) = \text{eig-space}\Big(\mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{E})\mathcal{P}_\Omega(\boldsymbol{X}+\boldsymbol{E})^\top\Big)$$

Large bias in diagonal entries:

$$\frac{1}{p^2}\mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\top] = \underbrace{\boldsymbol{X}\boldsymbol{X}^\top}_{\checkmark} + \underbrace{\left(\frac{1}{p}-1\right)\mathcal{P}_{\text{diag}}(\boldsymbol{X}\boldsymbol{X}^\top) + \frac{n}{p}\text{diag}\Big\{[\omega_i^{\star 2}]\Big\}}_{\text{potentially large diagonal matrix!}}$$

- a common issue under missing data or heteroskedastic noise

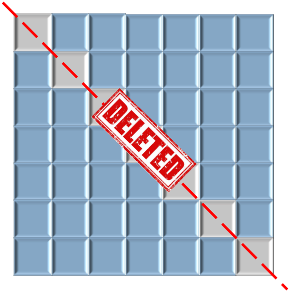
Two algorithms that take care of diagonals

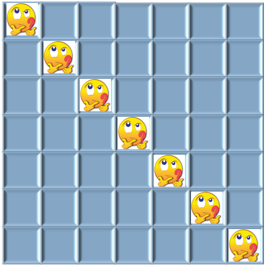


diagonal-deleted PCA:

- remove diag($\boldsymbol{Y}\boldsymbol{Y}^\top$)
- compute top-r eigen-space

Two algorithms that take care of diagonals





diagonal-deleted PCA:

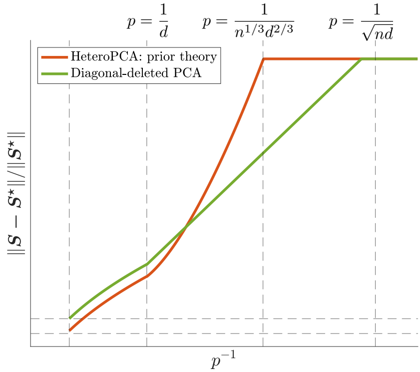
- remove diag$(\boldsymbol{Y}\boldsymbol{Y}^\top)$
- compute top-r eigen-space

HeteroPCA (Zhang et al '18)

- iteratively estimate diag$(\boldsymbol{Y}\boldsymbol{Y}^\top)$
- compute top-r eigen-space

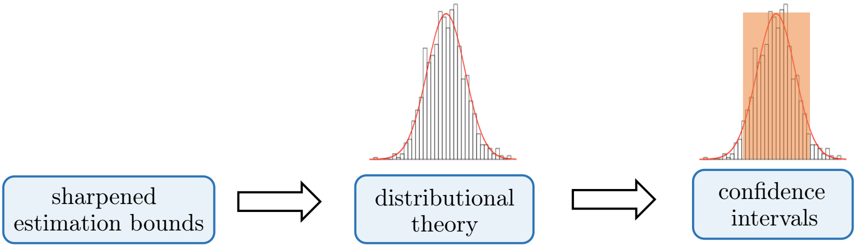
prior theory

(noiseless, $n > d$)



	$\|\cdot\|$ estimation error bounds	min sample size requirement
HeteroPCA (Zhang et al. '18)	$\frac{1}{\sqrt{nd^2p^3}} + \frac{1}{\sqrt{np}}$	$n^{\frac{2}{3}} d^{\frac{1}{3}}$
diagonal-deleted PCA (Cai et al. 19)	$\frac{1}{\sqrt{ndp^2}} + \frac{1}{\sqrt{np}} + \frac{1}{d}$	$\sqrt{nd}$

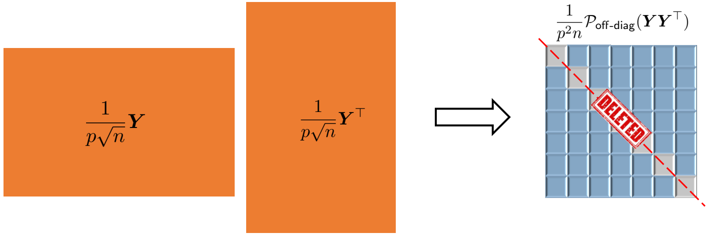
Our contributions: estimation and inference based on HeteroPCA



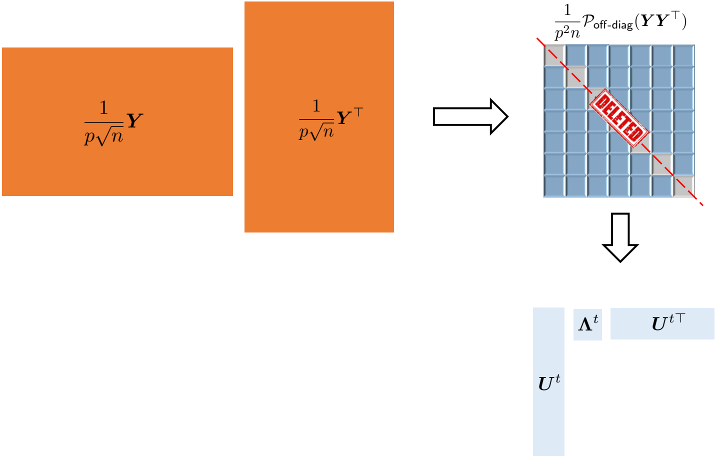
HeteroPCA (Zhang, Cai, Wu '18)

$$\frac{1}{p\sqrt{n}}\boldsymbol{Y} \quad \frac{1}{p\sqrt{n}}\boldsymbol{Y}^{\top}$$

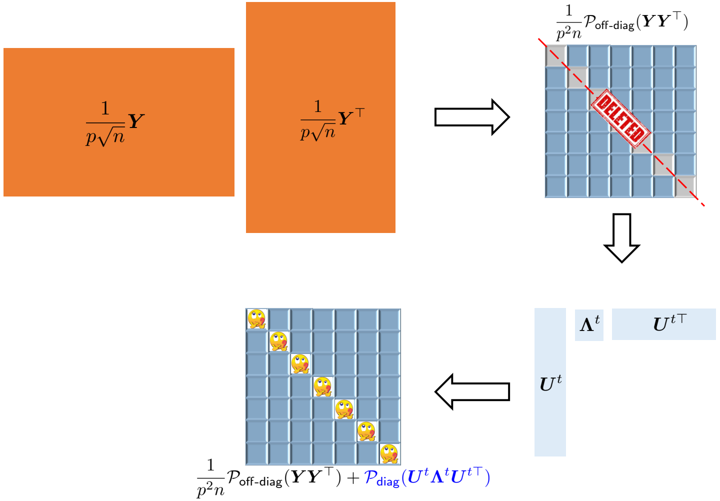
HeteroPCA (Zhang, Cai, Wu ’18)



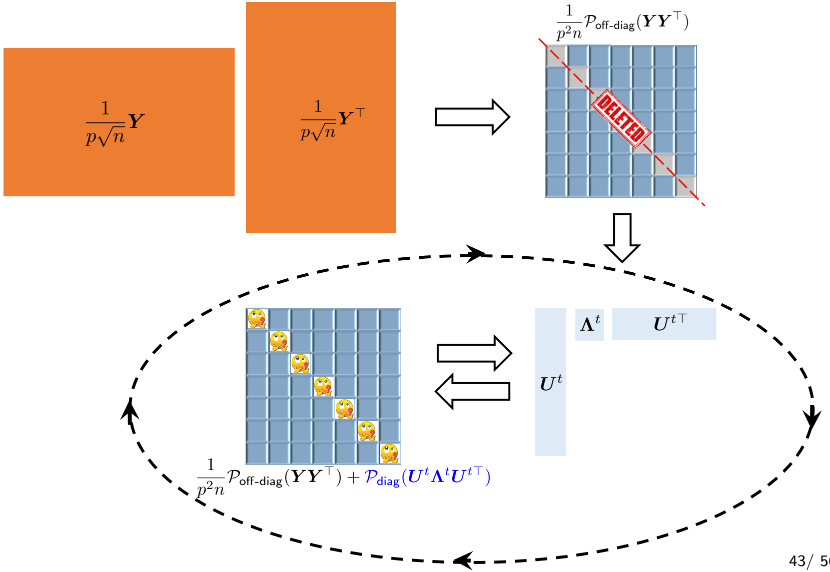
HeteroPCA (Zhang, Cai, Wu ’18)



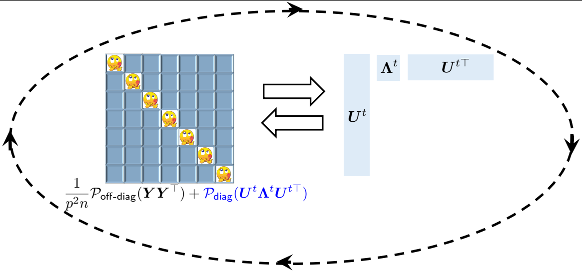
HeteroPCA (Zhang, Cai, Wu '18)



HeteroPCA (Zhang, Cai, Wu '18)



HeteroPCA (Zhang, Cai, Wu '18)

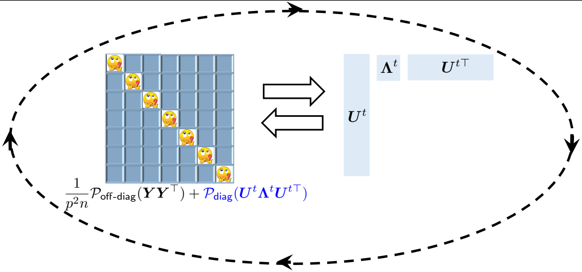


- **Initialize:** $G^0 = \frac{1}{np^2}\mathcal{P}_{\text{off-diag}}(YY^\top)$
- **Iterative update:** for $t = 0, 1, \ldots, t_0$

$$(U^t, \Lambda^t) = \text{eigs}(G^t, r)$$

$$G^{t+1} = G^0 + \mathcal{P}_{\text{diag}}(U^t \Lambda^t U^{t\top})$$

HeteroPCA (Zhang, Cai, Wu '18)



- **Initialize:** $\boldsymbol{G}^0 = \frac{1}{np^2}\mathcal{P}_{\text{off-diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)$
- **Iterative update:** for $t = 0, 1, \ldots, t_0$

$$(\boldsymbol{U}^t, \boldsymbol{\Lambda}^t) = \text{eigs}(\boldsymbol{G}^t, r)$$

$$\boldsymbol{G}^{t+1} = \boldsymbol{G}^0 + \mathcal{P}_{\text{diag}}(\boldsymbol{U}^t\boldsymbol{\Lambda}^t\boldsymbol{U}^{t\top})$$

- **Output:** $\boldsymbol{U} := \boldsymbol{U}^{t_0} \longrightarrow$ estimate of $\boldsymbol{U}^\star$

$$\boldsymbol{S} := \boldsymbol{U}^{t_0}\boldsymbol{\Lambda}^{t_0}\boldsymbol{U}^{t_0\top} \longrightarrow \text{estimate of } \boldsymbol{S}^\star = \boldsymbol{U}^\star\boldsymbol{\Lambda}^\star\boldsymbol{U}^{\star\top}$$

Sharpened estimation guarantees for HeteroPCA

Assumptions (omitting log factors)

- rank $r = O(1)$, incoherence $\mu = O(1)$, cond. number $\kappa = O(1)$
- signal-to-noise ratio (SNR) cannot be too low:

$$\frac{\omega_{\max}^2}{\lambda_r(\boldsymbol{S}^\star)} \lesssim \min\left\{\frac{np}{n+d}, \sqrt{\frac{np^2}{n+d}}\right\}$$

- sampling rate exceeds certain threshold

$$p \gtrsim \max\left\{\frac{1}{\sqrt{nd}},\ \frac{1}{n}\right\}$$

Sharpened estimation guarantees for HeteroPCA

Theorem 4 (Yan, Chen, Fan '21)

With high prob., we have

$$\|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star) - \boldsymbol{U}^\star\| \lesssim \zeta_{\mathsf{op}}, \quad \|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star) - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \frac{1}{\sqrt{d}}\,\zeta_{\mathsf{op}}$$

$$\|\boldsymbol{S} - \boldsymbol{S}^\star\| \lesssim \zeta_{\mathsf{op}}\,\lambda_1^\star, \quad \|\boldsymbol{S} - \boldsymbol{S}^\star\|_\infty \lesssim \frac{1}{d}\,\zeta_{\mathsf{op}}\lambda_1^\star$$

where $\zeta_{\mathsf{op}} := \frac{1}{\sqrt{nd}\,p} + \frac{\omega_{\max}^2}{p\,\lambda_r^\star}\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{np}} + \frac{\omega_{\max}}{\sqrt{\lambda_r^\star}}\sqrt{\frac{d}{np}}$

Sharpened estimation guarantees for HeteroPCA

Theorem 4 (Yan, Chen, Fan '21)

With high prob., we have

$$\|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star) - \boldsymbol{U}^\star\| \lesssim \zeta_{\mathsf{op}}, \quad \|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star) - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \frac{1}{\sqrt{d}}\,\zeta_{\mathsf{op}}$$

$$\|\boldsymbol{S} - \boldsymbol{S}^\star\| \lesssim \zeta_{\mathsf{op}}\,\lambda_1^\star, \quad \|\boldsymbol{S} - \boldsymbol{S}^\star\|_\infty \lesssim \frac{1}{d}\,\zeta_{\mathsf{op}}\lambda_1^\star$$

where $\zeta_{\mathsf{op}} := \frac{1}{\sqrt{nd}\,p} + \frac{\omega_{\max}^2}{p\,\lambda_r^\star}\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{np}} + \frac{\omega_{\max}}{\sqrt{\lambda_r^\star}}\sqrt{\frac{d}{np}}$

- fine-grained estimation guarantees ($\ell_{2,\infty}$ and ℓ_∞ bounds)

Sharpened estimation guarantees for HeteroPCA

Theorem 4 (Yan, Chen, Fan '21)

With high prob., we have

$$\|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top\boldsymbol{U}^\star)-\boldsymbol{U}^\star\| \lesssim \zeta_{\mathsf{op}}, \quad \|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top\boldsymbol{U}^\star)-\boldsymbol{U}^\star\|_{2,\infty} \lesssim \frac{1}{\sqrt{d}}\,\zeta_{\mathsf{op}}$$

$$\|\boldsymbol{S}-\boldsymbol{S}^\star\| \lesssim \zeta_{\mathsf{op}}\,\lambda_1^\star, \quad \|\boldsymbol{S}-\boldsymbol{S}^\star\|_\infty \lesssim \frac{1}{d}\,\zeta_{\mathsf{op}}\lambda_1^\star$$

where $\zeta_{\mathsf{op}} := \frac{1}{\sqrt{nd}\,p} + \frac{\omega_{\max}^2}{p\,\lambda_r^\star}\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{np}} + \frac{\omega_{\max}}{\sqrt{\lambda_r^\star}}\sqrt{\frac{d}{np}}$

- fine-grained estimation guarantees ($\ell_{2,\infty}$ and ℓ_∞ bounds)
- estimation errors are spread out across entries

Sharpened estimation guarantees for HeteroPCA

Theorem 4 (Yan, Chen, Fan '21)

With high prob., we have

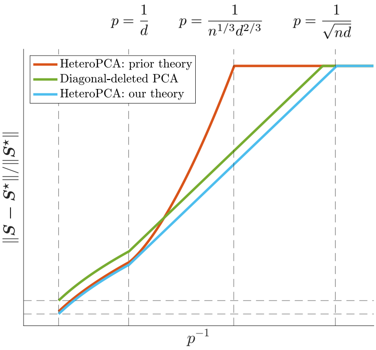
$$\|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top\boldsymbol{U}^\star) - \boldsymbol{U}^\star\| \lesssim \zeta_{\mathsf{op}}, \quad \|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top\boldsymbol{U}^\star) - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \frac{1}{\sqrt{d}}\,\zeta_{\mathsf{op}}$$

$$\|\boldsymbol{S} - \boldsymbol{S}^\star\| \lesssim \zeta_{\mathsf{op}}\,\lambda_1^\star, \quad \|\boldsymbol{S} - \boldsymbol{S}^\star\|_\infty \lesssim \frac{1}{d}\,\zeta_{\mathsf{op}}\lambda_1^\star$$

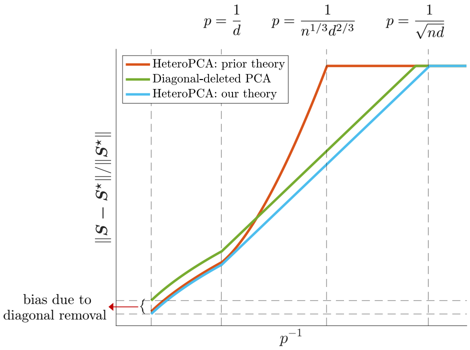
where $\zeta_{\mathsf{op}} := \frac{1}{\sqrt{nd}\,p} + \frac{\omega_{\max}^2}{p\,\lambda_r^\star}\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{np}} + \frac{\omega_{\max}}{\sqrt{\lambda_r^\star}}\sqrt{\frac{d}{np}}$

- fine-grained estimation guarantees ($\ell_{2,\infty}$ and ℓ_∞ bounds)
- estimation errors are spread out across entries
- our sample size and SNR conditions are minimax-optimal (in terms of achieving consistent estimation)

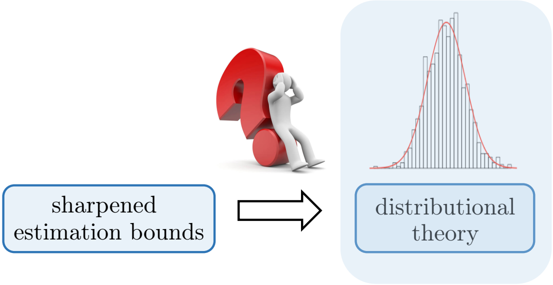
Sharpened estimation guarantees for HeteroPCA



Sharpened estimation guarantees for HeteroPCA



- diagonal-deleted PCA incurs some bias due to diagonal deletion
- $\underbrace{\text{HeteroPCA}}_{\text{method of choice}}$ achieves bias correction via iterative refinement



Can we obtain distributional characterizations for this appealing estimator HeteroPCA?

Distributional theory for U

Theorem 5 (Yan, Chen, Fan '21)

Consider any $1 \le l \le d$ s.t. $\|U^\star_{l,\cdot}\|_2$ is not too small. Under previous assumptions, we have

$$\sup_{\text{cvx set } \mathcal{C}} \left| \mathbb{P}\Big(\Big[\boldsymbol{U} \underbrace{\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star)}_{\text{global rotation}} - \boldsymbol{U}^\star \Big]_{l,\cdot} \in \mathcal{C} \Big) - \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^\star_{U,l}) \{\mathcal{C}\} \right| = o(1)$$

Distributional theory for $\boldsymbol{U}$

Theorem 5 (Yan, Chen, Fan '21)

Consider any $1 \leq l \leq d$ s.t. $\|\boldsymbol{U}^\star_{l,\cdot}\|_2$ is not too small. Under previous assumptions, we have

$$\sup_{\text{cvx set } \mathcal{C}} \left| \mathbb{P}\Big(\Big[\boldsymbol{U} \underbrace{\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star)}_{\textit{global rotation}} - \boldsymbol{U}^\star \Big]_{l,\cdot} \in \mathcal{C} \Big) - \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^\star_{U,l})\left\{\mathcal{C}\right\} \right| = o\left(1\right)$$

- Each row of $\boldsymbol{U}$ is approximately Gaussian
  — nearly unbiased + tractable covariance

Distributional theory for U

Theorem 5 (Yan, Chen, Fan '21)

Consider any $1 \le l \le d$ s.t. $\|U^\star_{l,\cdot}\|_2$ is not too small. Under previous assumptions, we have

$$\sup_{\text{cvx set } \mathcal{C}} \left| \mathbb{P}\Big(\Big[U \underbrace{\mathrm{sgn}(U^\top U^\star)}_{\text{global rotation}} - U^\star \Big]_{l,\cdot} \in \mathcal{C} \Big) - \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^\star_{U,l}) \{\mathcal{C}\} \right| = o(1)$$

- Each row of U is approximately Gaussian

— nearly unbiased + tractable covariance

$$\boldsymbol{\Sigma}^\star_{U,l} := \left(\frac{1-p}{np} S^\star_{l,l} + \frac{\omega_l^{\star 2}}{np} \right) (\boldsymbol{\Lambda}^\star)^{-1} + \frac{2(1-p)}{np} U^{\star\top}_{l,\cdot} U^\star_{l,\cdot}$$
$$+ (\boldsymbol{\Lambda}^\star)^{-1} U^{\star\top} \mathrm{diag}\left\{ [d^\star_{l,i}]_{1\le i\le d} \right\} U^\star (\boldsymbol{\Lambda}^\star)^{-1}$$
$$d^\star_{l,i} := \frac{1}{np^2} \left[\omega_l^{\star 2} + (1-p) S^{\star 2}_{l,l} \right] \left[\omega_i^{\star 2} + (1-p) S^{\star 2}_{i,i} \right] + \frac{2(1-p)^2}{np^2} S^{\star 2}_{l,i}$$

Distributional theory for U

Theorem 5 (Yan, Chen, Fan '21)

Consider any $1 \le l \le d$ *s.t.* $\|\boldsymbol{U}^{\star}_{l,\cdot}\|_2$ *is not too small. Under previous assumptions, we have*

$$\sup_{\text{cvx set } \mathcal{C}} \left| \mathbb{P}\Big(\Big[\boldsymbol{U} \underbrace{\mathsf{sgn}(\boldsymbol{U}^{\top}\boldsymbol{U}^{\star})}_{\textit{global rotation}} - \boldsymbol{U}^{\star} \Big]_{l,\cdot} \in \mathcal{C} \Big) - \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\star}_{U,l}) \{\mathcal{C}\} \right| = o(1)$$

- **Key observations:**

$$\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^{\top}\boldsymbol{U}^{\star}) - \boldsymbol{U}^{\star} \approx \Big[\underbrace{\boldsymbol{E}\boldsymbol{X}^{\top}}_{\text{linear term}} + \underbrace{\mathcal{P}_{\text{off-diag}}(\boldsymbol{E}\boldsymbol{E}^{\top})}_{\text{quadratic term}} \Big] \boldsymbol{U}^{\star} (\boldsymbol{\Lambda}^{\star})^{-1}$$

Distributional theory for S

Theorem 6 (Yan, Chen, Fan '21)

Consider any (i,j) s.t. $\|\boldsymbol{U}^{\star}_{i,\cdot}\|_2$ and $\|\boldsymbol{U}^{\star}_{j,\cdot}\|_2$ are not too small. Under previous assumptions, we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_{i,j}-S^{\star}_{i,j}}{\sqrt{v^{\star}_{i,j}}}\leq t\right)-\Phi\left(t\right)\right|=o\left(1\right)$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$

Distributional theory for S

Theorem 6 (Yan, Chen, Fan '21)

Consider any (i, j) s.t. $\|\boldsymbol{U}^{\star}_{i,\cdot}\|_2$ and $\|\boldsymbol{U}^{\star}_{j,\cdot}\|_2$ are not too small. Under previous assumptions, we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_{i,j}-S^{\star}_{i,j}}{\sqrt{v^{\star}_{i,j}}}\leq t\right)-\Phi\left(t\right)\right|=o\left(1\right)$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$

- Each entry of $\boldsymbol{S}$ is approximately Gaussian
  — nearly unbiased + tractable variance

Distributional theory for S

Theorem 6 (Yan, Chen, Fan '21)

Consider any (i, j) s.t. $\|\boldsymbol{U}^{\star}_{i,\cdot}\|_2$ and $\|\boldsymbol{U}^{\star}_{j,\cdot}\|_2$ are not too small. Under previous assumptions, we have

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_{i,j} - S^{\star}_{i,j}}{\sqrt{v^{\star}_{i,j}}} \leq t \right) - \Phi\left(t\right) \right| = o\left(1\right)$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$

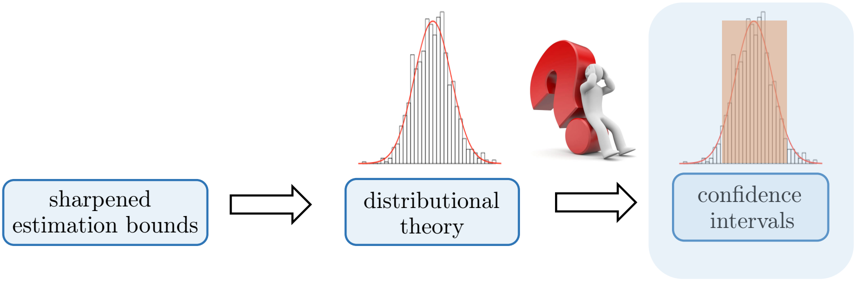
- Each entry of $\boldsymbol{S}$ is approximately Gaussian
  — nearly unbiased + tractable variance

For any $i \neq j$:

$$\begin{aligned} v^{\star}_{i,j} \coloneqq & \frac{2-p}{np} S^{\star}_{i,i} S^{\star}_{j,j} + \frac{4-3p}{np} S^{\star 2}_{i,j} + \frac{1}{np}\left(\omega_i^{\star 2} S^{\star}_{j,j} + \omega_j^{\star 2} S^{\star}_{i,i}\right) \\ & + \frac{2\left(1-p\right)^2}{np^2} \left[\sum_{k=1}^{d} S^{\star 2}_{i,k} \left(\boldsymbol{U}^{\star}_{k,\cdot} \boldsymbol{U}^{\star\top}_{j,\cdot}\right)^2 + S^{\star 2}_{j,k} \left(\boldsymbol{U}^{\star}_{k,\cdot} \boldsymbol{U}^{\star\top}_{i,\cdot}\right)^2 \right] \\ & + \frac{1}{np^2} \sum_{k=1}^{d} \left[\omega_i^{\star 2} + \left(1-p\right) S^{\star}_{i,i}\right] \left[\omega_k^{\star 2} + \left(1-p\right) S^{\star}_{k,k}\right] \left(\boldsymbol{U}^{\star}_{k,\cdot} \boldsymbol{U}^{\star\top}_{j,\cdot}\right)^2 \\ & + \frac{1}{np^2} \sum_{k=1}^{d} \left[\omega_j^{\star 2} + \left(1-p\right) S^{\star}_{j,j}\right] \left[\omega_k^{\star 2} + \left(1-p\right) S^{\star}_{k,k}\right] \left(\boldsymbol{U}^{\star}_{k,\cdot} \boldsymbol{U}^{\star\top}_{i,\cdot}\right)^2 \end{aligned}$$

For any $1 \leq i \leq d$:

$$\begin{aligned} v^{\star}_{i,i} \coloneqq & \frac{12-9p}{np} S^{\star 2}_{i,i} + \frac{4}{np} \omega_i^{\star 2} S^{\star}_{i,i} + \frac{8\left(1-p\right)^2}{np^2} \sum_{k=1}^{d} S^{\star 2}_{i,k} \left(\boldsymbol{U}^{\star}_{k,\cdot} \boldsymbol{U}^{\star\top}_{i,\cdot}\right)^2 \\ & + \frac{4}{np^2} \sum_{k=1}^{d} \left[\omega_i^{\star 2} + \left(1-p\right) S^{\star}_{i,i}\right] \left[\omega_k^{\star 2} + \left(1-p\right) S^{\star}_{k,k}\right] \left(\boldsymbol{U}^{\star}_{k,\cdot} \boldsymbol{U}^{\star\top}_{i,\cdot}\right)^2 \end{aligned}$$



How to compute confidence intervals in a data-driven manner (e.g., without prior knowledge of noise levels)?

Estimating unknown model parameters

- Compute estimate $(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{S})$ for $\big(\boldsymbol{U}^{\star}, \boldsymbol{\Lambda}^{\star}, \boldsymbol{S}^{\star}\big)$ via HeteroPCA

[1] $\{y_{i,j} : (i,j) \in \Omega\}$ are zero-mean r.v.s with common variance $S^{\star}_{i,i} + \omega_i^{\star 2}$

Estimating unknown model parameters

- Compute estimate $(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{S})$ for $\left(\boldsymbol{U}^\star, \boldsymbol{\Lambda}^\star, \boldsymbol{S}^\star\right)$ via HeteroPCA

- Estimate noise variances $\{\omega_i^{\star 2}\}_{i=1}^d$ via[1]

$$\omega_i^2 := \frac{\sum_{j=1}^n y_{i,j}^2 \mathbb{1}_{(i,j)\in\Omega}}{\sum_{j=1}^n \mathbb{1}_{(i,j)\in\Omega}} - S_{i,i}$$

[1] $\{y_{i,j} : (i,j) \in \Omega\}$ are zero-mean r.v.s with common variance $S_{i,i}^\star + \omega_i^{\star 2}$

Estimating unknown model parameters

- Compute estimate $(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{S})$ for $\big(\boldsymbol{U}^{\star}, \boldsymbol{\Lambda}^{\star}, \boldsymbol{S}^{\star}\big)$ via HeteroPCA

- Estimate noise variances $\{\omega_i^{\star 2}\}_{i=1}^{d}$ via[1]

$$\omega_i^2 := \frac{\sum_{j=1}^{n} y_{i,j}^2 \mathbb{1}_{(i,j)\in\Omega}}{\sum_{j=1}^{n} \mathbb{1}_{(i,j)\in\Omega}} - S_{i,i}$$

- Compute "plug-in" estimate $v_{i,j}$ for $v_{i,j}^{\star}$

[1] $\{y_{i,j} : (i,j) \in \Omega\}$ are zero-mean r.v.s with common variance $S_{i,i}^{\star} + \omega_i^{\star 2}$

Entrywise confidence intervals for $S^\star$

For any target coverage level $1-\alpha$ and each (i,j), compute

$$\mathsf{CI}_{i,j}^{1-\alpha} := \underbrace{\left[S_{i,j} \pm \Phi^{-1}(1-\alpha/2)\sqrt{v_{i,j}}\right]}_{\text{since } S_{i,j} \approx \mathcal{N}(S^\star_{i,j}, v^\star_{i,j}) \approx \mathcal{N}(S^\star_{i,j}, v_{i,j})}$$

Entrywise confidence intervals for $S^\star$

For any target coverage level $1-\alpha$ and each (i,j), compute

$$\mathsf{CI}_{i,j}^{1-\alpha} := \underbrace{\left[S_{i,j} \pm \Phi^{-1}\left(1-\alpha/2\right)\sqrt{v_{i,j}}\right]}_{\text{since } S_{i,j} \approx \mathcal{N}(S_{i,j}^\star, v_{i,j}^\star) \approx \mathcal{N}(S_{i,j}^\star, v_{i,j})}$$

Theorem 7 (Yan, Chen, Fan '21)

Suppose previous conditions hold and $\frac{\omega_{\max}}{\omega_{\min}} = O(1)$. *Then we have*

$$\mathbb{P}\left(S_{i,j}^\star \in \mathsf{CI}_{i,j}^{1-\alpha}\right) = 1-\alpha+o\left(1\right)$$

Entrywise confidence intervals for $S^\star$

For any target coverage level $1-\alpha$ and each (i,j), compute

$$\mathsf{CI}_{i,j}^{1-\alpha} := \underbrace{\left[S_{i,j} \pm \Phi^{-1}(1-\alpha/2)\sqrt{v_{i,j}}\right]}_{\text{since } S_{i,j} \approx \mathcal{N}(S_{i,j}^\star, v_{i,j}^\star) \approx \mathcal{N}(S_{i,j}^\star, v_{i,j})}$$

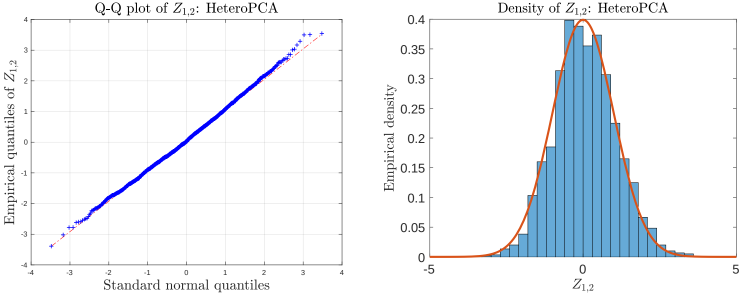
Theorem 7 (Yan, Chen, Fan '21)

Suppose previous conditions hold and $\frac{\omega_{\max}}{\omega_{\min}} = O(1)$. *Then we have*

$$\mathbb{P}\left(S_{i,j}^\star \in \mathsf{CI}_{i,j}^{1-\alpha}\right) = 1 - \alpha + o(1)$$

- adaptive to unknown noise levels
- adaptive to noise heteroskedasticity

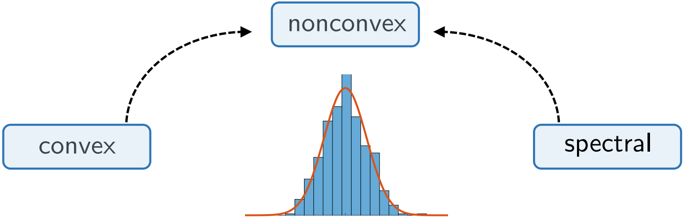
Numerical verification



$$n = 2000,\ d = 100,\ p = 0.6,\ r = 3,\ \omega_1^\star, \cdots, \omega_d^\star \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[0.025, 0.1],$$

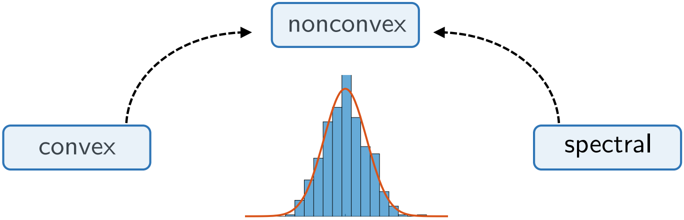
$$Z_{1,2} = \frac{S_{1,2} - S_{1,2}^\star}{\sqrt{v_{1,2}}}$$

Concluding remarks



uncertainty quantification for low-rank models

Concluding remarks



uncertainty quantification for low-rank models

- improve dependency on rank & cond. number
- general sampling patterns
- small eigen-gaps
- inference for functionals of eigenvectors

Papers:

"Inference and uncertainty quantification for noisy matrix completion," *Proceedings of the National Academy of Sciences (PNAS)*, Y. Chen, J. Fan, C. Ma, Y. Yan, vol. 116, no. 46, pp. 22931-22937, 2019

"Noisy matrix completion: understanding statistical guarantees for convex relaxation via nonconvex optimization," *SIAM Journal on Optimization*, Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, 2020

"Inference for Heteroskedastic PCA with Missing Data," Y. Yan, Y. Chen, J. Fan, arxiv:2107.12365, 2021

C. Cai, G. Li, Y. Chi, H. V. Poor, Y. Chen, "Subspace Estimation from Unbalanced and Incomplete Data Matrices: $\ell_{2,\infty}$ Statistical Guarantees," *Annals of Statistics*, 2021