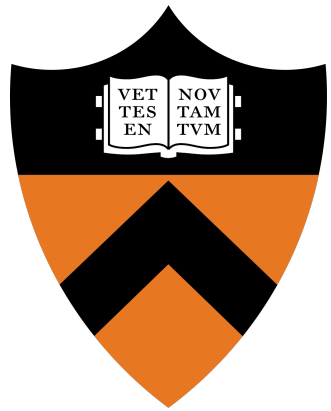
Inference and Uncertainty Quantification for Low-Rank Models



Yuling Yan

Princeton University

A ubiquitous low-complexity model

Composition C by Piet Mondrian

reconstructing low-rank structure
from imperfect measurements

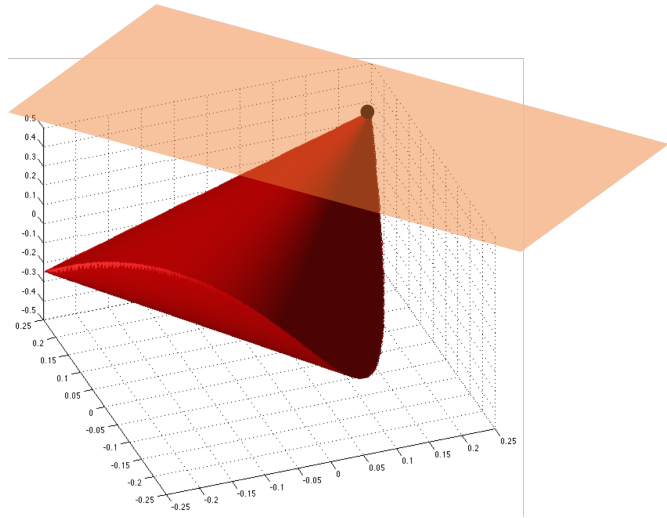
A ubiquitous low-complexity model

Composition C by Piet Mondrian

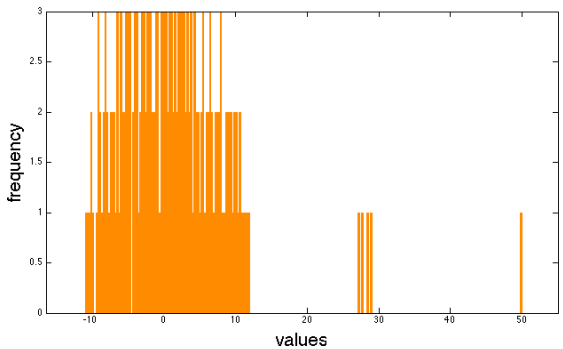
reconstructing low-rank structure
from imperfect measurements

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

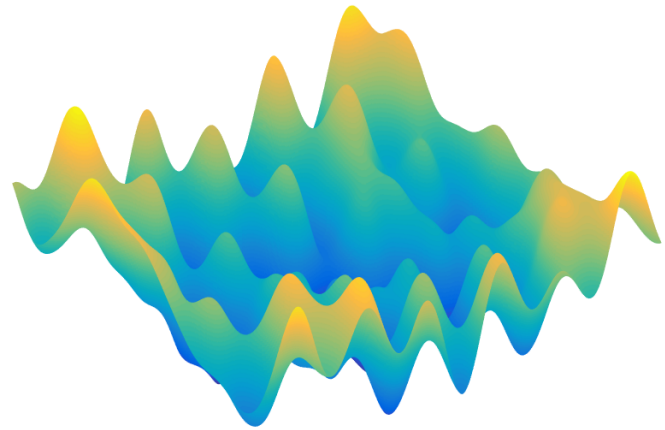
Various estimation schemes have been proposed



convex relaxation

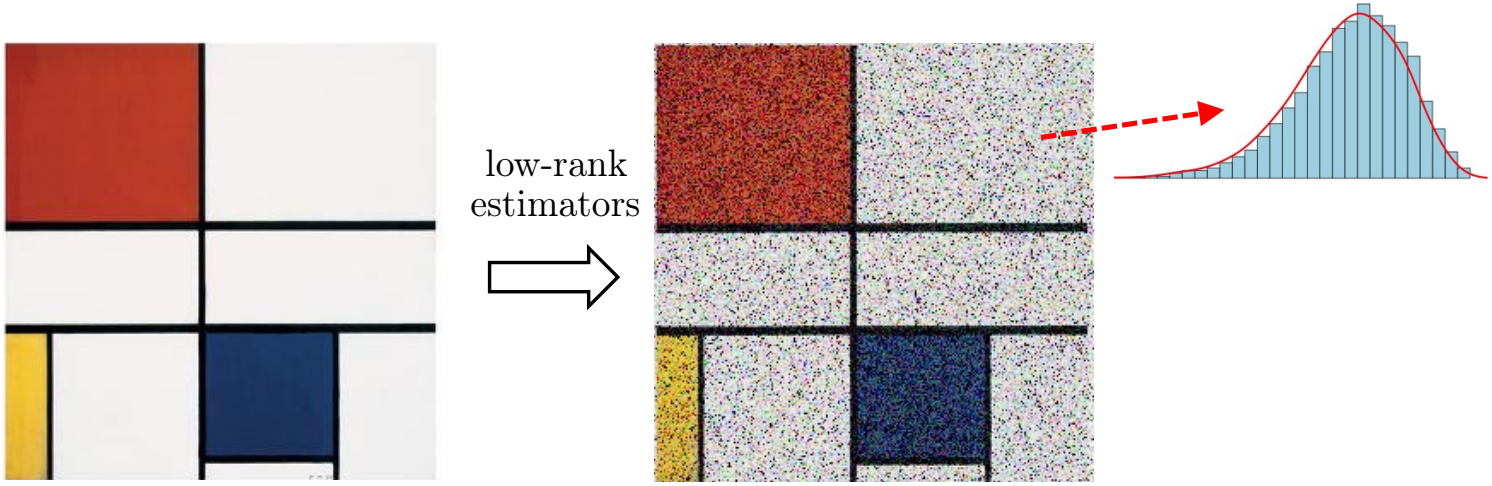


spectral methods

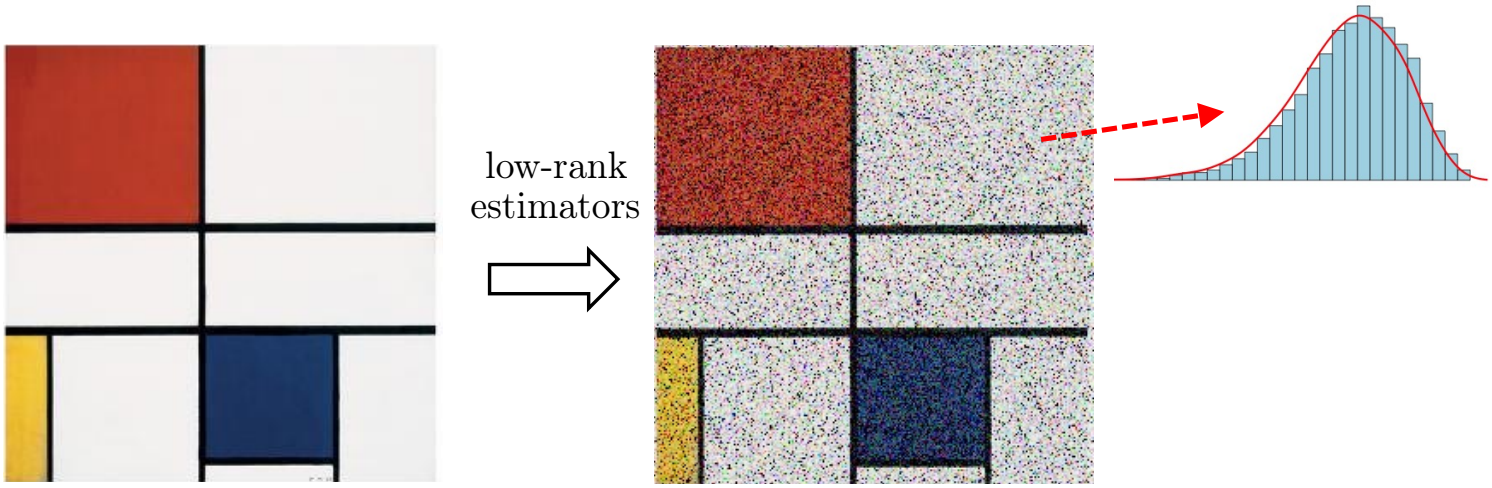


nonconvex optimization

One step further: reasoning about uncertainty?

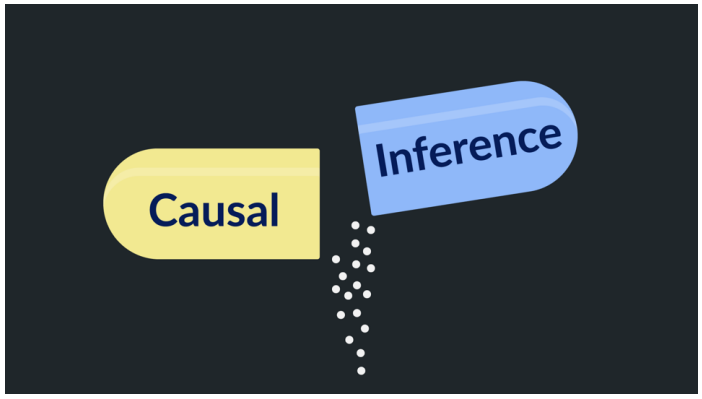


One step further: reasoning about uncertainty?

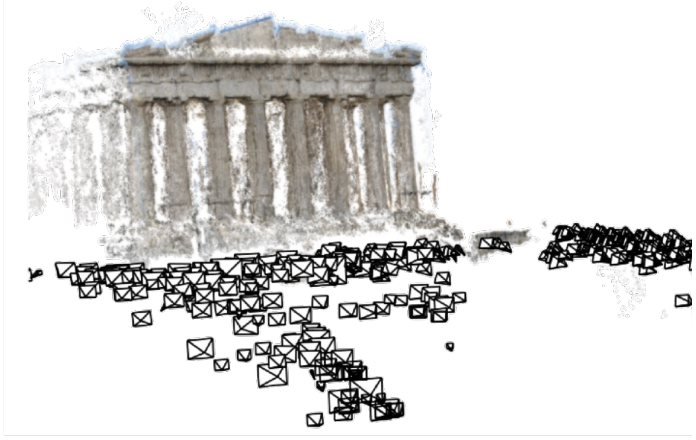


How to assess uncertainty, or "confidence", of obtained low-rank estimates due to imperfect data acquisition?

- noise
- missing data
- outliers
- ...

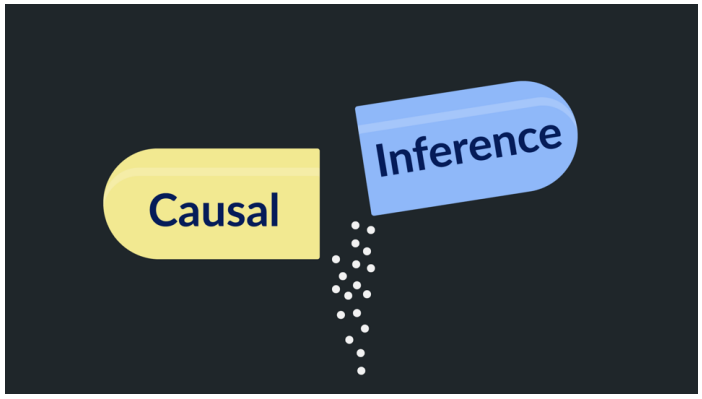


causal inference with panel data

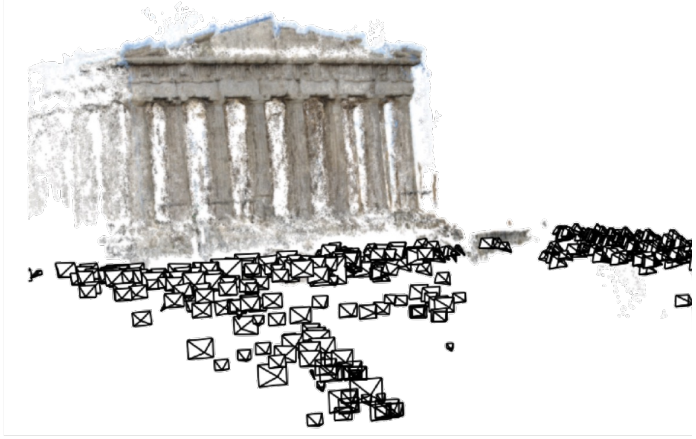


structure from motion

anomaly detection



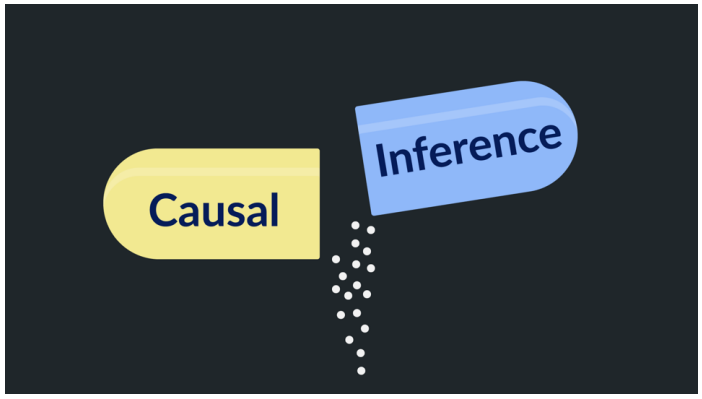
causal inference with panel data



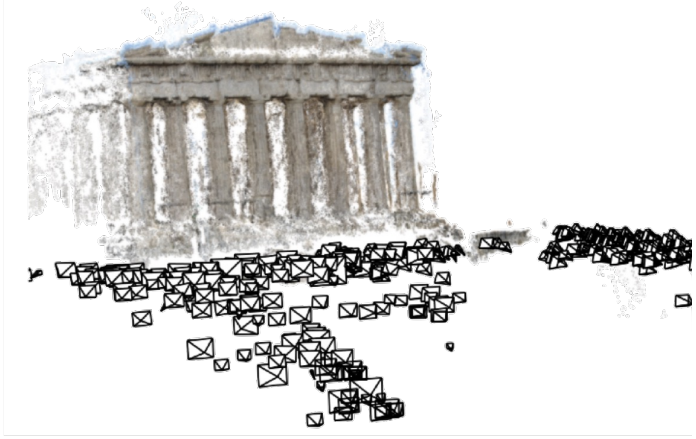
structure from motion

anomaly detection

"Matrix completion methods for causal panel data models," S. Athey, M. Bayati, N. Doudchenko, G. Imbens, K. Khosravi *Journal of the American Statistical Association*



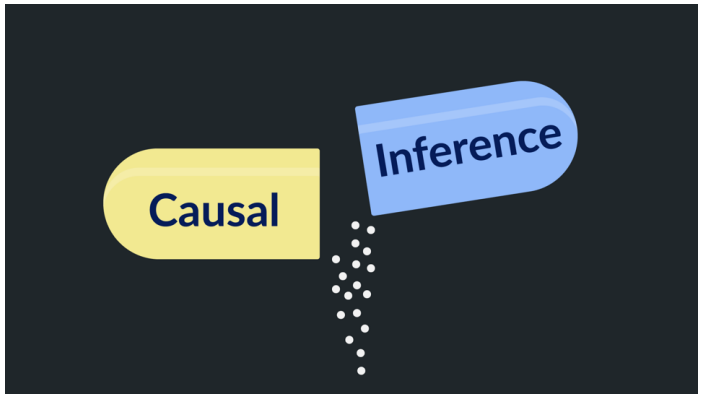
causal inference with panel data



structure from motion

anomaly detection

"A Closed-Form Uncertainty Propagation in Non-Rigid Structure From Motion," J. Song, M. Patel, A. Jasour, M. Ghaffari, *IEEE Robotics and Automation Letters*



causal inference with panel data

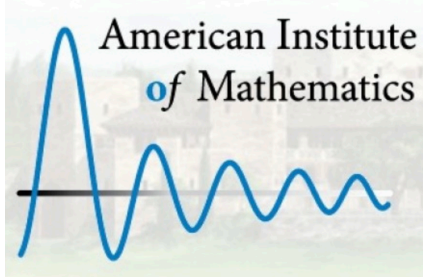
structure from motion

anomaly detection

"Near-optimal entrywise anomaly detection for low-rank matrices with sub-exponential noise," V. Farias, A. Li, A. Jasour, T. Peng, *ICML*

INFERENCE IN HIGH DIMENSIONAL REGRESSION

organized by
Peter Buehlmann, Andrea Montanari, and Jonathan Taylor



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

This talk: two recent examples

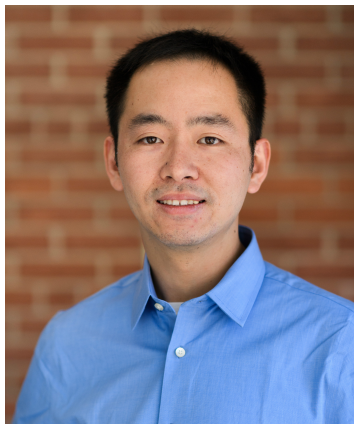
1. Inference for noisy matrix completion

   — convex optimization, nonconvex optimization

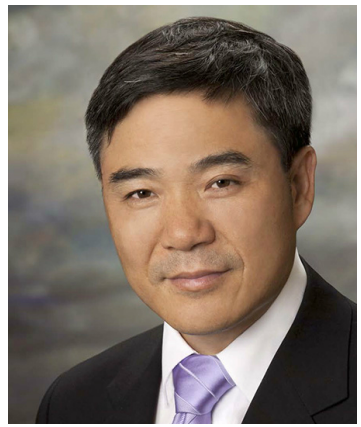
2. Inference for heteroskedastic PCA with missing data

   — spectral methods

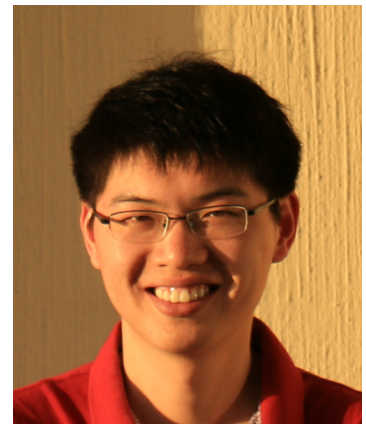
Part 1: Inference for noisy matrix completion



Yuxin Chen
UPenn Wharton



Jianqing Fan
Princeton ORFE



Cong Ma
U Chicago Stats

Noisy low-rank matrix completion

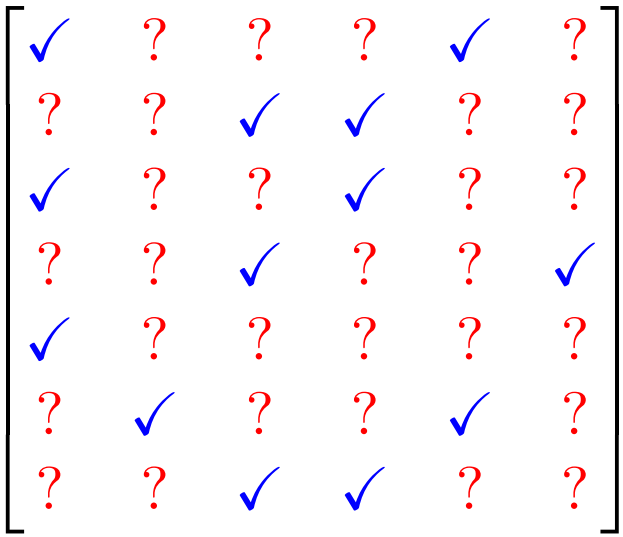


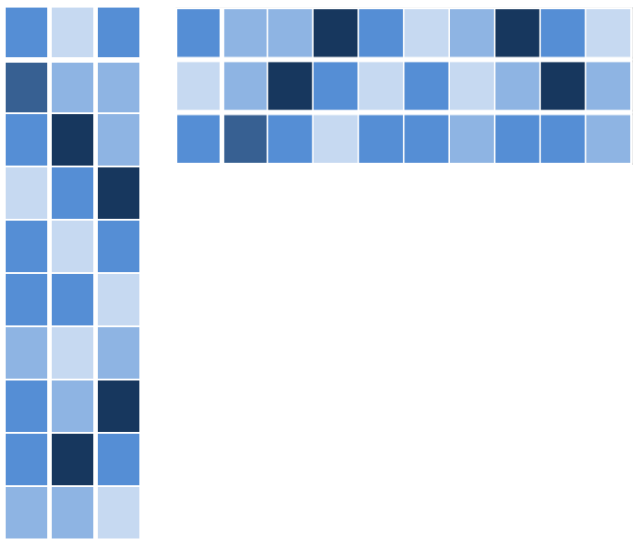
figure credit: E. Candès, M. Soltanolkotabi

Given partial noisy entries of a low-rank matrix $M^\star$, fill in missing entries

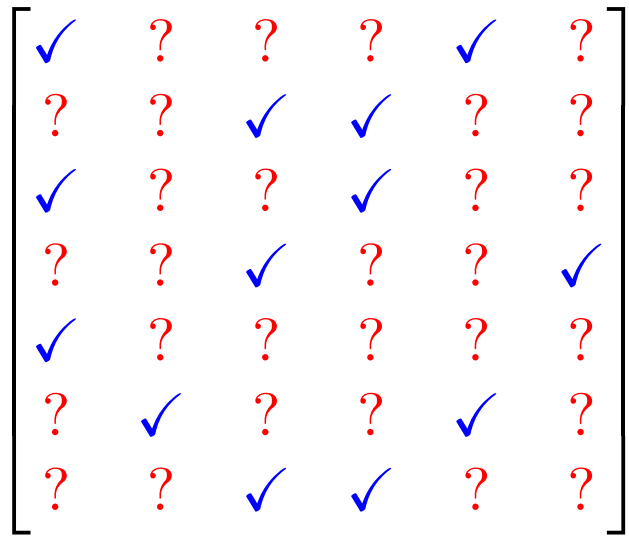
Noisy low-rank matrix completion

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

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



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



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 natural least squares formulation:

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

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:

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

Challenges

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

- convex estimate $\boldsymbol{M}^{\text{cvx}}$ is biased

Challenges

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

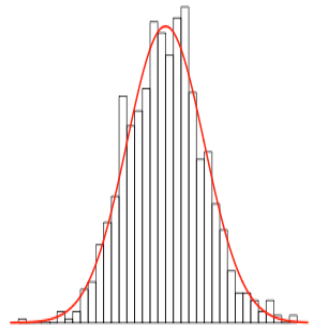
- convex estimate $\boldsymbol{M}^{\text{cvx}}$ is biased
- highly challenging to pin down distributions of obtained estimates

Challenges

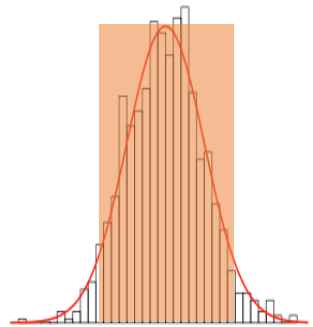
$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z}} \underbrace{\sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2}_{\text{empirical loss}} + \lambda \|\boldsymbol{Z}\|_*$$

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

sharpened estimation bounds

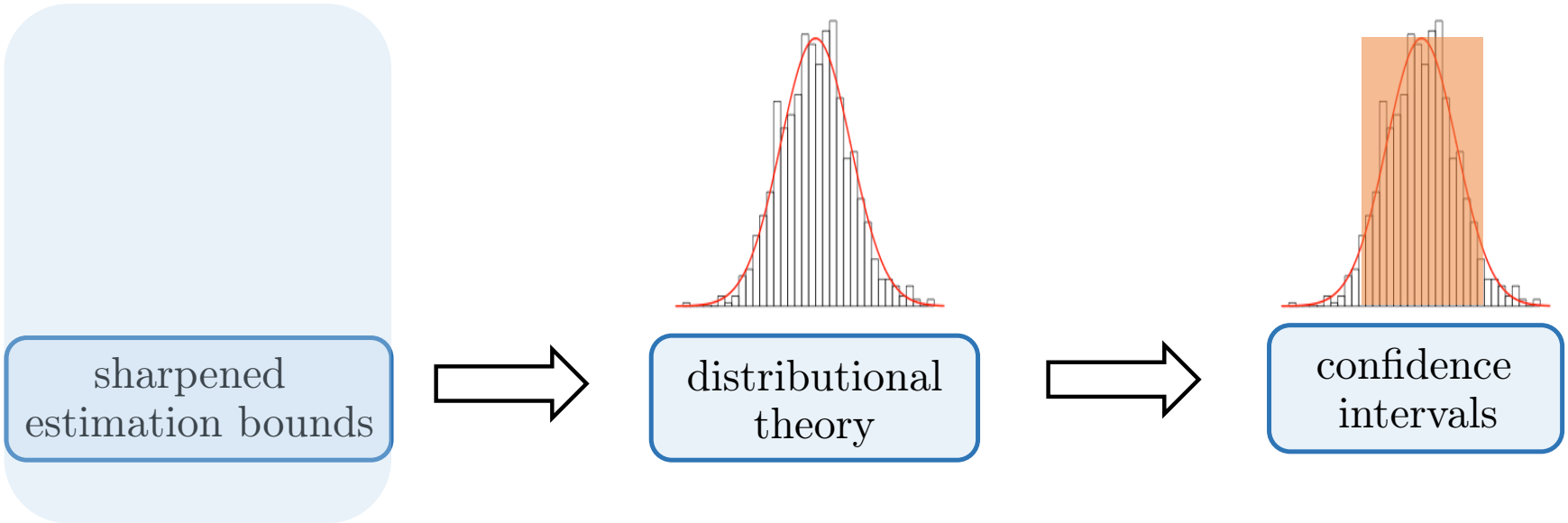


distributional theory



confidence intervals

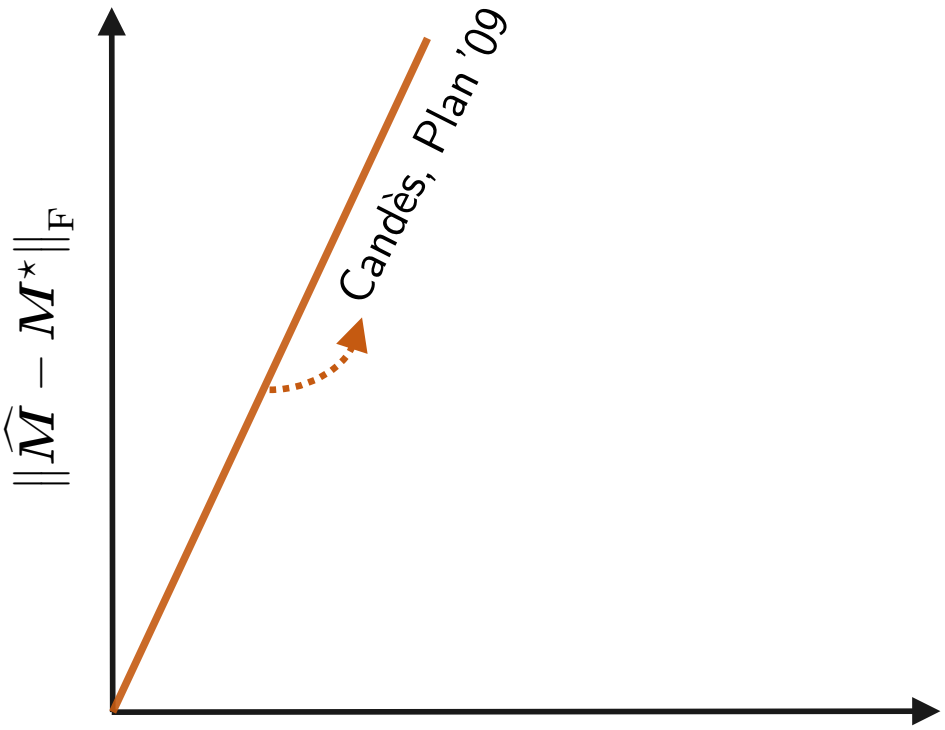
Step 1: sharpening estimation guarantees



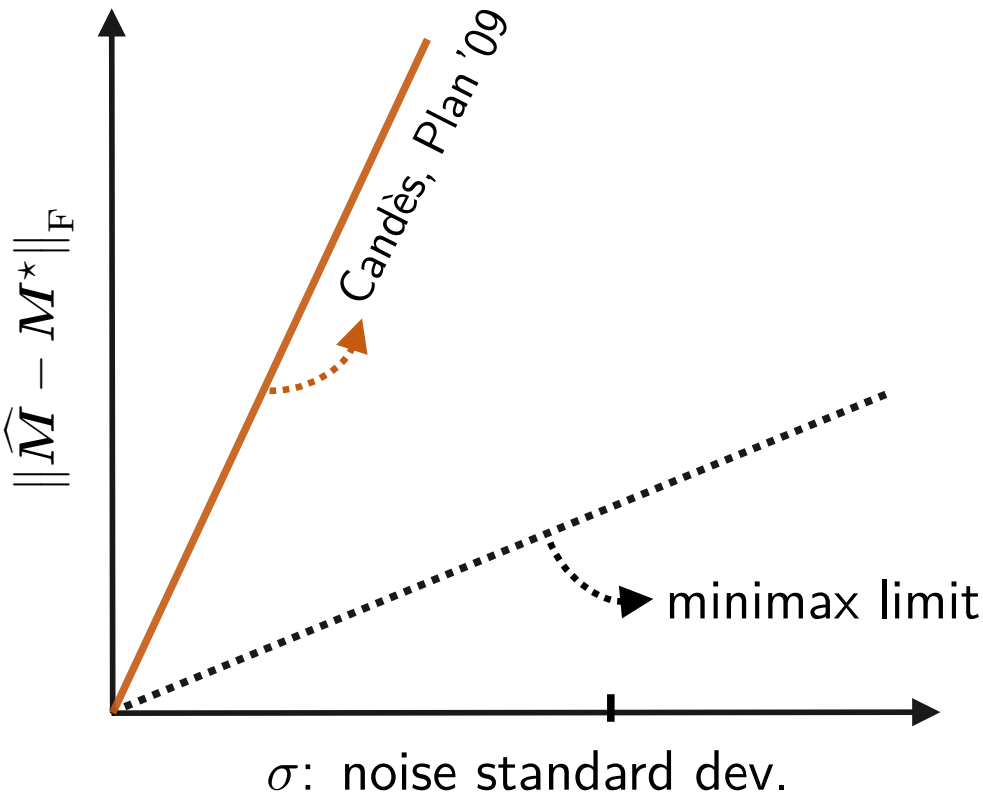
Prior statistical guarantees for convex relaxation

- **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 $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, well-conditioned

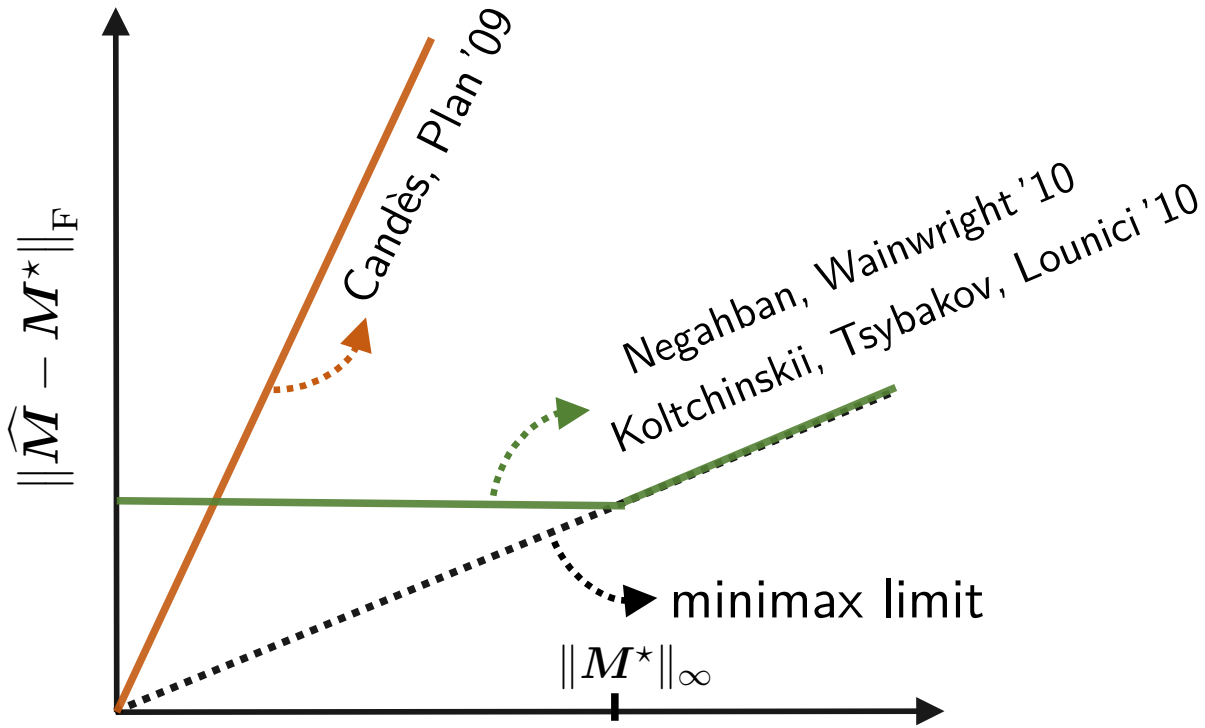
Candès, Plan '09	$\sigma n^{1.5}$



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

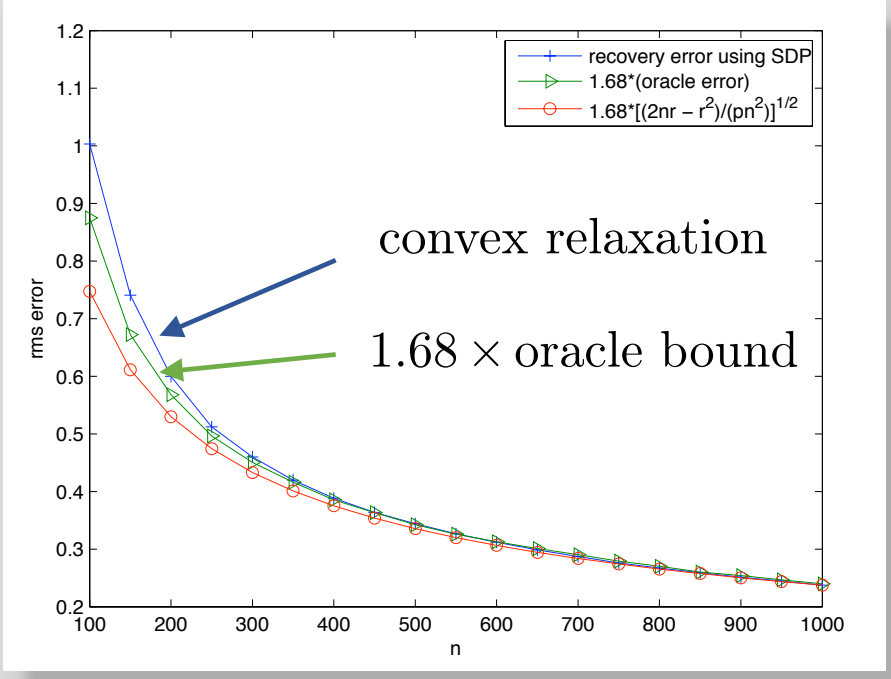


minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$
Koltchinskii, Tsybakov, Lounici '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$



Matrix Completion with Noise

Emmanuel J. Candès and Yaniv Plan



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

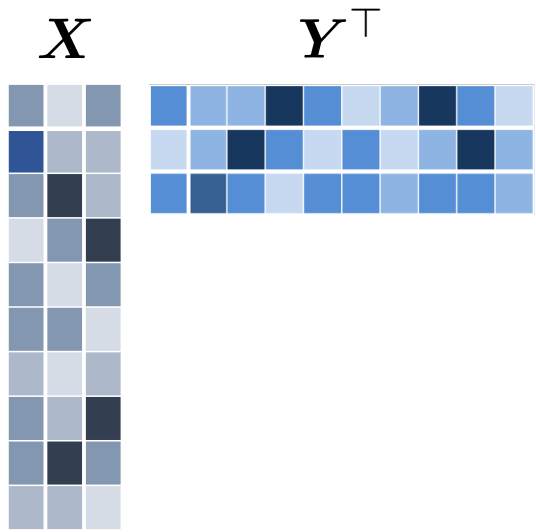
dual certification (golfing scheme)

dual certification (golfing scheme)

nonconvex optimization

A detour: nonconvex optimization

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



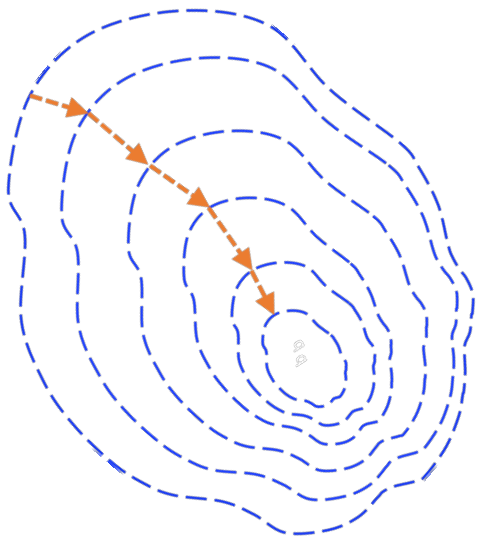
$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \underbrace{\sum_{(i,j)\in\Omega} \left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{i,j} - M_{i,j}\right]^2}_{\text{squared loss}} + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

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

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



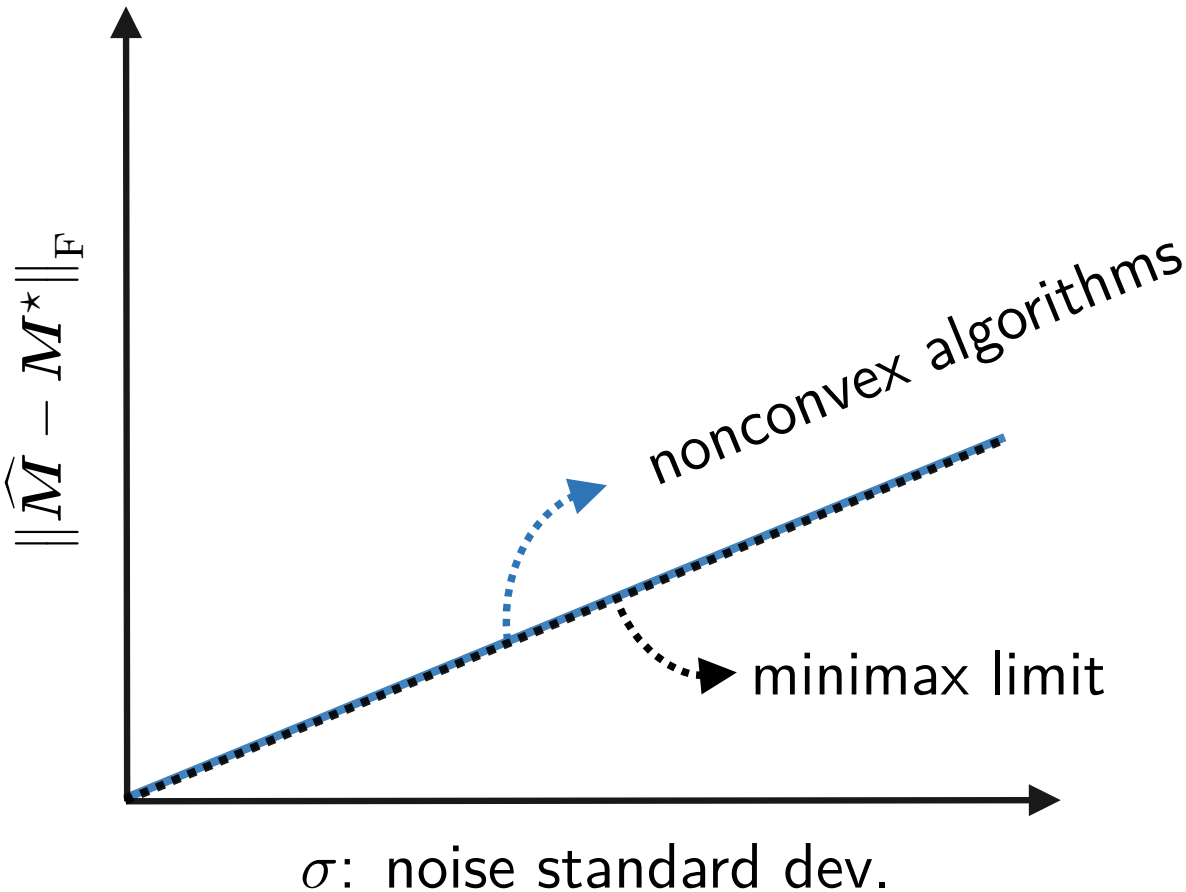
- **suitable initialization:** $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$
- **gradient descent:** for $t = 0, 1, \ldots$

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

A detour: nonconvex optimization

—*Ma, Wang, Chi, Chen '20*

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



A motivating experiment

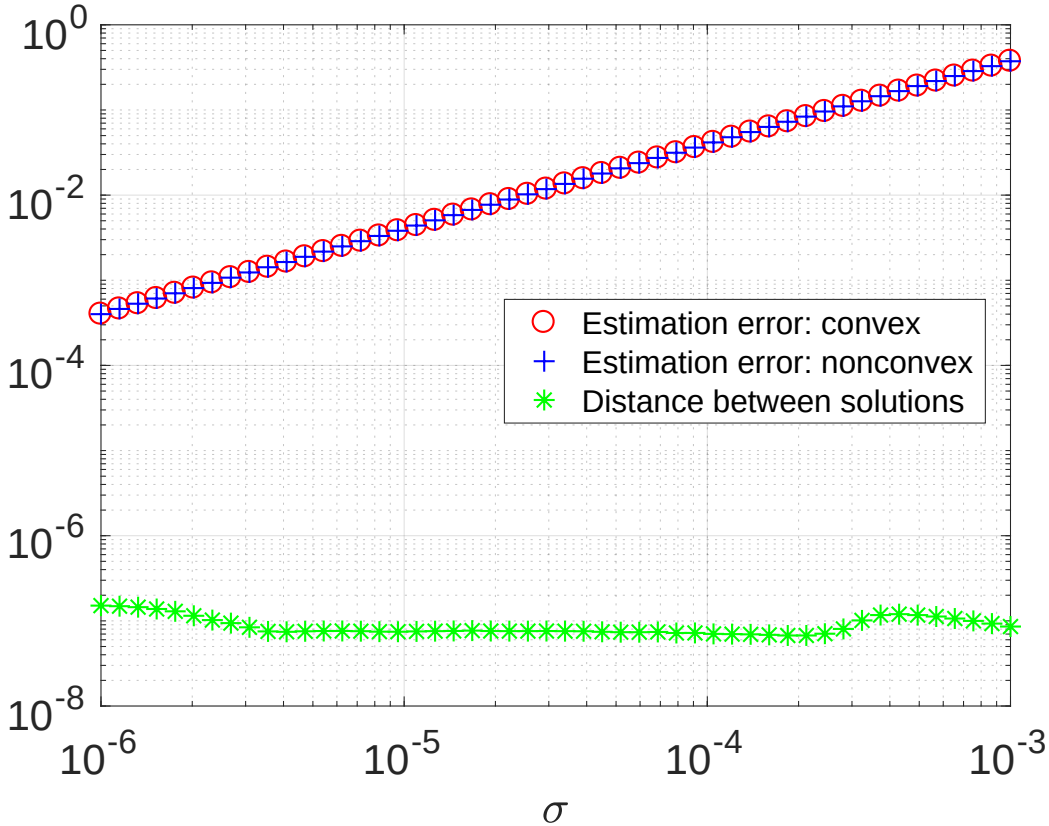
convex: $\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_*$

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

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

A motivating experiment

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



Convex and nonconvex solutions are exceedingly close!

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

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

convex

nonconvex

$$\Downarrow$$

$$\text{stability}\Big(\text{convex}\Big) \approx \text{stability}\Big(\text{nonconvex}\Big)$$

Sharpened estimation guarantees

Assumptions (omitting log factors)

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{1}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 satisfying $\sigma \lesssim \sqrt{np}\|\boldsymbol{M}^\star\|_\infty$
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

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

Sharpened estimation guarantees

Assumptions (omitting log factors)

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{1}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 satisfying $\sigma \lesssim \sqrt{np}\|\boldsymbol{M}^\star\|_\infty$
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

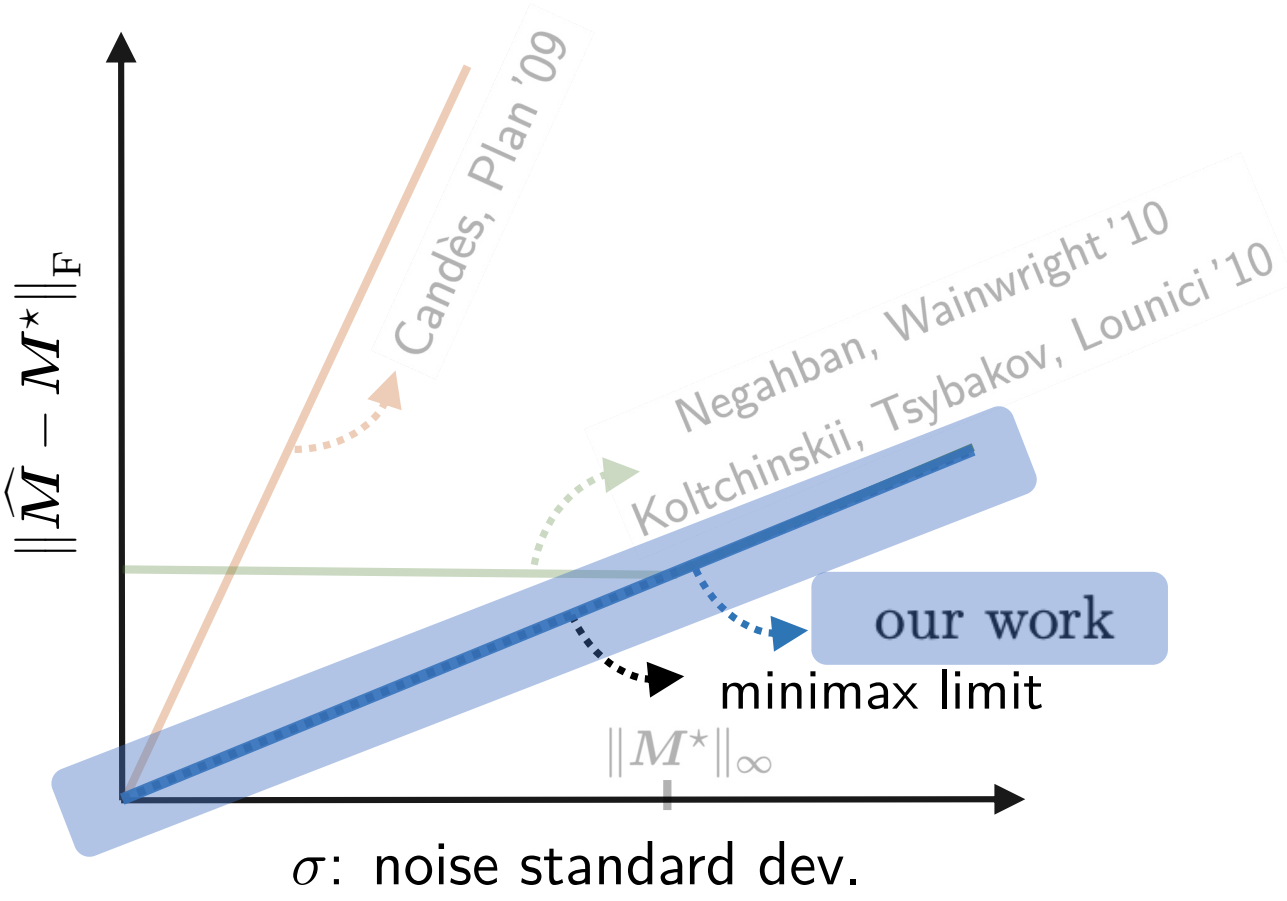
$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Theorem 1

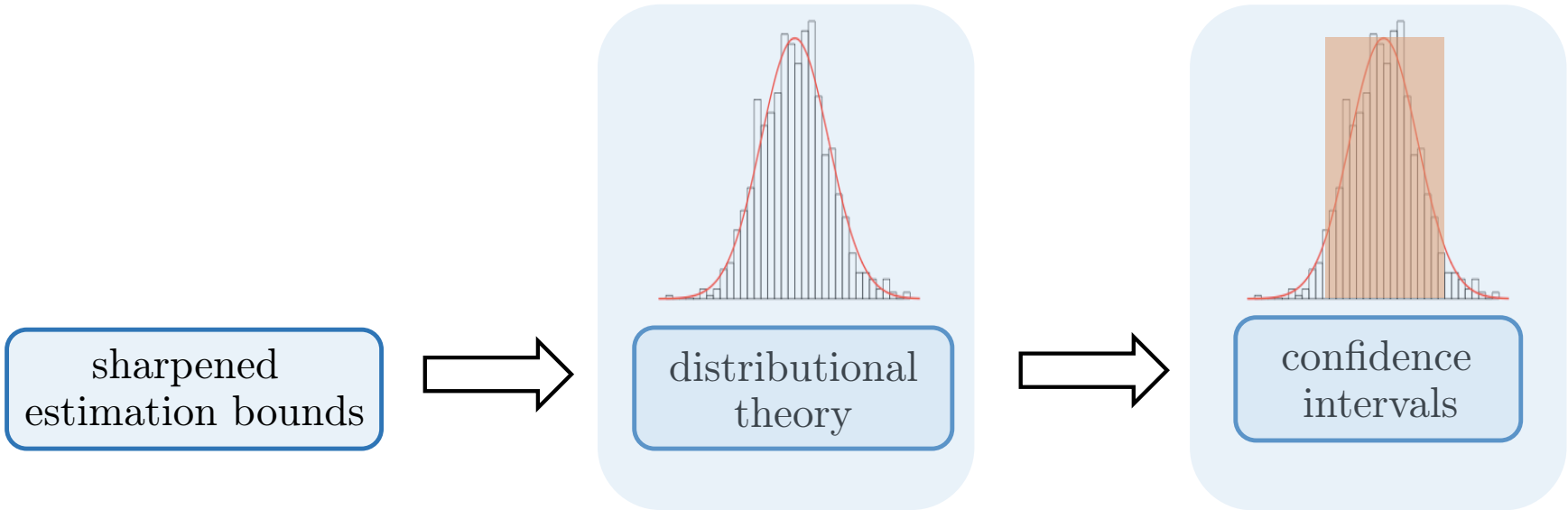
With high prob., any minimizer $\boldsymbol{M}^{\text{cvx}}$ of convex program obeys

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

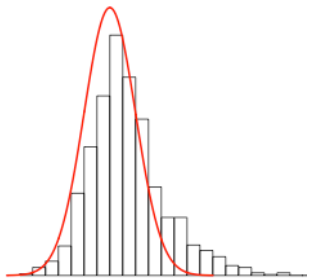
$$\left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^{\star}\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{\tfrac{n}{p}}: \quad \text{minimax optimal}$$



Step 2: from estimation to inference . . .



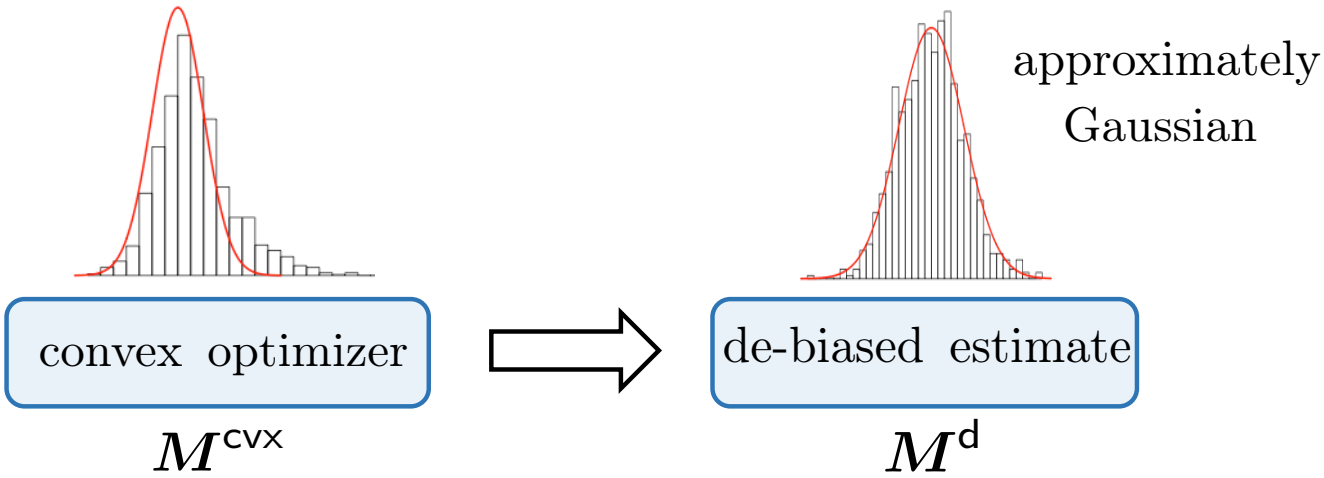
— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



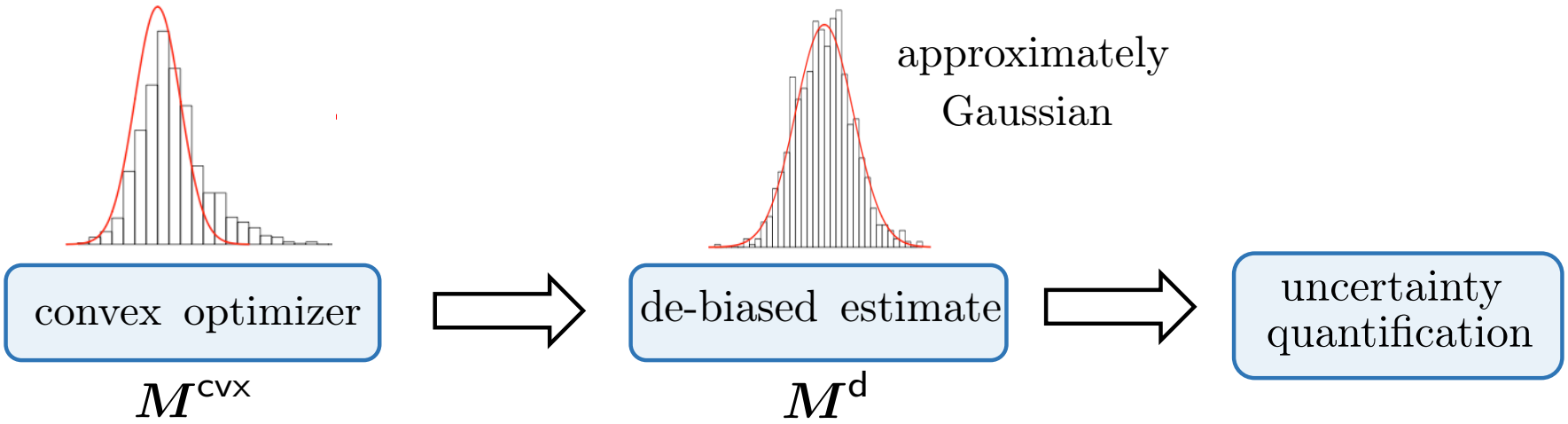
convex optimizer

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

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

De-biasing convex estimate

$$\boldsymbol{M}^{\mathsf{CVX}} \quad \overset{\text{de-biasing}}{\longrightarrow} \quad \boldsymbol{M}^{\mathsf{CVX}} + \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E}) - \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\mathsf{CVX}})$$

De-biasing convex estimate

$$\boldsymbol{M}^{\mathsf{CVX}} \quad \overset{\text{de-biasing}}{\longrightarrow} \quad \boldsymbol{M}^{\mathsf{CVX}} + \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E}) - \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\mathsf{CVX}})$$

$\mathcal{I}$ $\mathbf{0}$ $\mathcal{I}$

De-biasing convex estimate

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

($\mathcal{I}$ over $\frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E})$; $\mathbf{0}$ over $\boldsymbol{E}$; $\mathcal{I}$ over $\frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\mathsf{cvx}})$)

De-biasing convex estimate

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

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

De-biasing convex estimate

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

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

Distributional theory

$$\boldsymbol{M}^\star \xrightarrow{\text{rank-}r\text{ svd}} \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$$

Theorem 2

Consider any (i,j) s.t. $\|\boldsymbol{U}^\star_{i,\cdot}\|_2 + \|\boldsymbol{V}^\star_{j,\cdot}\|_2$ is not too small. Then

$$M^{\mathsf{d}}_{i,j} - M^\star_{i,j} \sim \mathcal{N}\Big(0, \frac{\sigma^2}{p}\big(\|\boldsymbol{U}^\star_{i,\cdot}\|_2^2 + \|\boldsymbol{V}^\star_{j,\cdot}\|_2^2\big)\Big) + \textit{negligible term}$$

Distributional theory

$$\boldsymbol{M}^\star \xrightarrow{\text{rank-}r\text{ svd}} \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$$

Theorem 2

Consider any (i,j) s.t. $\|\boldsymbol{U}^\star_{i,\cdot}\|_2 + \|\boldsymbol{V}^\star_{j,\cdot}\|_2$ is not too small. Then

$$M^{\mathsf{d}}_{i,j} - M^\star_{i,j} \sim \mathcal{N}\Big(0, \frac{\sigma^2}{p}\big(\|\boldsymbol{U}^\star_{i,\cdot}\|_2^2 + \|\boldsymbol{V}^\star_{j,\cdot}\|_2^2\big)\Big) + \textit{negligible term}$$

- key decomposition:

$$M^{\mathsf{d}}_{i,j} - M^\star_{i,j} \approx \underbrace{\frac{\sigma}{\sqrt{p}}(\boldsymbol{V}^\star_{j,\cdot}\boldsymbol{z} + \boldsymbol{U}^\star_{i,\cdot}\boldsymbol{w})}_{\text{first-order Gaussian term}} + \underbrace{\frac{\sigma^2}{p}\boldsymbol{z}^\top \boldsymbol{\Sigma}^\star \boldsymbol{w}}_{\text{second-order term}}$$

where $\boldsymbol{z}, \boldsymbol{w} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r)$

Distributional theory

$$\boldsymbol{M}^\star \xrightarrow{\text{rank-}r\text{ svd}} \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \xrightarrow{\text{bal. decom.}} \boldsymbol{X}^\star \boldsymbol{Y}^{\star\top}$$

Theorem 2

Consider any (i,j) s.t. $\|\boldsymbol{U}^\star_{i,\cdot}\|_2 + \|\boldsymbol{V}^\star_{j,\cdot}\|_2$ is not too small. Then

$$M^{\mathsf{d}}_{i,j} - M^\star_{i,j} \sim \mathcal{N}\Big(0, \frac{\sigma^2}{p}\big(\|\boldsymbol{U}^\star_{i,\cdot}\|_2^2 + \|\boldsymbol{V}^\star_{j,\cdot}\|_2^2\big)\Big) + \textit{negligible term}$$

- oracle information: $\{\boldsymbol{X}^\star_{k,\cdot}\}_{k:\,k\neq i}$, $\{\boldsymbol{Y}^\star_{k,\cdot}\}_{k:\,k\neq j}$

Distributional theory

$$M^\star \xrightarrow{\text{rank-}r\text{ svd}} U^\star \Sigma^\star V^{\star\top} \xrightarrow{\text{bal. decom.}} X^\star Y^{\star\top}$$

Theorem 2

Consider any (i,j) s.t. $\|U^\star_{i,\cdot}\|_2 + \|V^\star_{j,\cdot}\|_2$ is not too small. Then

$$M^{\mathsf{d}}_{i,j} - M^\star_{i,j} \sim \mathcal{N}\Big(0, \frac{\sigma^2}{p}\big(\|U^\star_{i,\cdot}\|_2^2 + \|V^\star_{j,\cdot}\|_2^2\big)\Big) + \textit{negligible term}$$

- oracle information: $\{X^\star_{k,\cdot}\}_{k:\,k\neq i}$, $\{Y^\star_{k,\cdot}\}_{k:\,k\neq j}$
- parameters: $X^\star_{i,\cdot}$, $Y^\star_{j,\cdot}$

Distributional theory

$$M^\star \xrightarrow{\text{rank-}r\text{ svd}} U^\star \Sigma^\star V^{\star\top} \xrightarrow{\text{bal. decom.}} X^\star Y^{\star\top}$$

Theorem 2

Consider any (i,j) s.t. $\|U^\star_{i,\cdot}\|_2 + \|V^\star_{j,\cdot}\|_2$ is not too small. Then

$$M^{\mathsf{d}}_{i,j} - M^\star_{i,j} \sim \mathcal{N}\Big(0, \frac{\sigma^2}{p}\big(\|U^\star_{i,\cdot}\|_2^2 + \|V^\star_{j,\cdot}\|_2^2\big)\Big) + \textit{negligible term}$$

- oracle information: $\{X^\star_{k,\cdot}\}_{k:\,k\neq i}$, $\{Y^\star_{k,\cdot}\}_{k:\,k\neq j}$
- parameters: $X^\star_{i,\cdot}$, $Y^\star_{j,\cdot}$
- estimate $M^\star_{i,j} = X^\star_{i,\cdot}(Y^\star_{j,\cdot})^\top$ based on a linear model

$$M_{i,k} = X^\star_{i,\cdot}(Y^\star_{k,\cdot})^\top + E_{i,k}, \qquad (i,k)\in\Omega$$
$$M_{k,j} = X^\star_{k,\cdot}(Y^\star_{j,\cdot})^\top + E_{k,j}, \qquad (k,j)\in\Omega$$

Distributional theory

$$\boldsymbol{M}^\star \xrightarrow{\text{rank-}r\text{ svd}} \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \xrightarrow{\text{bal. decom.}} \boldsymbol{X}^\star \boldsymbol{Y}^{\star\top}$$

Theorem 2

Consider any (i,j) s.t. $\|\boldsymbol{U}^\star_{i,\cdot}\|_2 + \|\boldsymbol{V}^\star_{j,\cdot}\|_2$ is not too small. Then

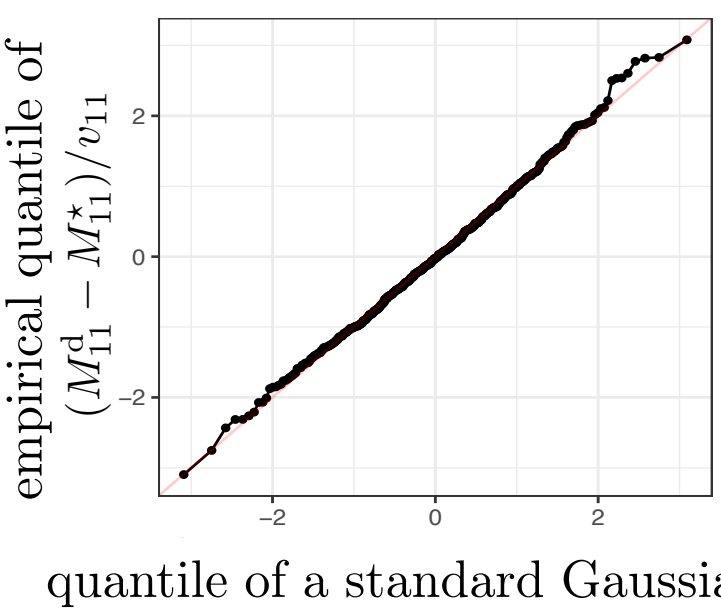
$$M^{\mathsf{d}}_{i,j} - M^\star_{i,j} \sim \mathcal{N}\big(\mathbf{0}, \text{Cramer-Rao}\big) + \textit{negligible term}$$

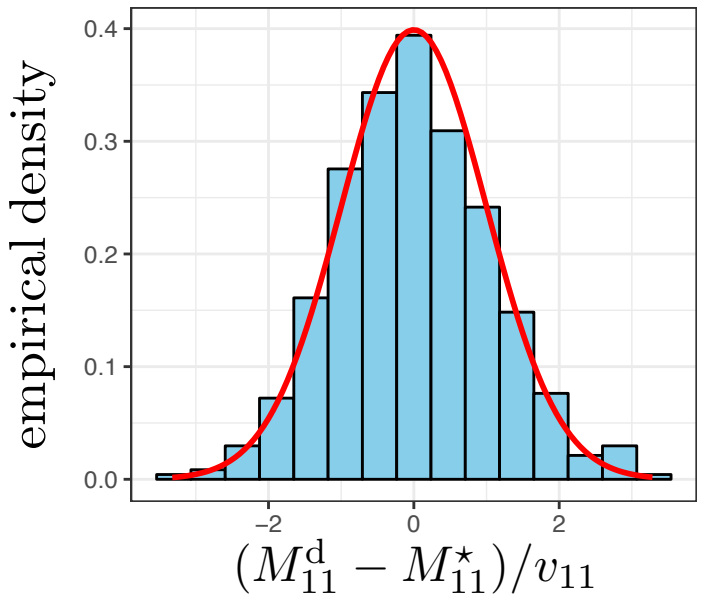
- oracle information: $\{\boldsymbol{X}^\star_{k,\cdot}\}_{k:\,k\neq i}$, $\{\boldsymbol{Y}^\star_{k,\cdot}\}_{k:\,k\neq j}$
- parameters: $\boldsymbol{X}^\star_{i,\cdot}$, $\boldsymbol{Y}^\star_{j,\cdot}$
- estimate $M^\star_{i,j} = \boldsymbol{X}^\star_{i,\cdot}(\boldsymbol{Y}^\star_{j,\cdot})^\top$ based on a linear model

$$M_{i,k} = \boldsymbol{X}^\star_{i,\cdot}(\boldsymbol{Y}^\star_{k,\cdot})^\top + E_{i,k}, \qquad (i,k)\in\Omega$$
$$M_{k,j} = \boldsymbol{X}^\star_{k,\cdot}(\boldsymbol{Y}^\star_{j,\cdot})^\top + E_{k,j}, \qquad (k,j)\in\Omega$$

— *asymptotically optimal!*

Numerical experiments





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

A bit of analysis

$$\boldsymbol{M}^{\mathsf{d}} = \mathsf{proj}_{\mathsf{rank}\text{-}r}\Big(\boldsymbol{M}^{\mathsf{cvx}} + \tfrac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\mathsf{cvx}}) \Big)$$

- challenging to characterize the distribution of $\boldsymbol{M}^{\mathsf{d}}$

A bit of analysis

$$\boldsymbol{M}^{\mathsf{d}} = \mathsf{proj}_{\mathsf{rank}\text{-}r}\Big(\boldsymbol{M}^{\mathsf{cvx}} + \tfrac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\mathsf{cvx}}) \Big)$$

- challenging to characterize the distribution of $\boldsymbol{M}^{\mathsf{d}}$
- solution: resort to nonconvex optimization

nonconvex optimization

A bit of analysis

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

A bit of analysis

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

$$\Downarrow$$

$$\text{nonconvex solution } (\boldsymbol{X},\boldsymbol{Y}) \implies \boldsymbol{M}^{\text{cvx}} \approx \boldsymbol{X}\boldsymbol{Y}^\top$$

A bit of analysis

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \sum_{(i,j)\in\Omega} \Big[\big(\boldsymbol{X}\boldsymbol{Y}^\top\big)_{i,j} - M_{i,j}\Big]^2 + \tfrac{\lambda}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \tfrac{\lambda}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2$$

$$\Downarrow$$

$$\text{nonconvex solution } (\boldsymbol{X},\boldsymbol{Y}) \implies \boldsymbol{M}^{\mathsf{cvx}} \approx \boldsymbol{X}\boldsymbol{Y}^\top$$

$$\Downarrow$$

$$\boldsymbol{X}^{\mathrm{d}} = \boldsymbol{X}\Big[\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\Big]^{1/2}, \quad \boldsymbol{Y}^{\mathrm{d}} = \boldsymbol{Y}\Big[\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{Y}\boldsymbol{Y}^\top)^{-1}\Big]^{1/2}$$

A bit of analysis

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \sum_{(i,j)\in\Omega} \Big[\big(\boldsymbol{X}\boldsymbol{Y}^\top\big)_{i,j} - M_{i,j}\Big]^2 + \tfrac{\lambda}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \tfrac{\lambda}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2$$

$$\Downarrow$$

$$\text{nonconvex solution } (\boldsymbol{X},\boldsymbol{Y}) \implies \boldsymbol{M}^{\mathsf{cvx}} \approx \boldsymbol{X}\boldsymbol{Y}^\top$$

$$\Downarrow$$

$$\boldsymbol{X}^{\mathrm{d}} = \boldsymbol{X}\Big[\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\Big]^{1/2}, \quad \boldsymbol{Y}^{\mathrm{d}} = \boldsymbol{Y}\Big[\boldsymbol{I}_r + \frac{\lambda}{p}(\boldsymbol{Y}\boldsymbol{Y}^\top)^{-1}\Big]^{1/2}$$

$$\Downarrow$$

$$\boldsymbol{M}^{\mathsf{d}} \approx \boldsymbol{X}^{\mathsf{d}}\boldsymbol{Y}^{\mathsf{d}\top}$$

A bit of analysis

with high prob., there exists global rotation matrix $\boldsymbol{H} \in \mathbb{R}^{r \times r}$ s.t.

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1})$$

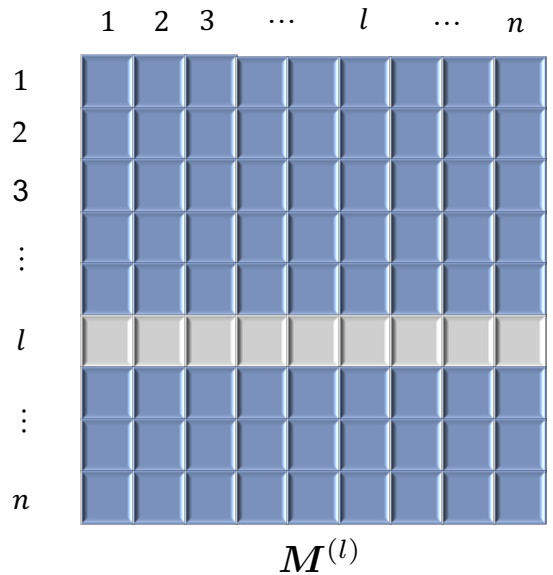
$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1})$$

A bit of analysis

with high prob., there exists global rotation matrix $\boldsymbol{H} \in \mathbb{R}^{r\times r}$ s.t.

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1})$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1})$$

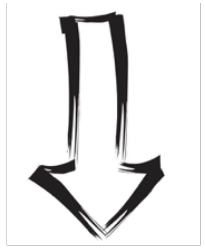


$$\Longrightarrow \{(\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)})\}_{t\geq 0}$$

Leave-one-out anlaysis: leave out a small amount of information from data and run GD...

convex

nonconvex



inference (convex) $\approx$ inference (nonconvex)

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

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

Back to estimation: de-biased estimator is optimal

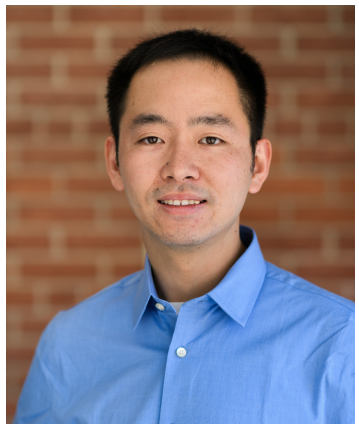
Distributional theory in turn allows us to track estimation accuracy

Theorem 3

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

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

Part 2: Inference for heteroskedastic PCA with missing data



Yuxin Chen
UPenn Wharton

Jianqing Fan
Princeton ORFE

Principal component analysis

Pearson '1901

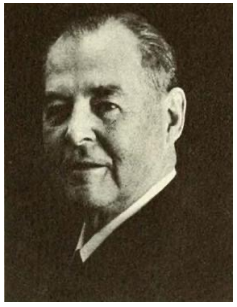
Hotelling '1933

Hotelling '1936

1900

2020

Karl Pearson



Harold Hotelling

Principal component analysis

A spiked covariance model:

$$\text{data} \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \underbrace{\boldsymbol{S}^\star}_{\text{low-rank}} + \sigma^2 \boldsymbol{I}_d)$$

Principal component analysis

A spiked covariance model:

$$\text{data} \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \underbrace{\boldsymbol{S}^\star}_{\text{low-rank}} + \sigma^2 \boldsymbol{I}_d)$$

Goal: estimate
- spiked covariance matrix $\boldsymbol{S}^\star$
- principal subspace $\boldsymbol{U}^\star$ — the eigenspace of $\boldsymbol{S}^\star$

Principal component analysis

A spiked covariance model:

$$\text{data} \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \underbrace{\boldsymbol{S}^\star}_{\text{low-rank}} + \sigma^2 \boldsymbol{I}_d)$$

Goal: estimate
- spiked covariance matrix $\boldsymbol{S}^\star$
- principal subspace $\boldsymbol{U}^\star$ — the eigenspace of $\boldsymbol{S}^\star$

- Johnstone '03
- Baik, Arous, Péché '05
- Paul '07
- Nadler '08
- Donoho, Gavish, Johnstone '18
- Bao, Ding, Wang, Wang '22
- Zhang, Cai, Wu '22
- ...

Principal component analysis

A spiked covariance model:

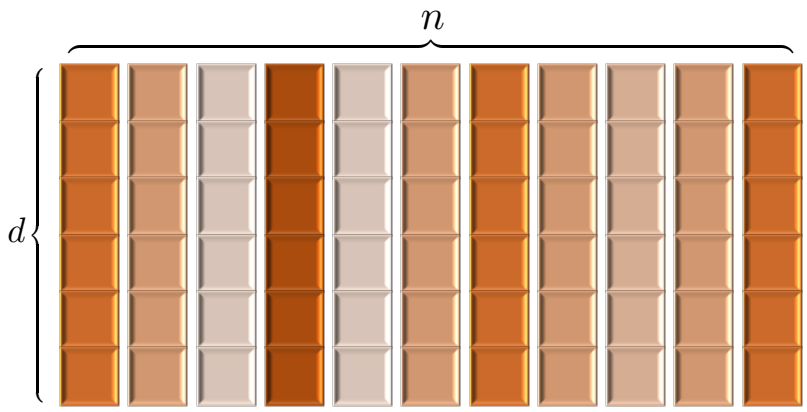
$$\text{data} \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \underbrace{\boldsymbol{S}^\star}_{\text{low-rank}} + \sigma^2 \boldsymbol{I}_d)$$

Goal: estimate
- spiked covariance matrix $\boldsymbol{S}^\star$
- principal subspace $\boldsymbol{U}^\star$ — the eigenspace of $\boldsymbol{S}^\star$

- Johnstone '03
- Baik, Arous, Péché '05
- Paul '07
- Nadler '08
- Donoho, Gavish, Johnstone '18
- Bao, Ding, Wang, Wang '22
- Zhang, Cai, Wu '22
- ...

Today's talk: inference for PCA in high dimension

Principal component analysis



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

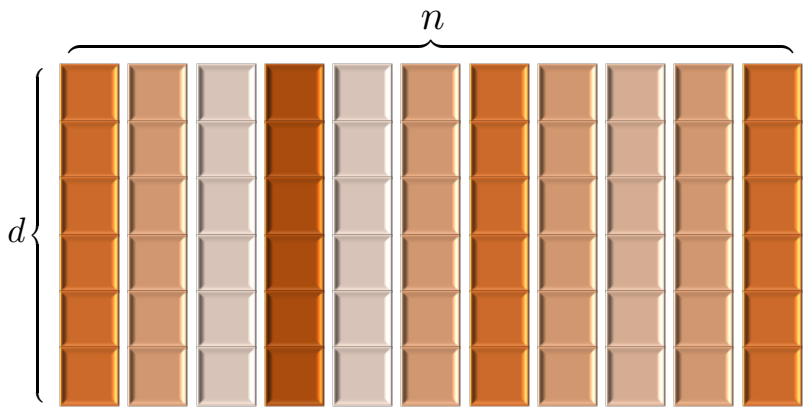
- Ground-truth data

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

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

Principal component analysis

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



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

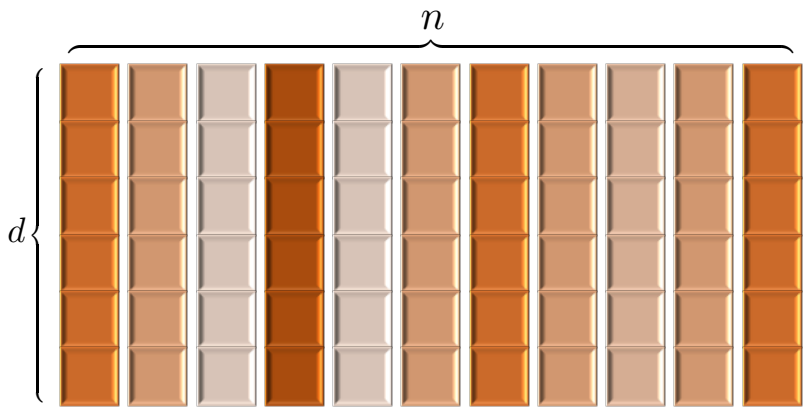
- Ground-truth data

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

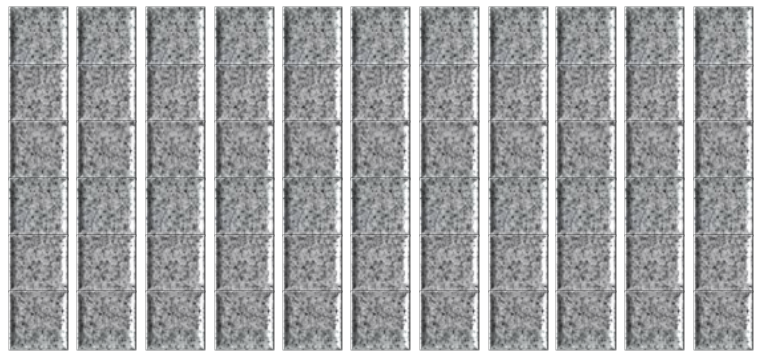
where $\boldsymbol{S}^\star = \boldsymbol{U}^\star \boldsymbol{\Lambda}^\star \boldsymbol{U}^{\star\top} \in \mathbb{R}^{d \times d}$ has rank $r \ll d$

Principal component analysis

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



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



noise matrix: $\boldsymbol{E}$

- Ground-truth data

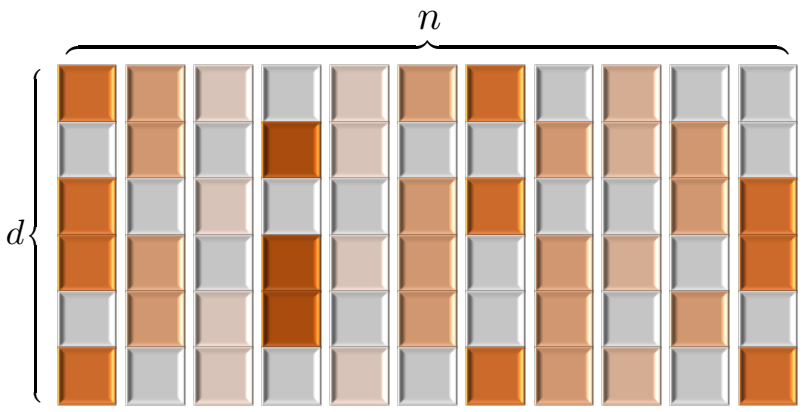
$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}, \qquad \boldsymbol{x}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}^\star)$$

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

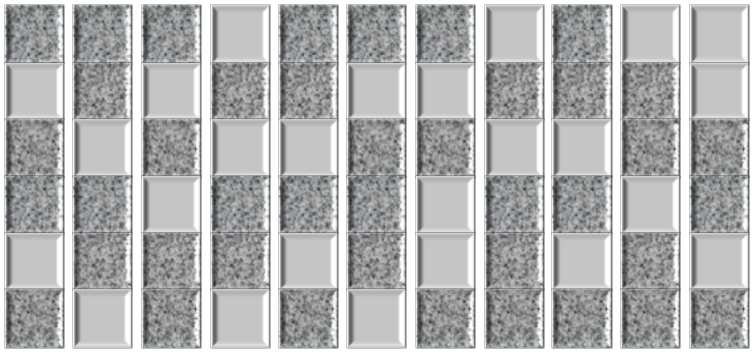
- Noisy observations: $\boldsymbol{X} + \boldsymbol{E}$ (a.k.a. spiked covariance model)

Principal component analysis

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



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



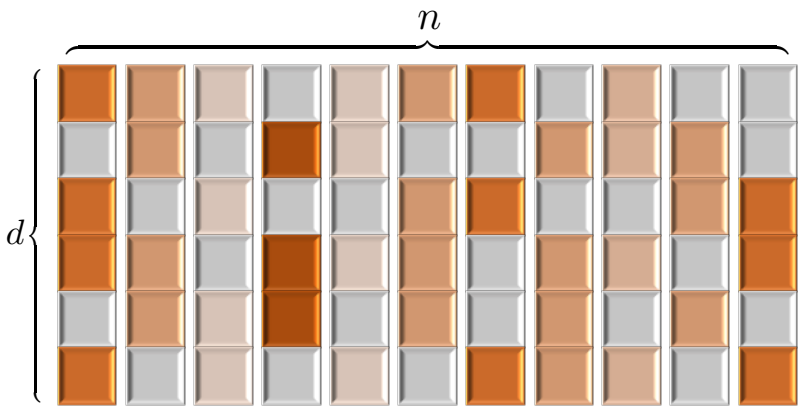
noise matrix: $\boldsymbol{E}$

- Incomplete observations $\longrightarrow$ sampling set Ω:

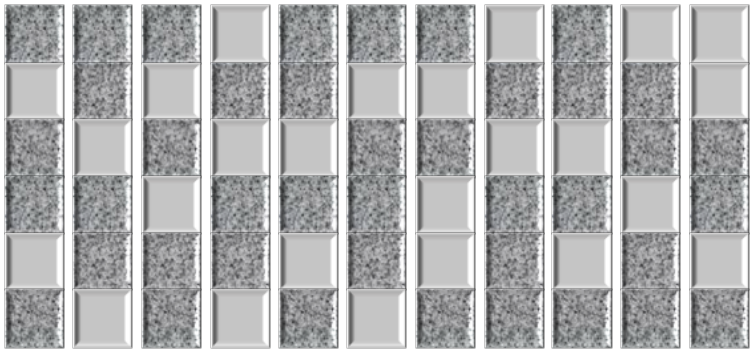
$$Y_{i,j} = \begin{cases} X_{i,j} + E_{i,j}, & (i,j) \in \Omega \\ 0, & \text{else} \end{cases} \qquad \text{or} \quad \boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{E})$$

Principal component analysis

$\mathrm{span}\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subseteq \boldsymbol{U}^\star$ (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

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

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

What we consider here ...

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

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

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

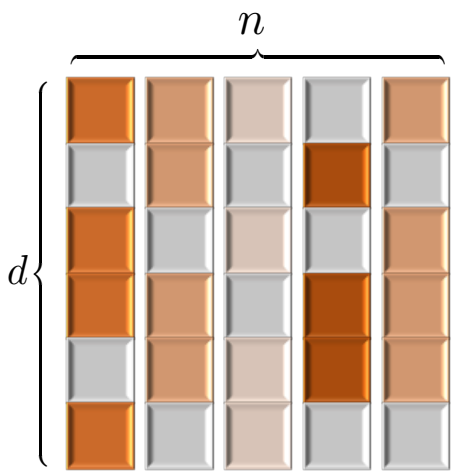
- **Random sampling**: $(i,j) \in \Omega$ independently with prob. p

What we consider here ...

Our focus: estimating/inferring column subspace of $\boldsymbol{X}$ from noisy and incomplete observations $\boldsymbol{Y} = \mathcal{P}_{\Omega}(\boldsymbol{X} + \boldsymbol{E})$

What we consider here ...

Our focus: estimating/inferring column subspace of $\boldsymbol{X}$ from noisy and incomplete observations $\boldsymbol{Y} = \mathcal{P}_{\Omega}(\boldsymbol{X} + \boldsymbol{E})$



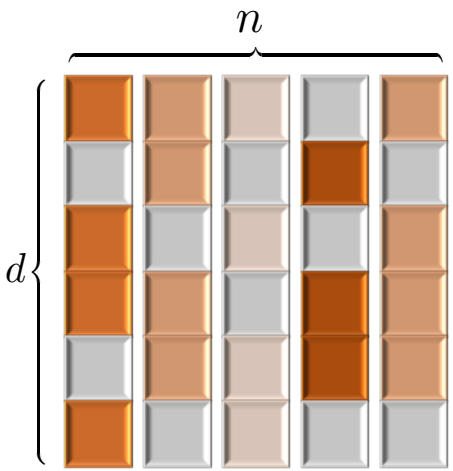
$n \lesssim d$: solvable via *matrix completion* methods

(e.g., Chen, Fan, Ma, Yan '19)

What we consider here . . .

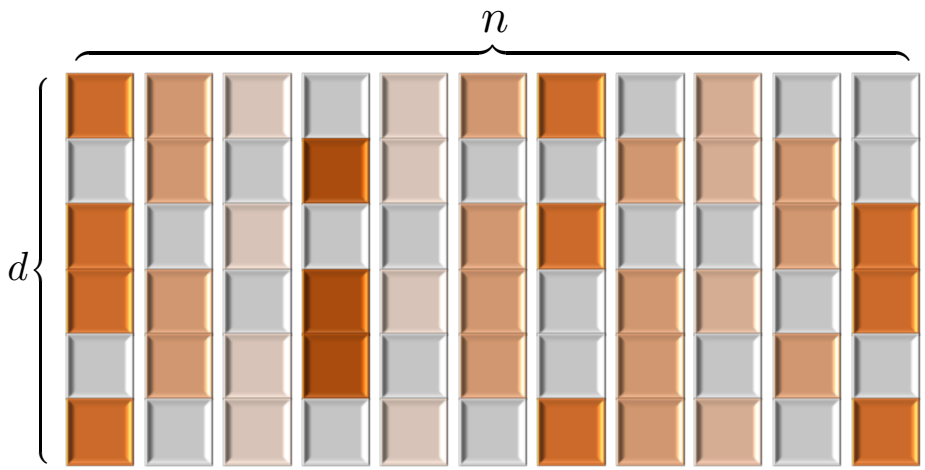
Our focus: estimating/inferring column subspace of $\boldsymbol{X}$ from noisy and incomplete observations $\boldsymbol{Y} = \mathcal{P}_{\Omega}(\boldsymbol{X} + \boldsymbol{E})$ when $\underbrace{n \gg d}$

more challenging regime



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

(e.g., Chen, Fan, Ma, Yan '19)



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

A natural SVD-based algorithm

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

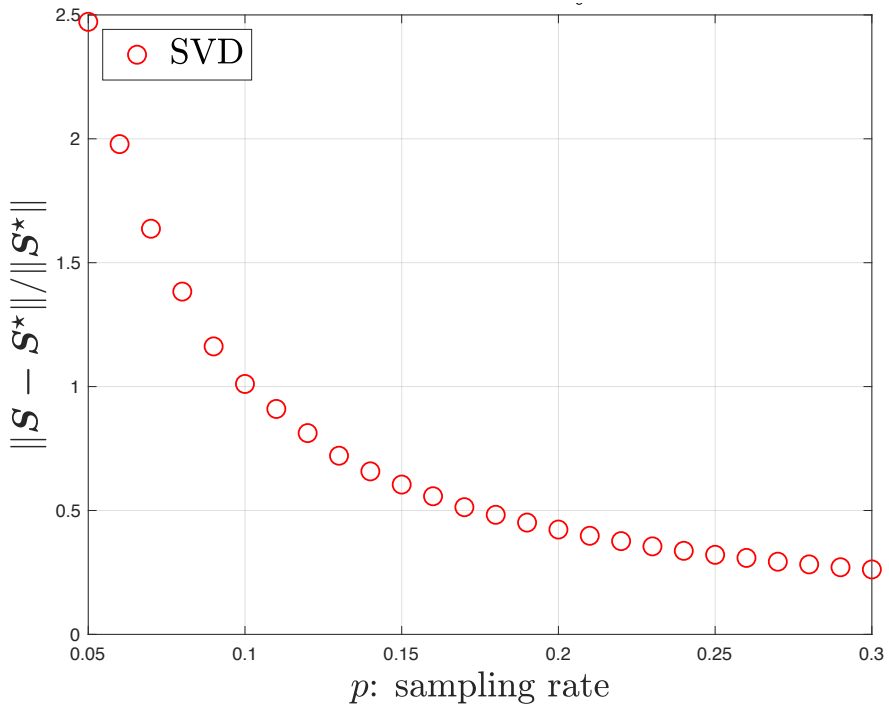
A natural SVD-based algorithm

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

Rationale: under zero-mean noise and random sampling, we have

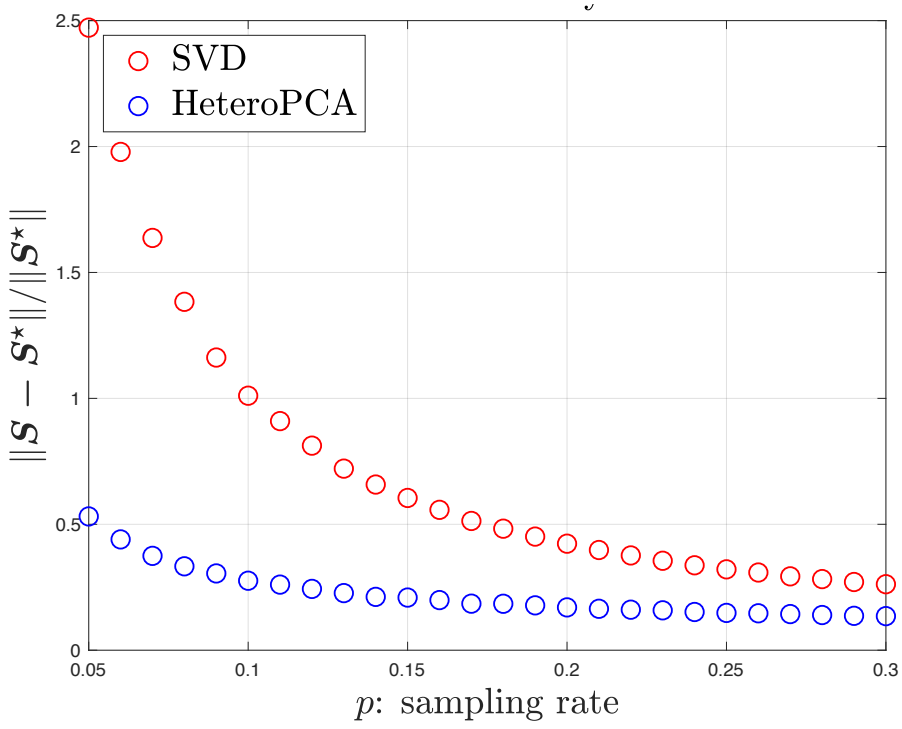
$$\text{col-space}(\,\mathbb{E}[\boldsymbol{Y}]\,) = \text{col-space}(\boldsymbol{X}) = \boldsymbol{U}^\star$$

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



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

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

Diagnosis: diagonal entries need special treatment

$$\text{col-space}(\boldsymbol{Y}) = \text{eig-space}(\boldsymbol{Y}\boldsymbol{Y}^\top)$$

Diagnosis: diagonal entries need special treatment

$$\text{col-space}(\boldsymbol{Y}) = \text{eig-space}(\boldsymbol{Y}\boldsymbol{Y}^\top)$$

Large bias in diagonal entries:

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

Diagnosis: diagonal entries need special treatment

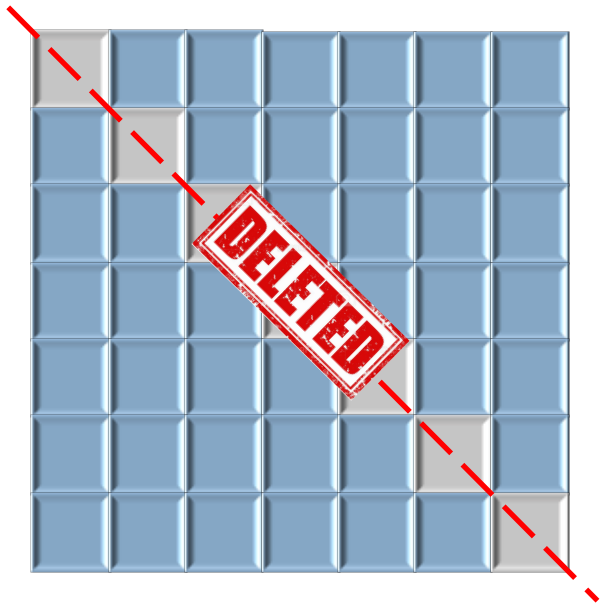
$$\text{col-space}(\boldsymbol{Y}) = \text{eig-space}(\boldsymbol{Y}\boldsymbol{Y}^\top)$$

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

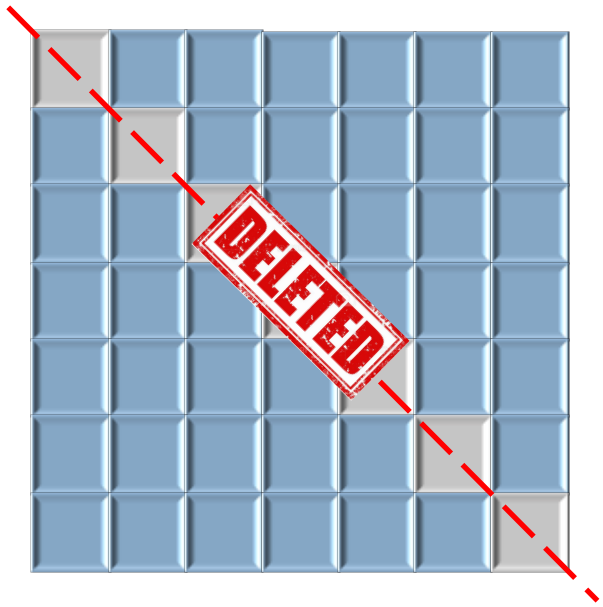
Two spectral algorithms that take care of diagonals



diagonal-deleted/reweighted PCA

- remove/reweight diag($\boldsymbol{Y}\boldsymbol{Y}^\top$)

Two spectral algorithms that take care of diagonals

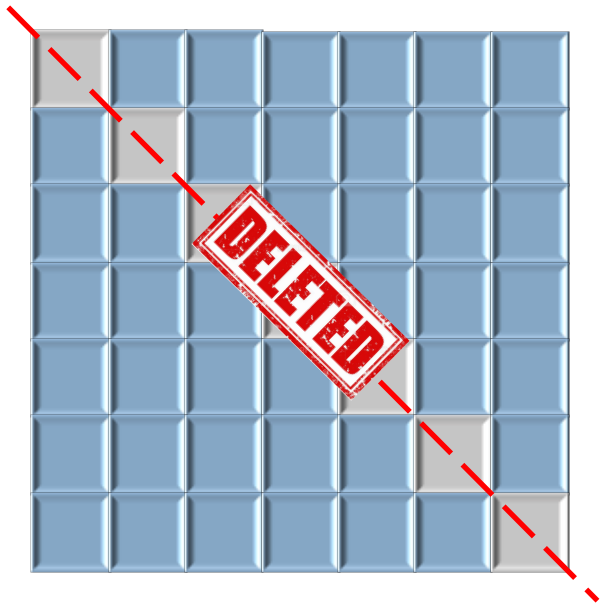


diagonal-deleted/reweighted PCA

- remove/reweight diag($\boldsymbol{Y}\boldsymbol{Y}^\top$)

  — Loh, Wainwright '12
  — Lounici '13 '14
  — Florescu and Perkins '16
  — Montanari and Sun '18
  — Zhu, Wang, Samworth '19
  — Cai, Li, Chi, Poor, Chen '19
  — ...

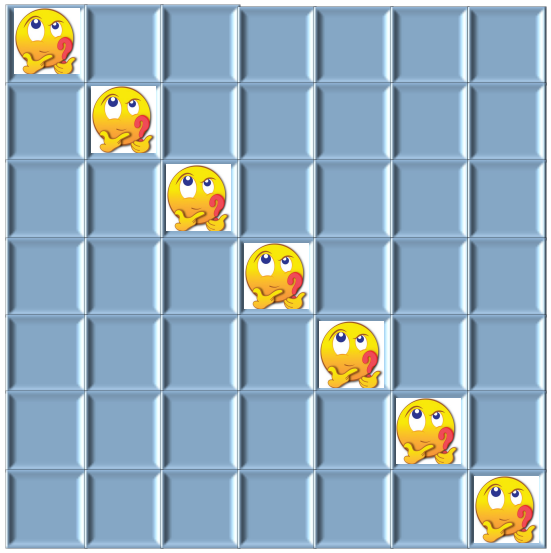
Two spectral algorithms that take care of diagonals



diagonal-deleted/reweighted PCA

- remove/reweight diag($\boldsymbol{Y}\boldsymbol{Y}^\top$)

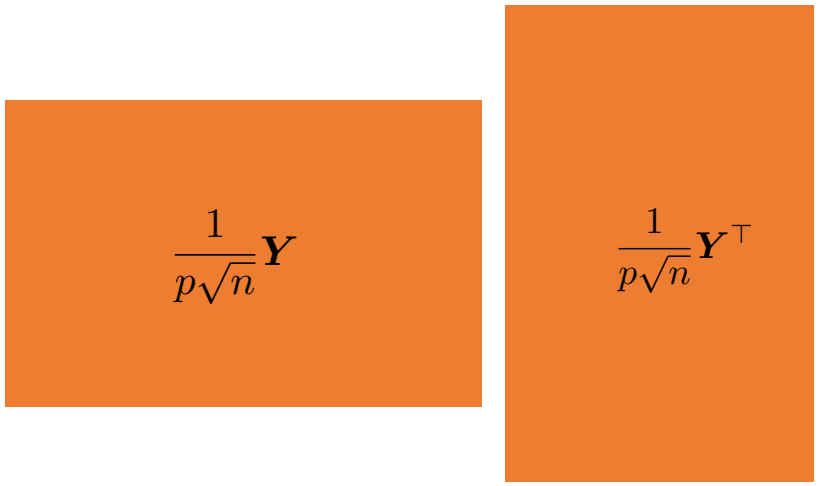
  — Loh, Wainwright '12
  — Lounici '13 '14
  — Florescu and Perkins '16
  — Montanari and Sun '18
  — Zhu, Wang, Samworth '19
  — Cai, Li, Chi, Poor, Chen '19
  — ...



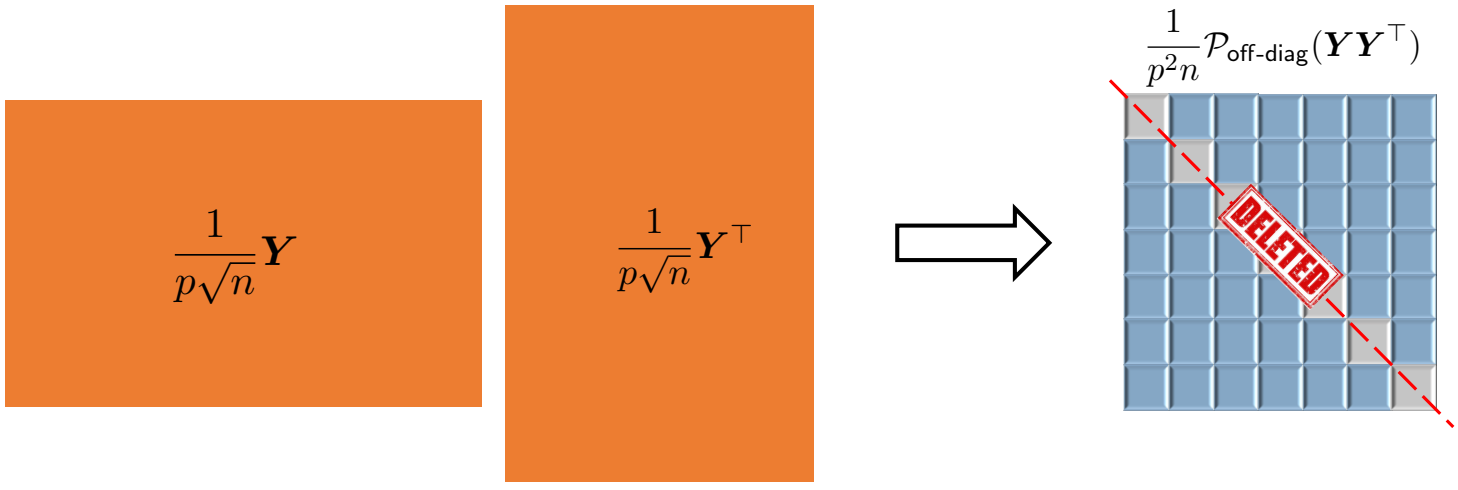
HeteroPCA (Zhang et al '18)

- iteratively estimate diag($\boldsymbol{Y}\boldsymbol{Y}^\top$)

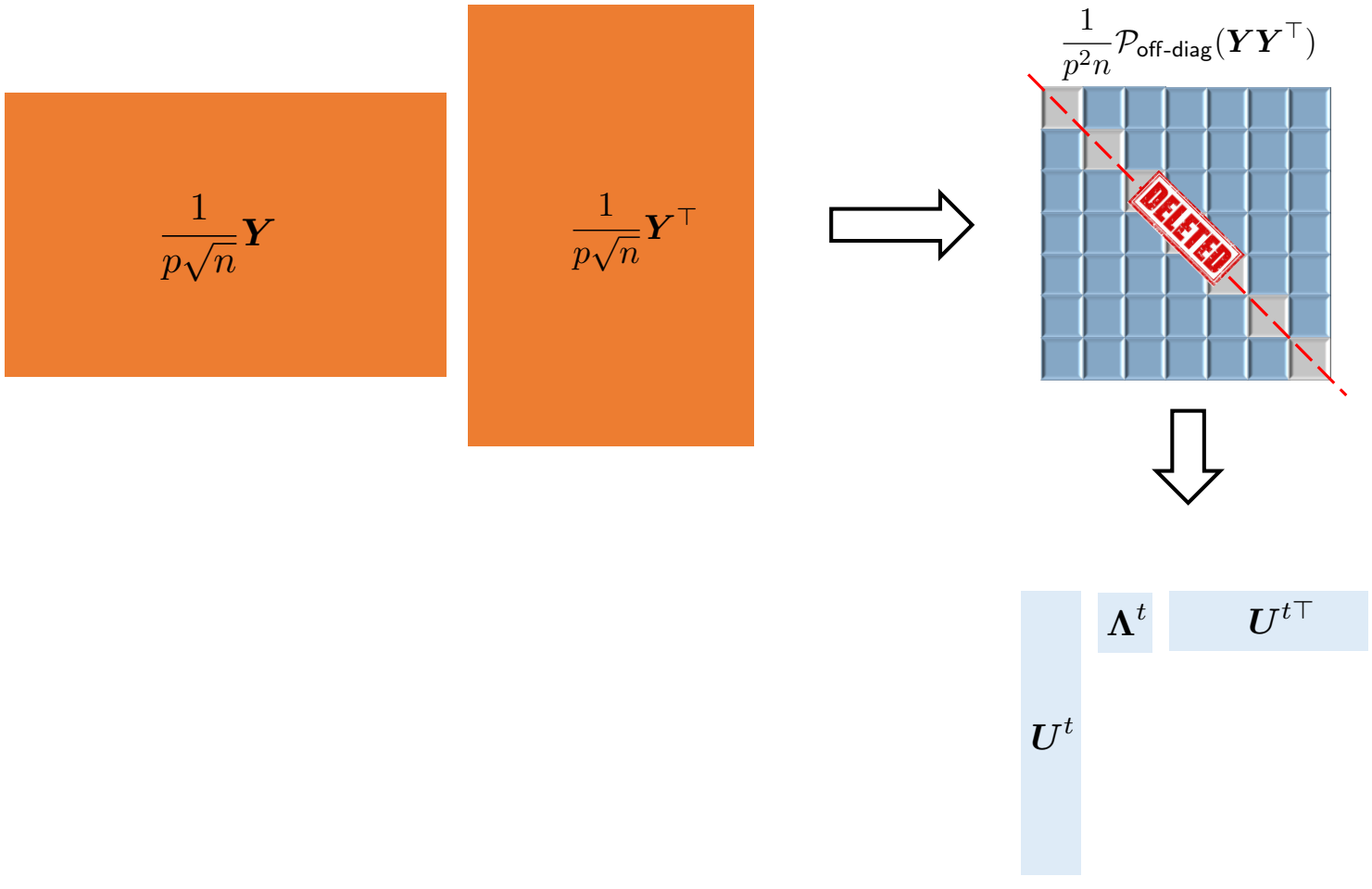
HeteroPCA (Zhang, Cai, Wu '18)



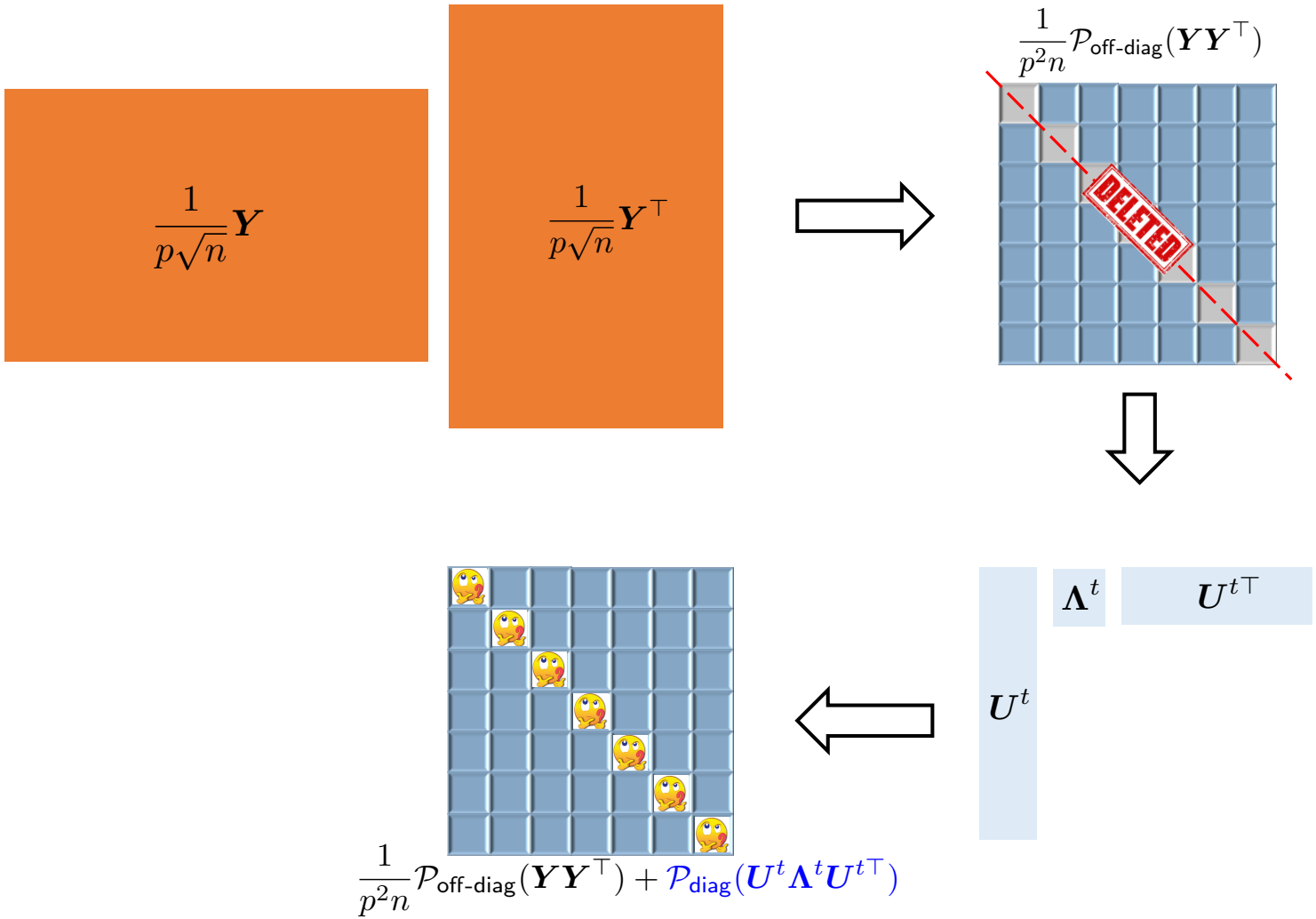
HeteroPCA (Zhang, Cai, Wu '18)



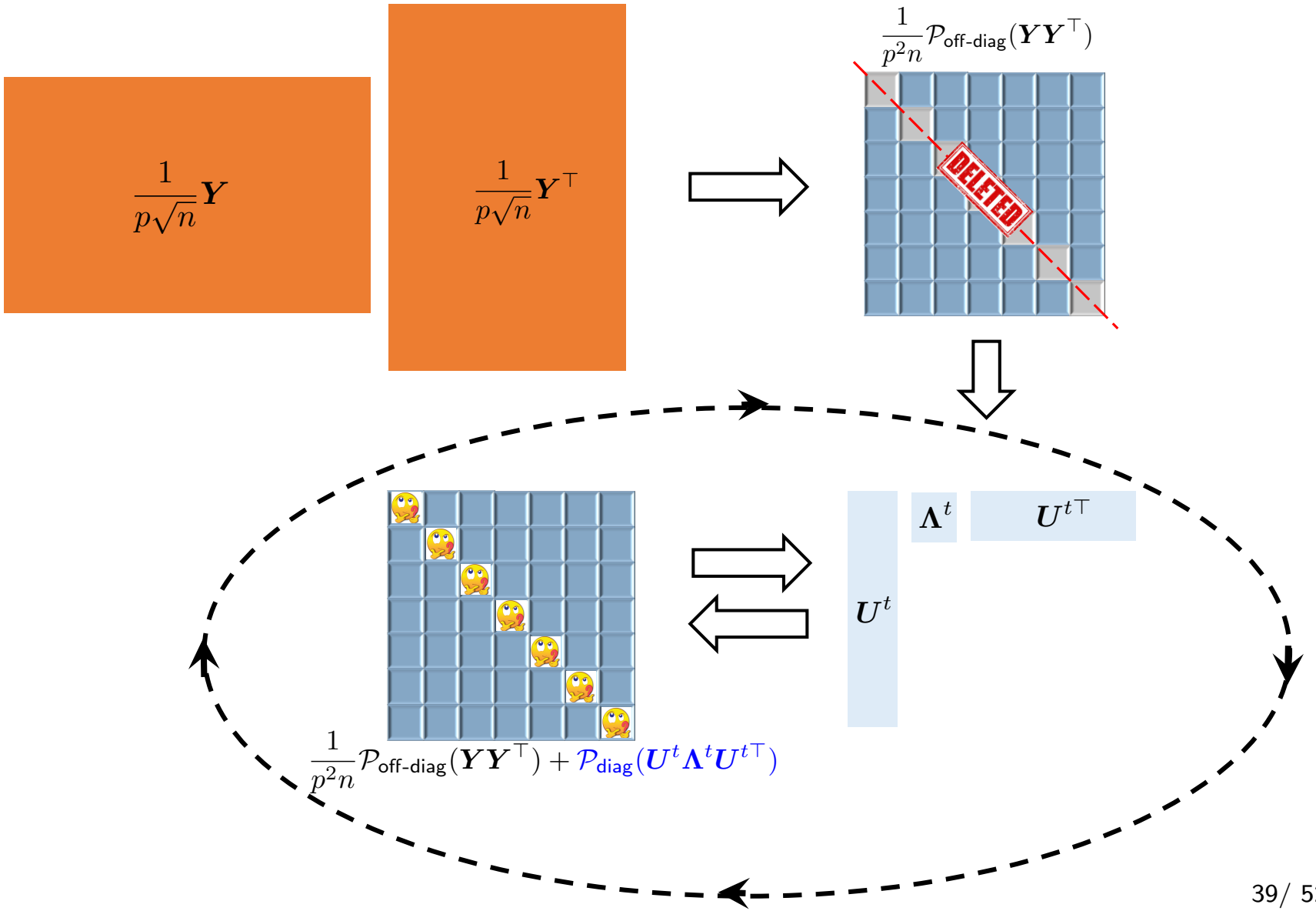
HeteroPCA (Zhang, Cai, Wu '18)



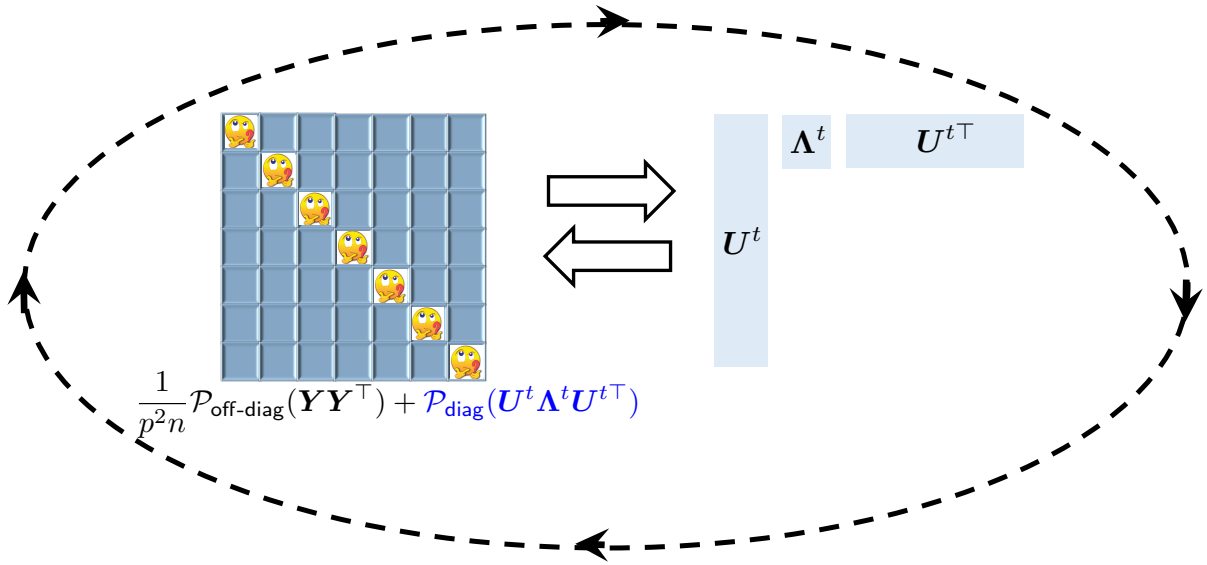
HeteroPCA (Zhang, Cai, Wu '18)



HeteroPCA (Zhang, Cai, Wu '18)



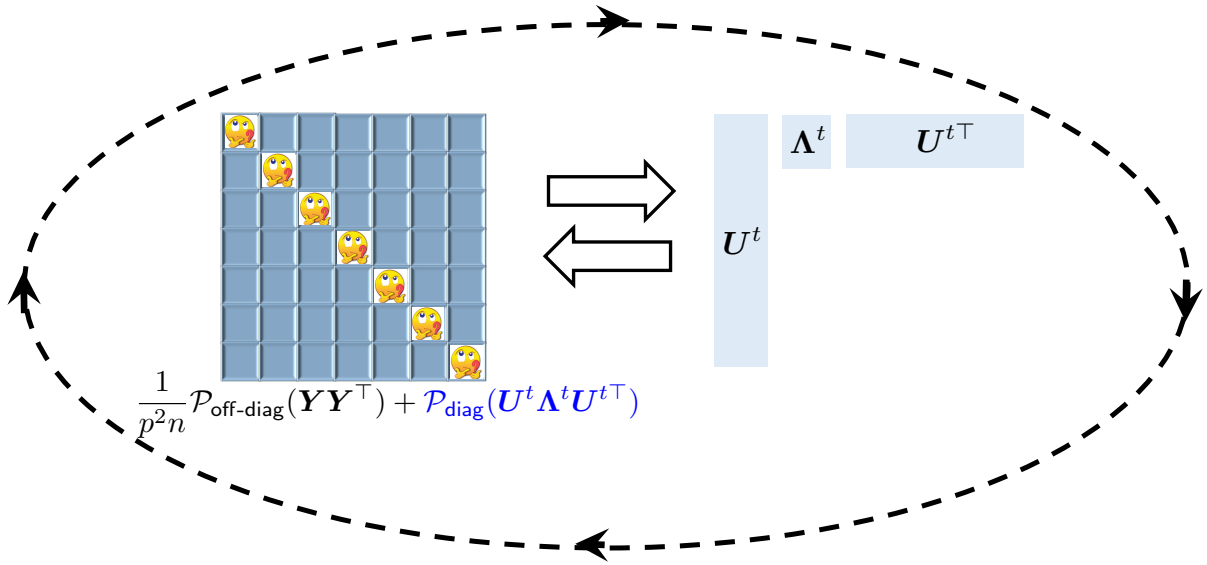
HeteroPCA (Zhang, Cai, Wu '18)



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

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

HeteroPCA (Zhang, Cai, Wu '18)

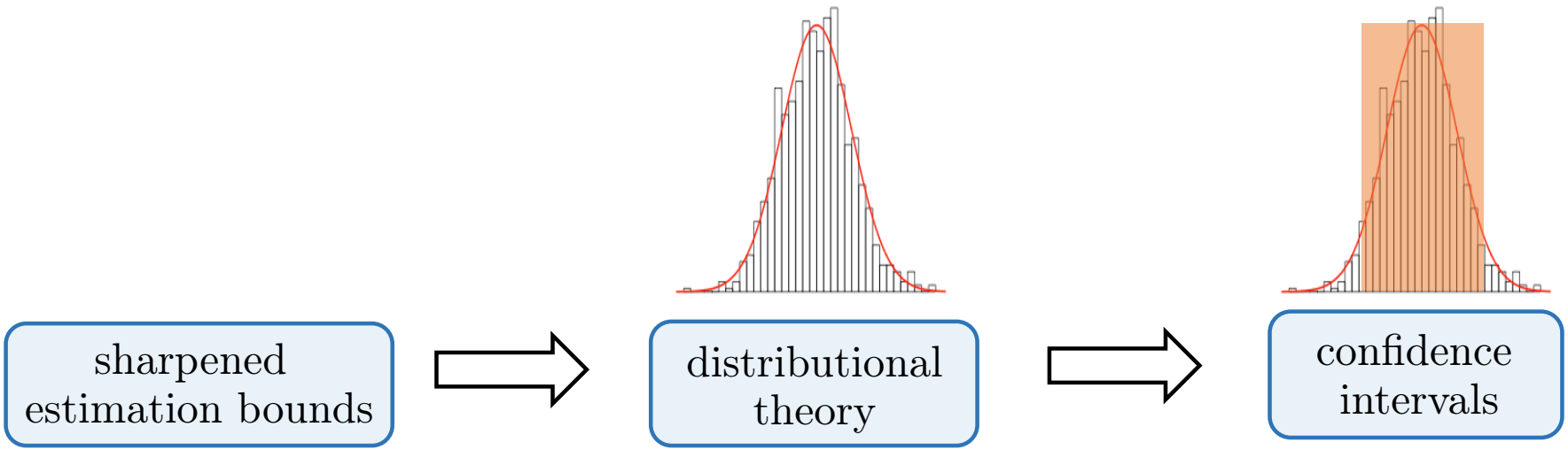


- **Initialize:** $\boldsymbol{G}^0 = \frac{1}{np^2}\mathcal{P}_{\text{off-diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)$
- **Iterative update:** for $t = 0, 1, \ldots, t_0$

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

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

Our contributions: estimation and inference based on HeteroPCA



Sharpened estimation guarantees for HeteroPCA

Assumptions (omitting log factors)

- $\boldsymbol{S}^\star$ has rank $r = O(1)$, incoherent, well-conditioned
- sampling rate exceeds certain threshold

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

- per-entry signal-to-noise ratio (SNR) cannot be too low:

$$\frac{\omega_{\max}^2}{\lambda_r(\boldsymbol{S}^\star)/d} \lesssim \min\left\{pn,\ p\sqrt{nd}\right\}$$

Sharpened estimation guarantees for HeteroPCA

Assumptions (omitting log factors)

- $\boldsymbol{S}^\star$ has rank $r = O(1)$, incoherent, well-conditioned
- sampling rate exceeds certain threshold

$$p \quad \gtrsim \quad \underbrace{\frac{1}{\sqrt{nd}}}_{\text{HeteroPCA}} \qquad \text{vs.} \qquad \underbrace{\frac{1}{d}}_{\text{matrix completion}} \qquad (n \gg d)$$

- per-entry signal-to-noise ratio (SNR) cannot be too low:

$$\frac{\omega_{\max}^2}{\lambda_r(\boldsymbol{S}^\star)/d} \quad \lesssim \quad \underbrace{p\sqrt{nd}}_{\text{HeteroPCA}} \qquad \text{vs.} \qquad \underbrace{pd}_{\text{matrix completion}} \qquad (n \gg d)$$

Sharpened estimation guarantees for HeteroPCA

Theorem 4

With high prob., there exists global rotation matrix $\boldsymbol{H} \in \mathcal{O}^{r\times r}$ *s.t.*

$$\|\boldsymbol{U}\boldsymbol{H} - \boldsymbol{U}^\star\| \lesssim \zeta, \quad \|\boldsymbol{U}\boldsymbol{H} - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \frac{1}{\sqrt{d}}\zeta$$

where

$$\zeta := \frac{1}{\sqrt{nd}\,p} + \frac{\omega_{\max}^2}{p\,\lambda_r^\star}\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{np}} + \frac{\omega_{\max}}{\sqrt{\lambda_r^\star}}\sqrt{\frac{d}{np}} = o(1).$$

Sharpened estimation guarantees for HeteroPCA

Theorem 4

With high prob., there exists global rotation matrix $\boldsymbol{H} \in \mathcal{O}^{r\times r}$ *s.t.*

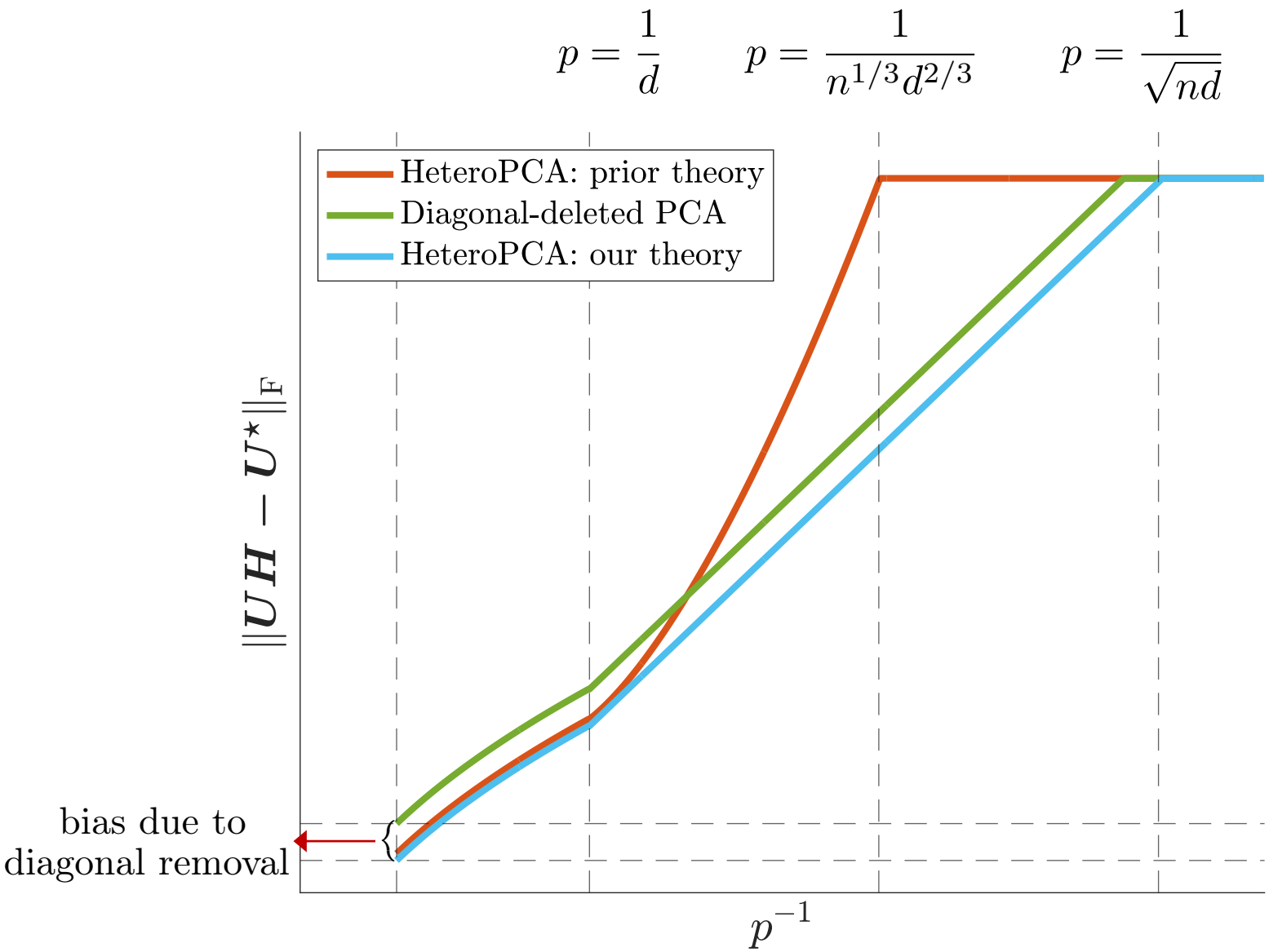
$$\|\boldsymbol{U}\boldsymbol{H} - \boldsymbol{U}^\star\| \lesssim \zeta, \quad \|\boldsymbol{U}\boldsymbol{H} - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \frac{1}{\sqrt{d}}\zeta$$

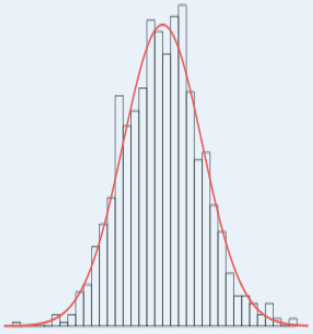
where

$$\zeta := \frac{1}{\sqrt{nd}\,p} + \frac{\omega_{\max}^2}{p\,\lambda_r^\star}\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{np}} + \frac{\omega_{\max}}{\sqrt{\lambda_r^\star}}\sqrt{\frac{d}{np}} = o(1).$$

- rotational ambiguity: for any orthonormal matrix $\boldsymbol{O} \in \mathcal{O}^{r\times r}$, $\boldsymbol{U}$ and $\boldsymbol{U}\boldsymbol{O}$ represents the same subspace
- global rotation matrix $\boldsymbol{H} = \arg\min_{\boldsymbol{O}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{O} - \boldsymbol{U}^\star\|_{\mathrm{F}}$

Sharpened estimation guarantees for HeteroPCA





sharpened estimation bounds

distributional theory

Given HeteroPCA is an appealing estimator, can we take one step further to obtain distributional characterizations?

Distributional theory for U

Theorem 5

Consider any $1 \leq l \leq d$. *Under previous assumptions, we have*

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

Distributional theory for U

Theorem 5

Consider any $1 \leq l \leq d$. *Under previous assumptions, we have*

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

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

Distributional theory for U

Theorem 5

Consider any $1 \leq l \leq d$. *Under previous assumptions, we have*

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

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

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

$$d^\star_{l,i} := \frac{1}{np^2} \left[\omega^{\star 2}_l + (1-p) S^{\star 2}_{l,l} \right] \left[\omega^{\star 2}_i + (1-p) S^{\star 2}_{i,i} \right] + \frac{2(1-p)^2}{np^2} S^{\star 2}_{l,i}$$

Distributional theory for U

Theorem 5

Consider any $1 \leq l \leq d$. *Under previous assumptions, we have*

$$\sup_{\text{cvx set } \mathcal{C}} \left| \mathbb{P}\Big((\boldsymbol{U}\boldsymbol{H} - \boldsymbol{U}^\star)_{l,\cdot} \in \mathcal{C} \Big) - \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^\star_{U,l}) \{\mathcal{C}\} \right| = o(1)$$

- **Key decomposition:**

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

  where $\boldsymbol{\Delta}$ is the "equivalent noise matrix"

$$\boldsymbol{\Delta} := \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{X} + \boldsymbol{E}) - \boldsymbol{X}$$

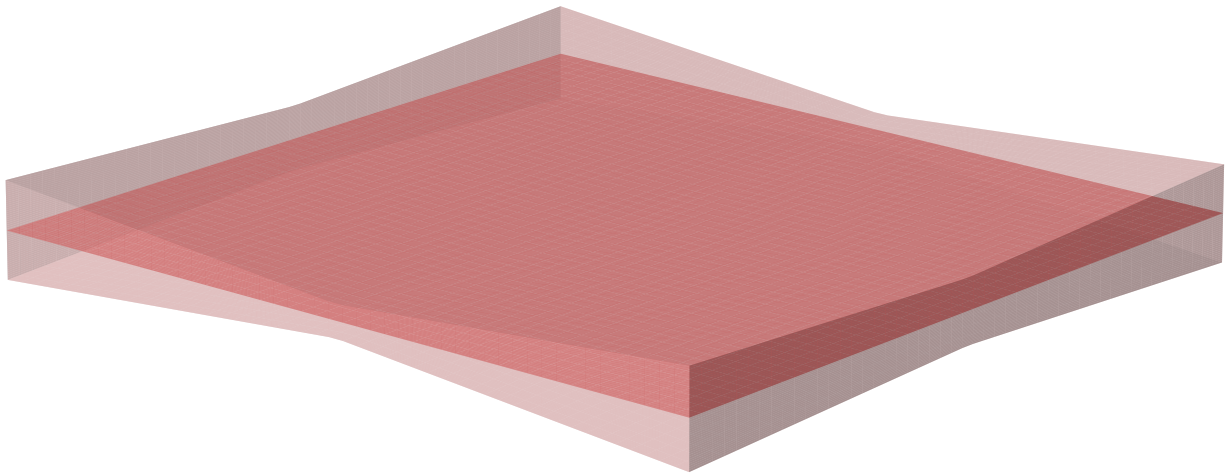
Distributional theory for U

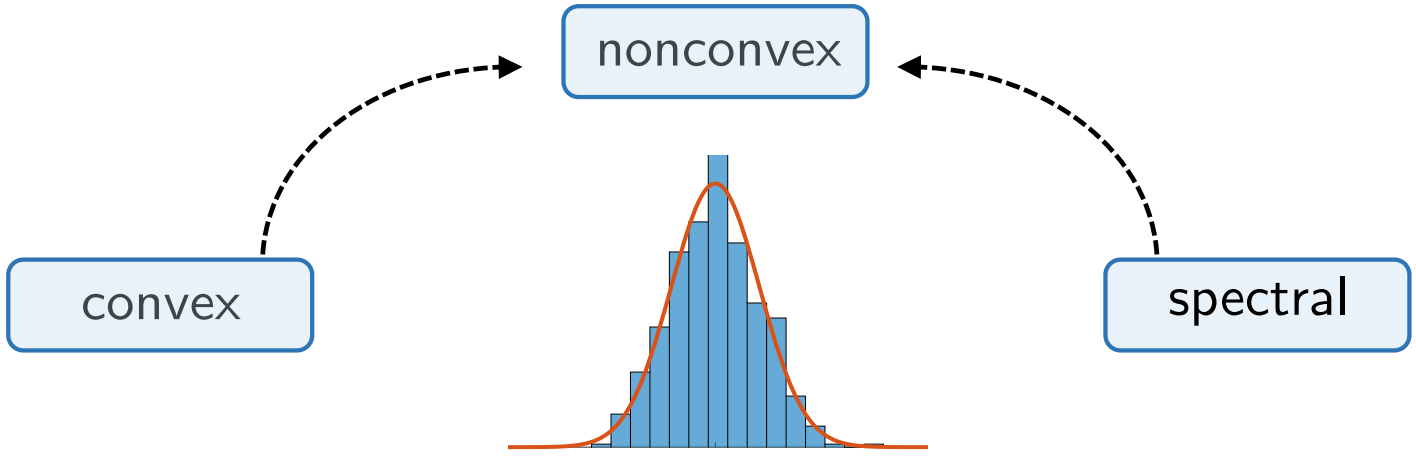
Theorem 5

Consider any $1 \leq l \leq d$. *Under previous assumptions, we have*

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

- confidence regions for the principal subspace





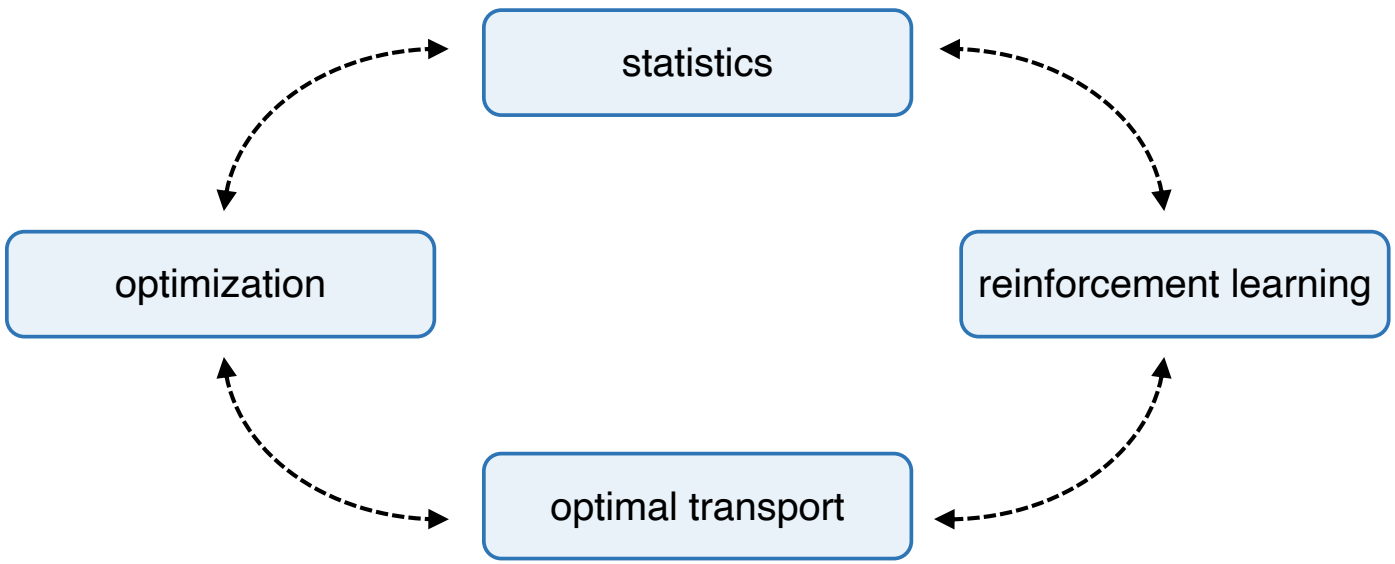
uncertainty quantification for low-rank models

"Inference and uncertainty quantification for noisy matrix completion," Y. Chen, J. Fan, C. Ma, Y. Yan, *Proceedings of National Academy of Sciences (PNAS)*

"Bridging convex and nonconvex optimization in robust PCA: noise, outliers, and missing data," Y. Chen, J. Fan, C. Ma, Y. Yan, *Annals of Statistics*

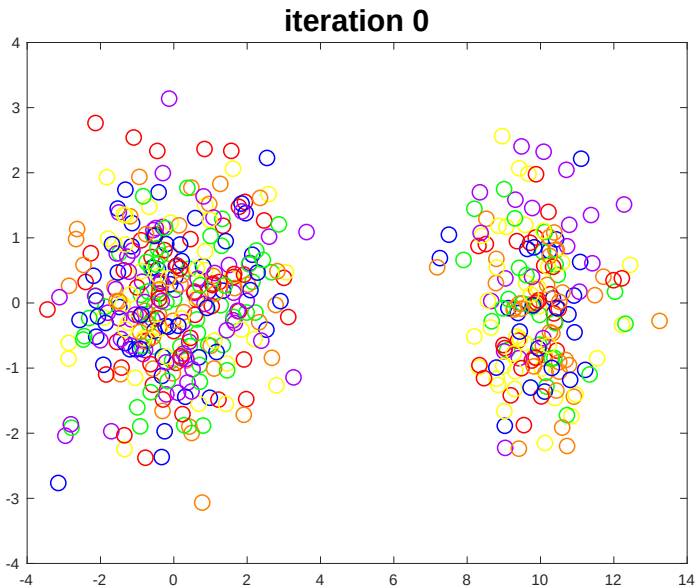
"Inference for heteroskedastic PCA with missing data," Y. Yan, Y. Chen, J. Fan

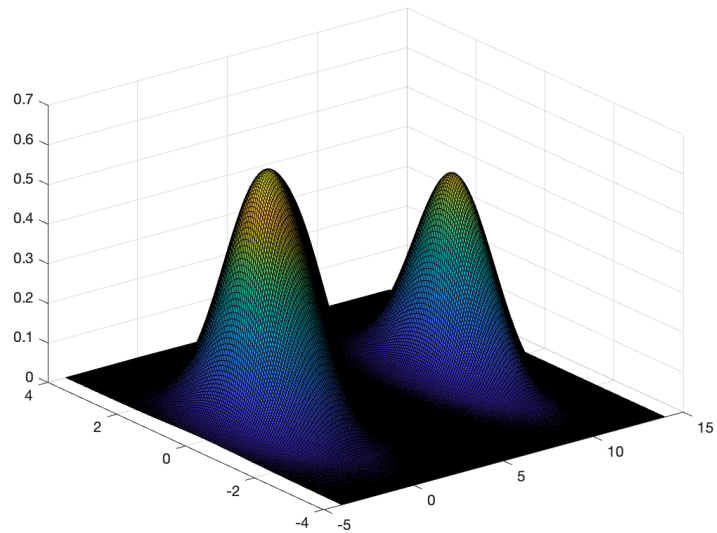
Other highlights of my research



Overparameterization for Gaussian mixture model

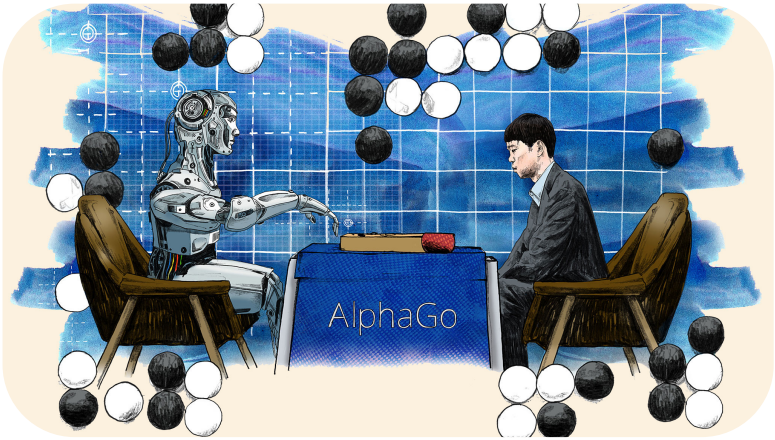
- **observations**: $X_1, \ldots, X_N \overset{\text{ind}}{\sim} \rho^\star * \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$
- **algorithm**: an interacting particle system where the weight and location of particles evolve according to the Wasserstein-Fisher-Rao gradient flow
- **theory**: convergence to nonparametric MLE in mean-field limit

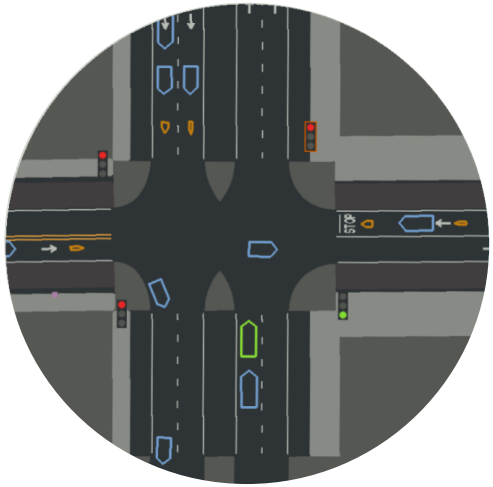


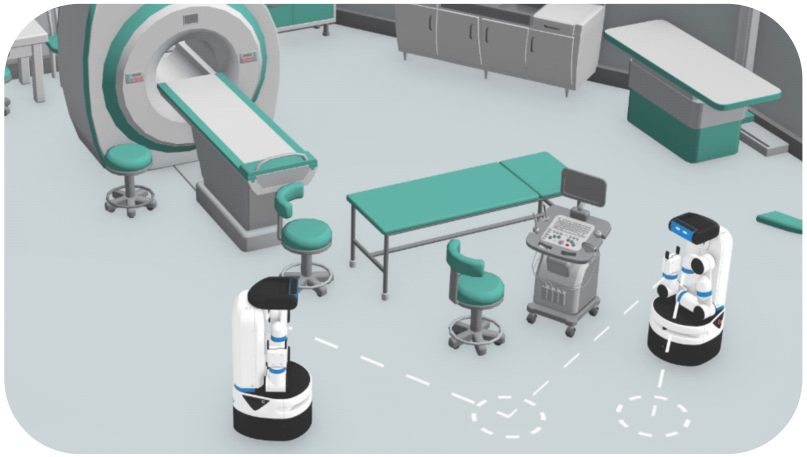


"Learning Gaussian mixtures using the Wasserstein-Fisher-Rao gradient flow," Y. Yan, K. Wang, P. Rigollet

Offline multi-agent reinforcement learning

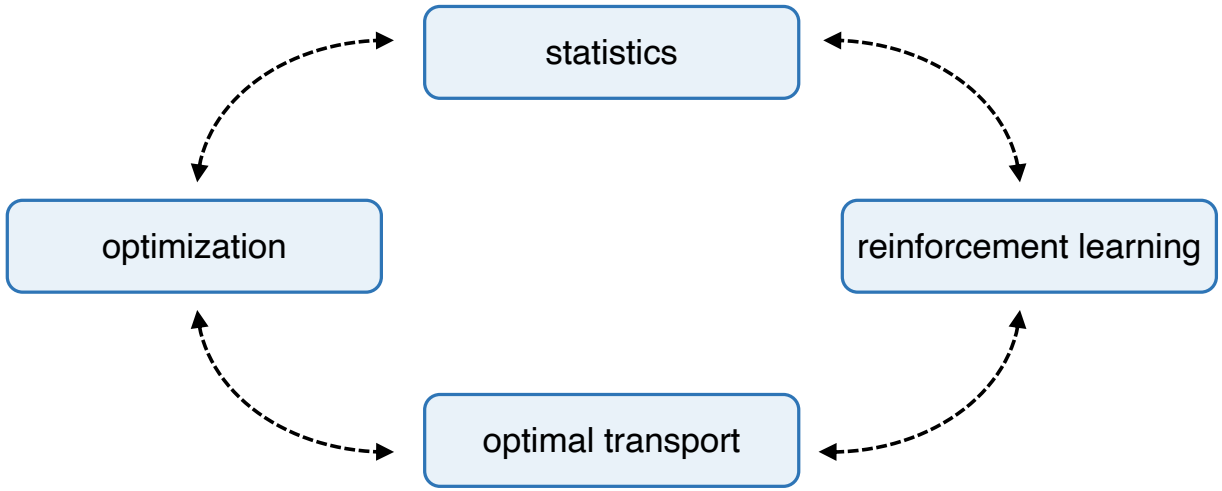






"Model-based reinforcement learning for offline zero-sum Markov games," Y. Yan, G. Li, Y. Chen, J. Fan

Many future opportunities...



- robust estimation/inference
- uncertainty quantification for deep learning
- multi-agent reinforcement learning
- applications in economics and social sciences $\cdots$

Thank you!