Estimation and Inference for Heteroskedastic PCA with Missing Data

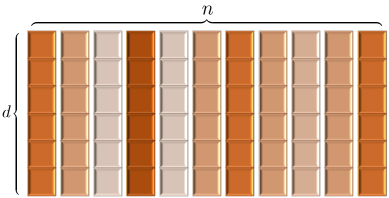
Yuxin Chen

Wharton Statistics & Data Science

Yuling Yan
Princeton ORFE

Jianqing Fan
Princeton ORFE

Principal component analysis



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

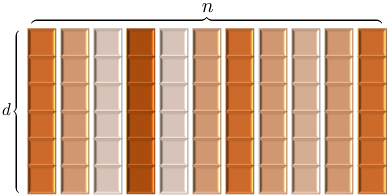
- Ground-truth data

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

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

Principal component analysis

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



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

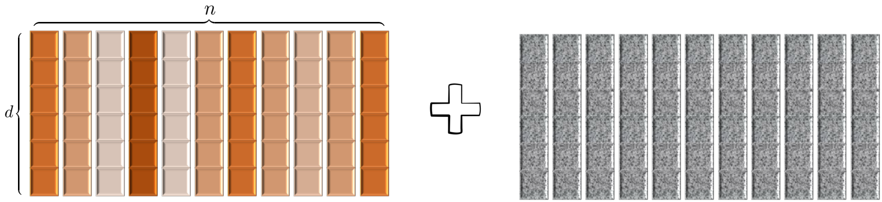
- Ground-truth data

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

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

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

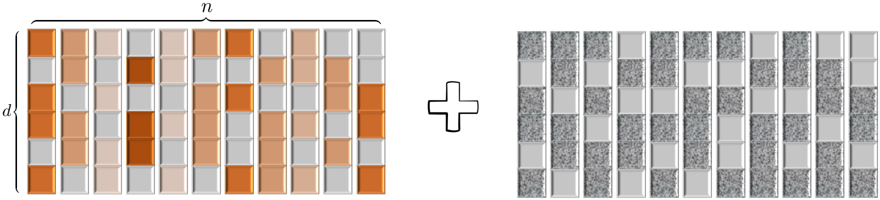
$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}, \qquad \boldsymbol{x}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}^\star)$$

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

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



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

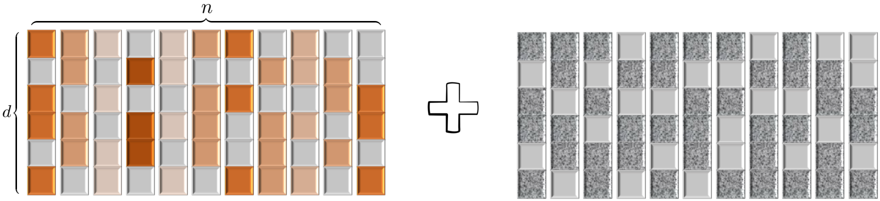
noise matrix: $\boldsymbol{E}$

- 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

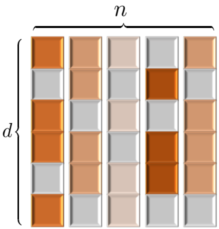
- **Random sampling**: $(i, j) \in \Omega$ independently with prob. p

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 $\underbrace{n \gg d}_{\text{more challenging regime}}$

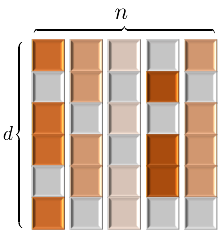


$n \lesssim d$: solvable via *matrix completion* methods

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

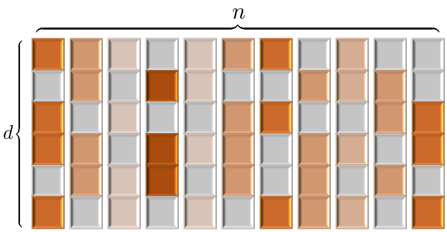
What we consider here ...

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



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

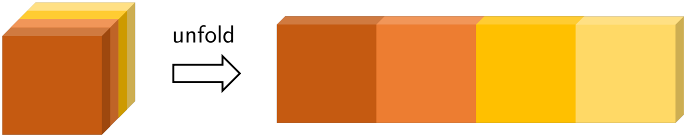
(e.g., Chen, Fan, Ma, Yan '19)



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

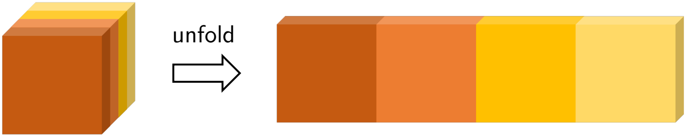
Applications beyond PCA

- Tensor completion

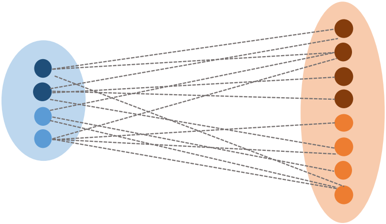


Applications beyond PCA

- Tensor completion



- One-sided community recovery in bipartite random graphs



A natural SVD-based algorithm

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

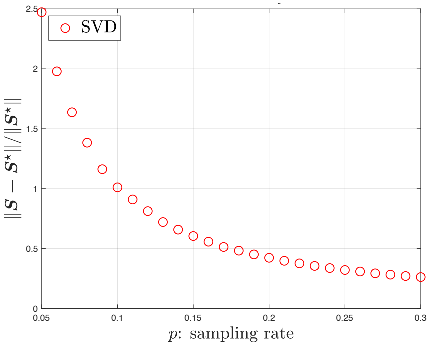
A natural SVD-based algorithm

- **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

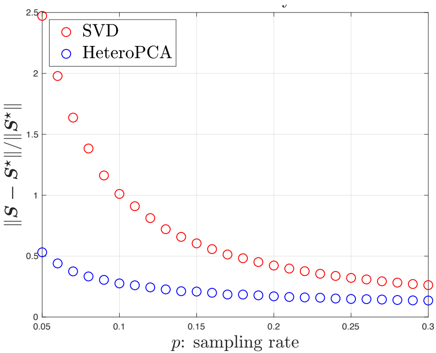
$$\text{col-space}(\,\mathbb{E}[\boldsymbol{Y}]\,) = \text{col-space}(\boldsymbol{X}) = \boldsymbol{U}^\star$$

Numerical suboptimality of SVD-based approach



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

Numerical suboptimality of SVD-based approach



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

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

Diagnosis: dagonal entries need special treatment

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

Diagnosis: dagonal entries need special treatment

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

Large bias in diagonal entries:

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

Diagnosis: dagonal entries need special treatment

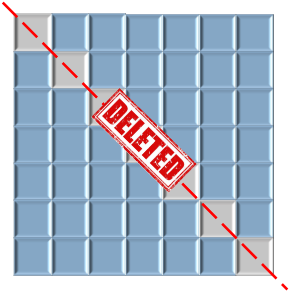
col-space$(\boldsymbol{Y})$ = eig-space$(\boldsymbol{Y}\boldsymbol{Y}^\top)$

Large bias in diagonal entries:

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

- a common issue under missing data or heteroskedastic noise

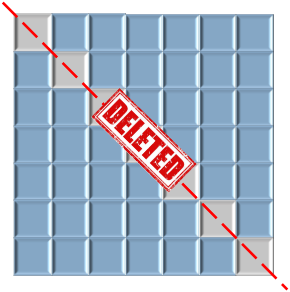
Two spectral algorithms that take care of diagonals



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

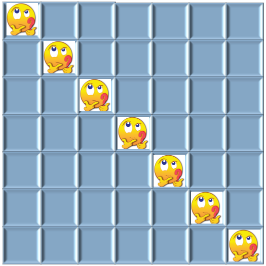
— Florescu and Perkins '16

— Montanari and Sun '18

— Zhu, Wang, Samworth '19

— Cai, Li, Chi, Poor, Chen '19

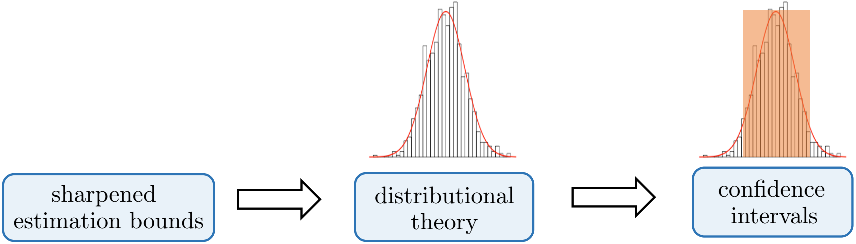
— ...



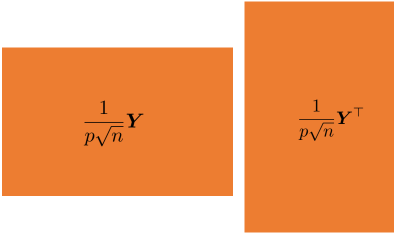
HeteroPCA (Zhang et al '18)

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

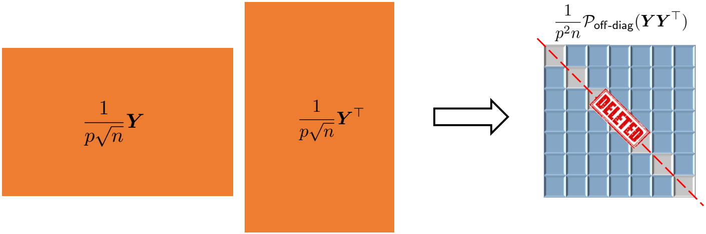
Our contributions: estimation and inference based on HeteroPCA



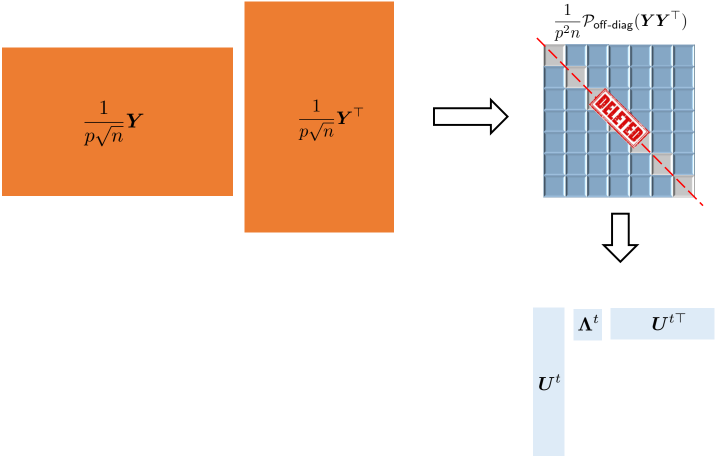
HeteroPCA (Zhang, Cai, Wu '18)



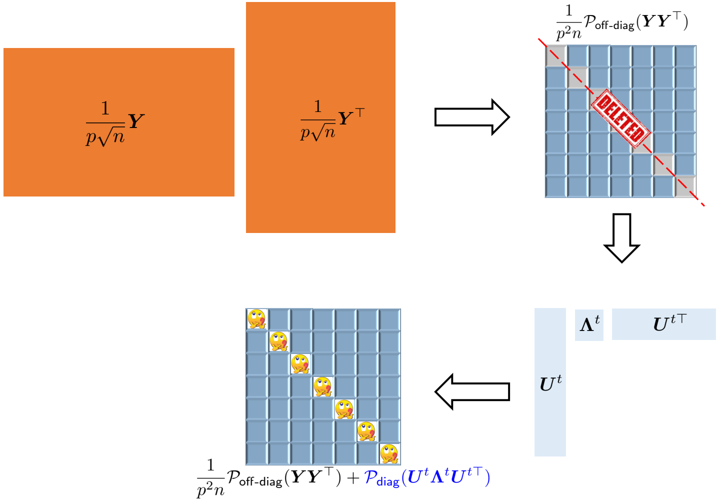
HeteroPCA (Zhang, Cai, Wu '18)



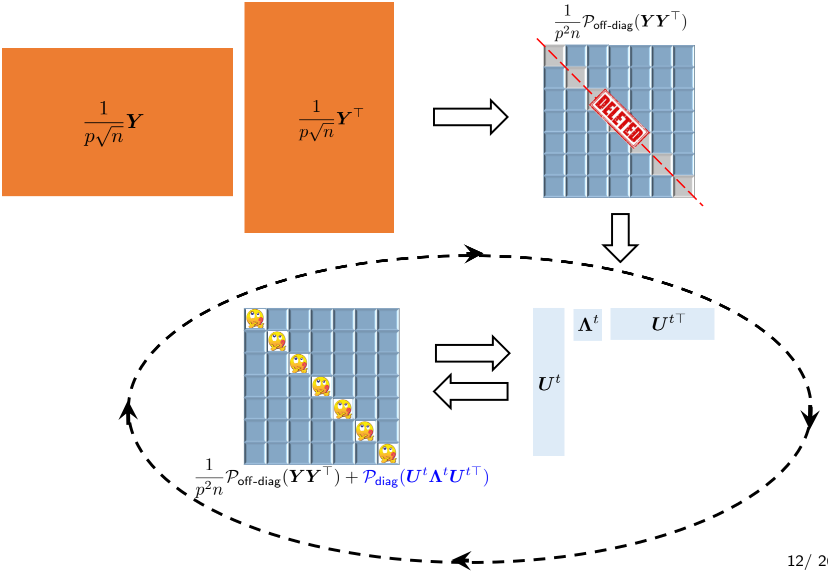
HeteroPCA (Zhang, Cai, Wu '18)



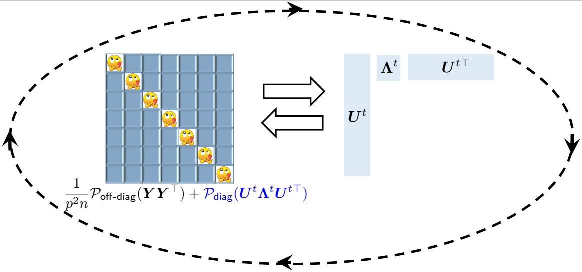
HeteroPCA (Zhang, Cai, Wu ’18)



HeteroPCA (Zhang, Cai, Wu '18)



HeteroPCA (Zhang, Cai, Wu '18)

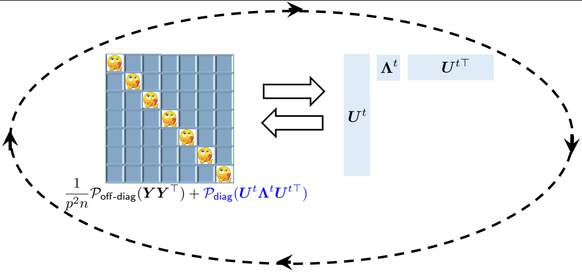


- **Initialize:** $G^0 = \frac{1}{np^2}\mathcal{P}_{\text{off-diag}}(YY^\top)$
- **Iterative update:** for $t = 0, 1, \ldots, t_0$

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

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

HeteroPCA (Zhang, Cai, Wu '18)



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

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

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

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

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

Sharpened estimation guarantees for HeteroPCA

Assumptions (omitting log factors)

- rank $r = O(1)$, incoherence $\mu = O(1)$, cond. number $\kappa = O(1)$
- 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

Theorem 1 (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 1 (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 1 (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 1 (Yan, Chen, Fan '21)

With high prob., we have

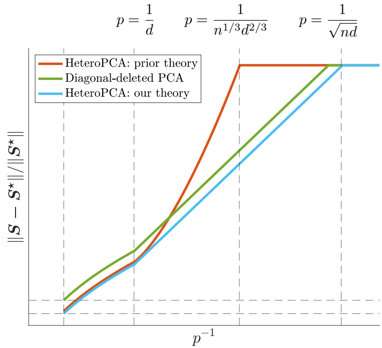
$$\|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star) - \boldsymbol{U}^\star\| \lesssim \zeta_{\mathsf{op}}, \quad \|\boldsymbol{U}\mathsf{sgn}(\boldsymbol{U}^\top \boldsymbol{U}^\star) - \boldsymbol{U}^\star\|_{2,\infty} \lesssim \frac{1}{\sqrt{d}}\,\zeta_{\mathsf{op}}$$

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

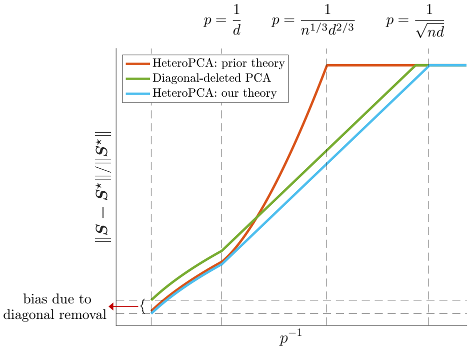
where $\zeta_{\mathsf{op}} := \frac{1}{\sqrt{nd}\,p} + \frac{\omega_{\max}^2}{p\,\lambda_r^\star}\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{np}} + \frac{\omega_{\max}}{\sqrt{\lambda_r^\star}}\sqrt{\frac{d}{np}}$

- 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)

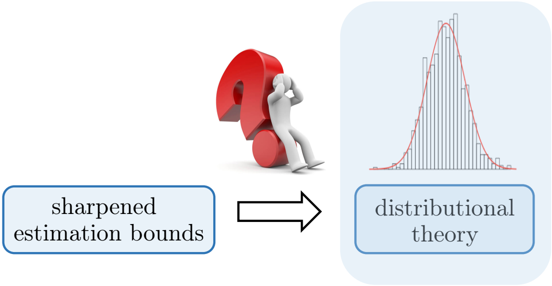
Sharpened estimation guarantees for HeteroPCA



Sharpened estimation guarantees for HeteroPCA



- diagonal-deleted PCA incurs some bias due to diagonal deletion
- $\underbrace{\text{HeteroPCA}}$ achieves bias correction via iterative refinement
  method of choice
- first $\ell_{2,\infty}$ and ℓ_∞ theory for HeteroPCA



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

Distributional theory for U

Theorem 2 (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 2 (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

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

- Each row of $\boldsymbol{U}$ is approximately Gaussian

  — nearly unbiased + tractable covariance

Distributional theory for $\boldsymbol{U}$

Theorem 2 (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

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

- 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 2 (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

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

- **Key observations:**

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

Distributional theory for S

Theorem 3 (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 3 (Yan, Chen, Fan '21)

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

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

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 3 (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

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

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

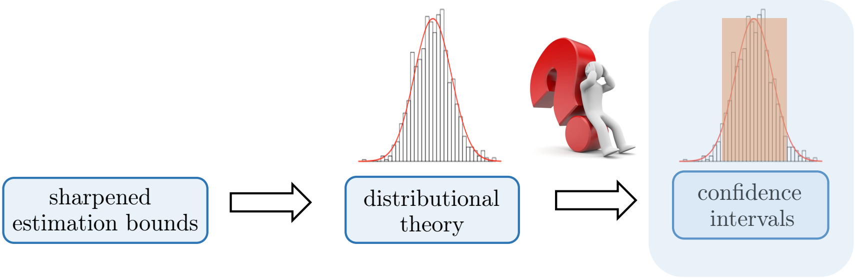
- Each entry of $\boldsymbol{S}$ is approximately Gaussian
  — nearly unbiased + tractable variance

For any $i \neq j$:

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

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

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



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 $\{\omega_i^{\star 2}\}_{i=1}^d$ in a data-driven manner[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 $\{\omega_i^{\star 2}\}_{i=1}^{d}$ in a data-driven manner[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

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

Entrywise confidence intervals for $S^\star$

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

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

Theorem 4 (Yan, Chen, Fan '21)

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

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

Entrywise confidence intervals for $S^\star$

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

$$\mathsf{CI}_{i,j}^{1-\alpha} \coloneqq \left[S_{i,j} \pm \Phi^{-1}(1-\alpha/2)\sqrt{v_{i,j}}\right]$$

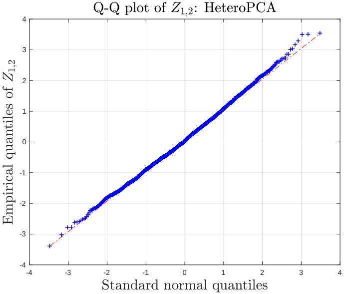
Theorem 4 (Yan, Chen, Fan '21)

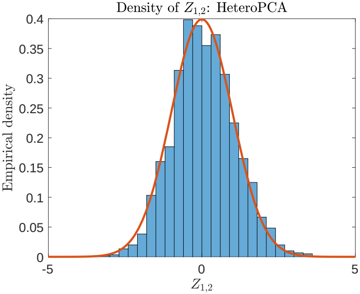
Suppose previous conditions hold and $\frac{\omega_{\max}}{\omega_{\min}} = O(1)$. *Then we have*

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

- 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

- Missing data and heterogeneous noise require special treatment
- HeteroPCA is $\underbrace{\text{provably effective}}_{\text{minimax optimal in some sense}}$ for estimation & inference

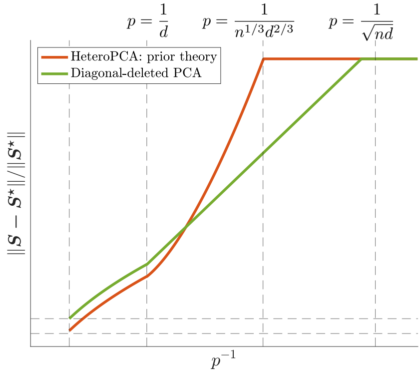
papers:

Y. Yan, Y. Chen, J. Fan, "Inference for Heteroskedastic PCA with Missing Data," 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

Backup slides

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)	$\frac{1}{\sqrt{ndp^2}} + \frac{1}{\sqrt{np}} + \frac{1}{d}$	$\sqrt{nd}$