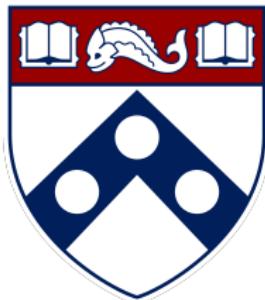


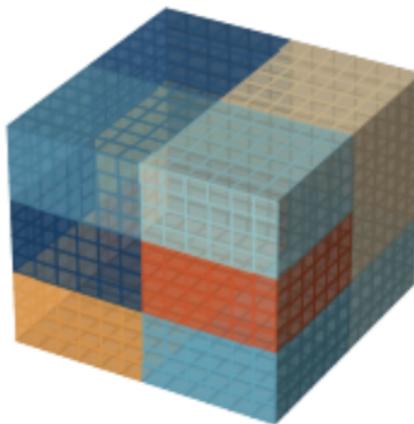
Towards more reliable tensor learning

- heteroskedastic tensor clustering and uncertainty quantification for low-rank tensors



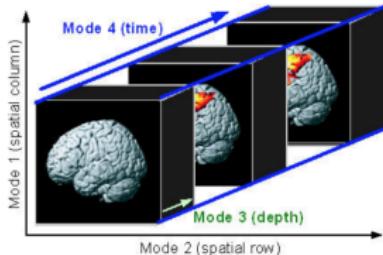
Yuchen Zhou, Wharton Statistics & Data Science

Tensors: high-order arrays



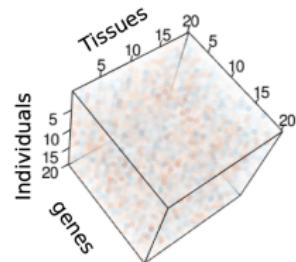
- vectors and matrices are order-1 and order-2 tensors
- $\mathcal{X} \in \mathbb{R}^{n_1 \times \dots \times n_d}$: order- d tensor

Tensors are everywhere

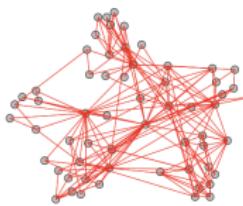


fMRI

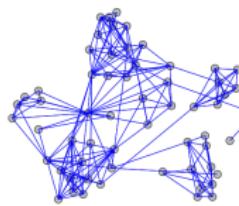
fig. credit: Lu et al. '13



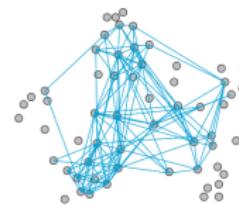
Multi-tissue gene expression
fig. credit: Wang et al. '19



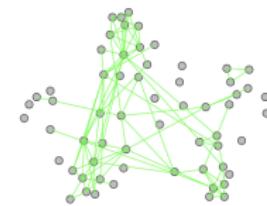
(a) Work.



(b) Lunch.



(c) Facebook.

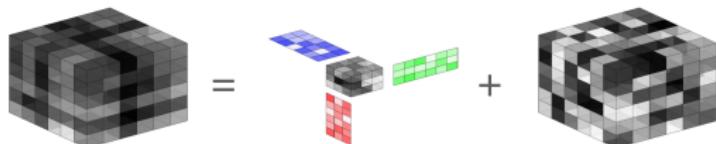


(d) Friend.

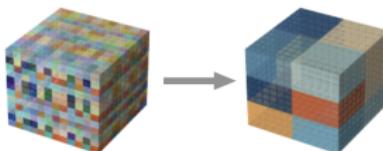
Multilayer network analysis
fig. credit: Kim and Lee '15

Various tensor problems

- Tensor SVD/PCA



- Tensor clustering



- Tensor regression

A diagram illustrating tensor regression. On the left is a blue square. To its right is an equals sign (=). To the right of the equals sign is a bracketed expression: \langle followed by a 3D tensor cube with a multi-colored, blocky pattern, a comma, a 3D tensor cube with a black and white checkered pattern, and \rangle . To the right of this expression is a plus sign (+). To the right of the plus sign is a gray square.

- ...

Foundamental challenges for tensor learning

It is NP-hard to compute

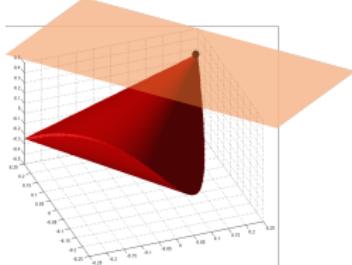
- **best low-rank approximation** of tensors
- tensor **spectral norm**
- tensor **nuclear norm**
- ...



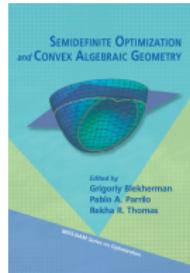
reference:

C. J. Hillar, L. H. Lim, "most tensor problems are NP-hard," *JACM*, 2013

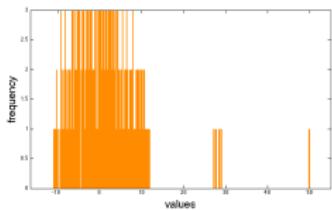
Prior arts



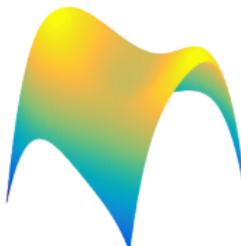
convex relaxation



sum-of-squares



spectral methods



nonconvex optimization

- De Lathauwer et al. '00
- Tomioka and Suzuki '13
- Zhou et al. '13
- Anandkumaret al. '14
- Montanari and Richard '14
- Yuan and Zhang '16
- Rauhut et al. '17
- Sun and Li '17
- Li et al. '18
- Montanari and Sun '18
- Arous et al. '19
- Raskutti et al. '19
- Chi et al. '20
- Zhang et al. '20
- Cai et al. '22
- Han et al. '22
- ...

Prior arts

Challenges remain:

- ideal assumptions (e.g., i.i.d. noise) are often violated
- uncertainty quantification

Prior arts

Challenges remain:

- ideal assumptions (e.g., i.i.d. noise) are often violated
- uncertainty quantification

Can we address these challenges using **computationally efficient** algorithms? Under **essential SNR assumptions?**

In this talk...

- **Heteroskedastic** tensor clustering
 - a detour: heteroskedastic PCA
- **Uncertainty quantification** for tensor learning

papers:

Y. Zhou, Y. Chen, "Deflated HeteroPCA: Overcoming the Curse of Ill-conditioning in Heteroskedastic PCA," under review at *Annals of Statistics*

Y. Zhou, Y. Chen, "Heteroskedastic Tensor Clustering," under review at *JRSSB*

D. Xia, A. Zhang, Y. Zhou, "Inference for Low-rank Tensors – No Need to Debias," *Annals of Statistics*, 2022

Preliminary: Tucker Low-rank Tensors

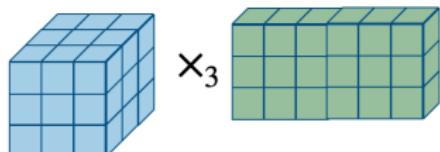
- Matrix SVD

$$X = U \Sigma V^T$$

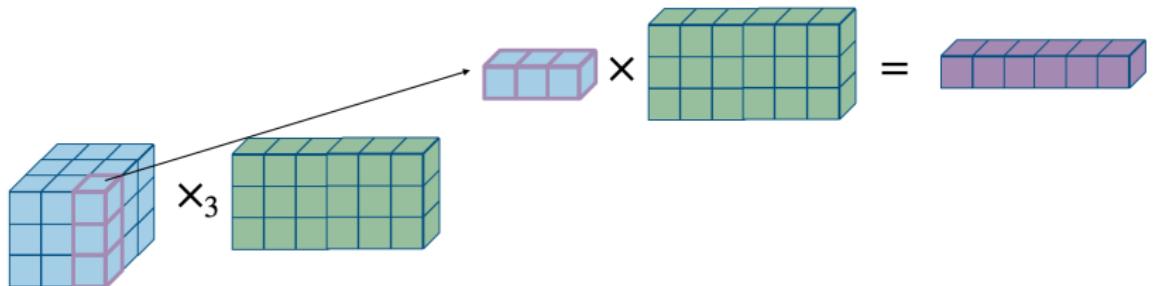
- Tucker decomposition

$$\mathcal{T} = U_1 \otimes U_2 \otimes U_3 \otimes \mathcal{S}$$

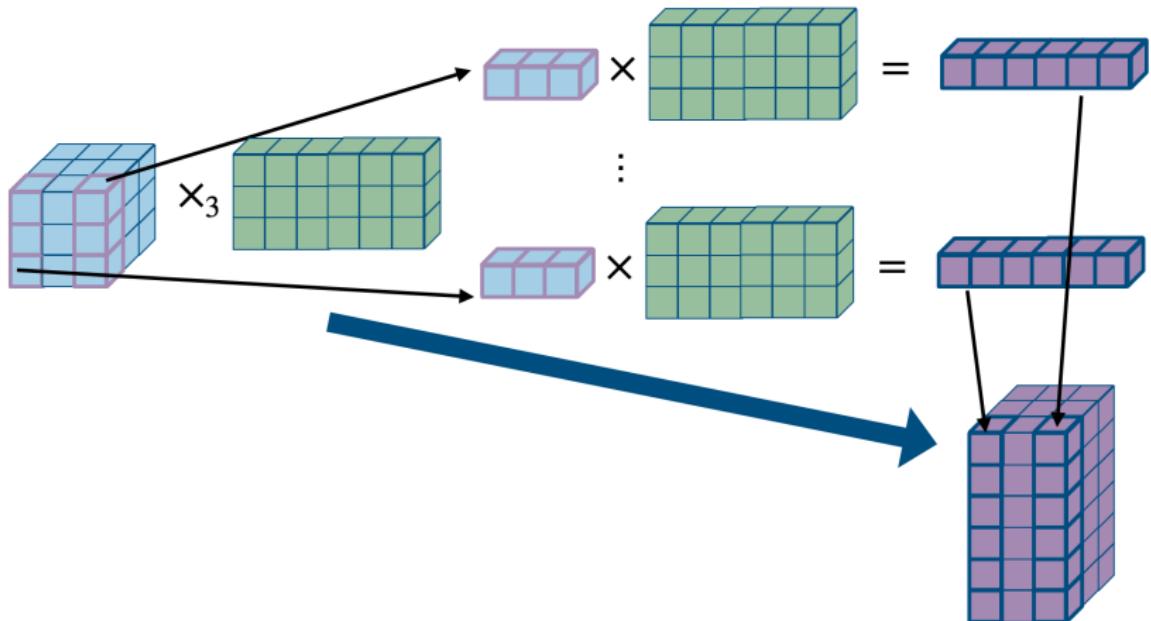
Preliminary: Tucker Low-rank Tensors



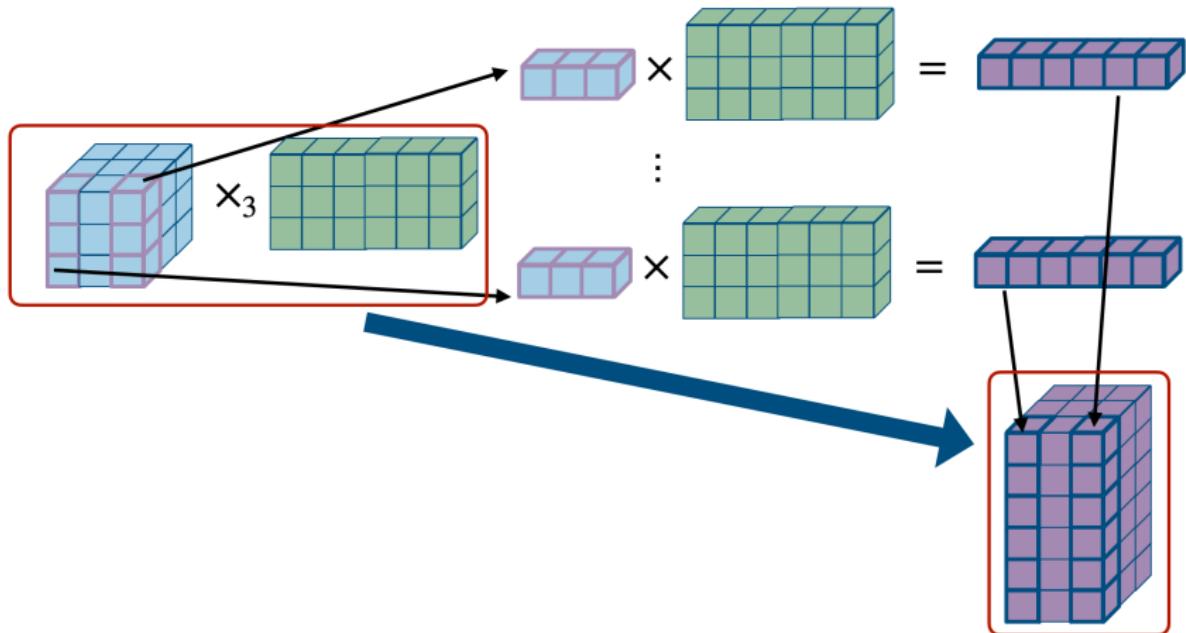
Preliminary: Tucker Low-rank Tensors



Preliminary: Tucker Low-rank Tensors



Preliminary: Tucker Low-rank Tensors

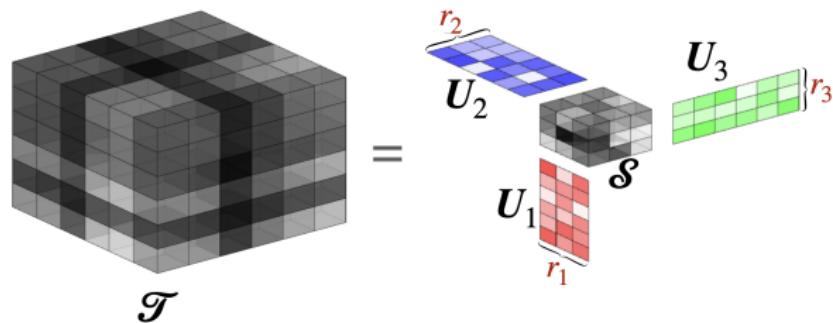


Preliminary: Tucker Low-rank Tensors

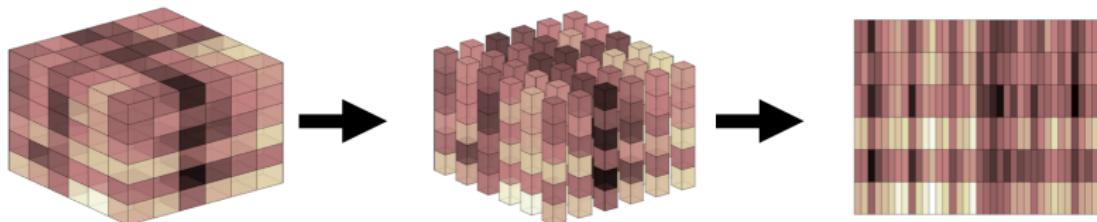
$\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ has Tucker rank (r_1, r_2, r_3) if

$$\mathcal{T} = \mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 = (\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) \cdot \mathcal{G},$$

where $\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and $\mathbf{U}_i \in \mathbb{O}_{n_i, r_i}$ for $i \in [3]$.

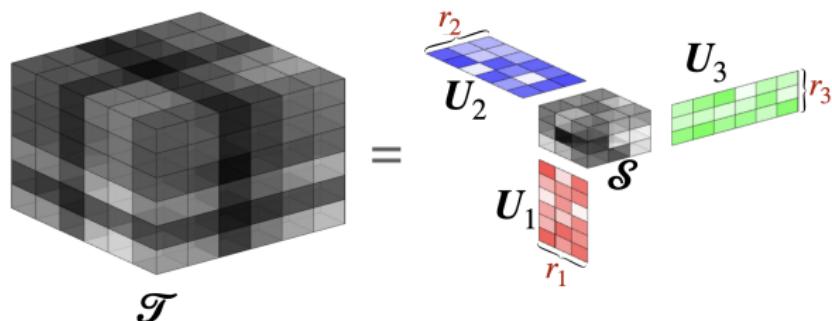


Preliminary: Tucker Low-rank Tensors



$\mathcal{M}_j(\mathcal{T}) \in \mathbb{R}^{n_j \times (n_1 n_2 n_3 / n_j)}$: *j*th matricization of \mathcal{T}

Preliminary: Tucker Low-rank Tensors



$\mathcal{M}_j(\mathcal{T}) \in \mathbb{R}^{n_j \times (n_1 n_2 n_3 / n_j)}$: *j*th matricization of \mathcal{T}

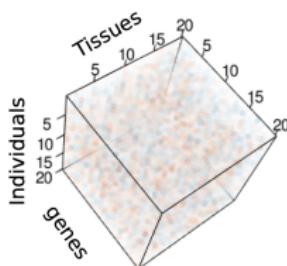
- $\text{rank}(\mathcal{M}_j(\mathcal{T})) = r_j, \quad \forall 1 \leq j \leq 3.$
- U_j is the column subspace of $\mathcal{M}_j(\mathcal{T})$.

Part 1: Heteroskedastic tensor clustering

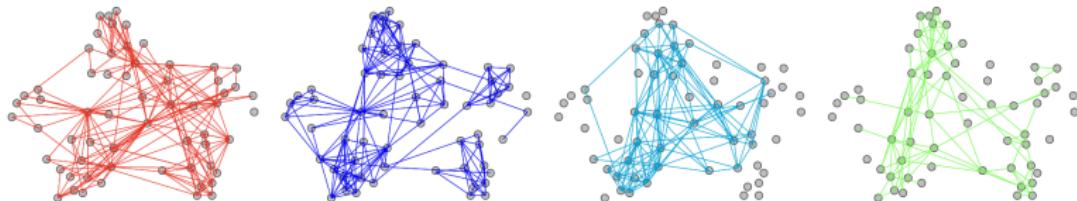


Yuxin Chen
Wharton Statistics & Data Science

Multiway clustering structure



Multi-tissue gene expression
fig. credit: Wang et al. '19



(a) Work.

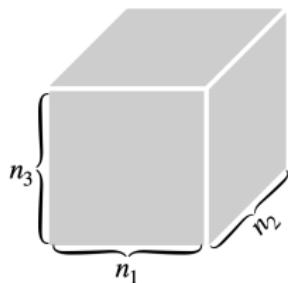
(b) Lunch.

(c) Facebook.

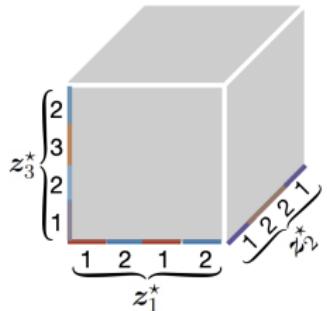
(d) Friend.

Multilayer network analysis
fig. credit: Kim and Lee '15

Tensor block model

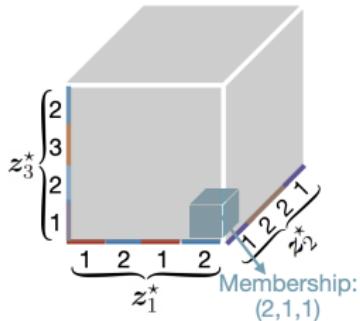


Tensor block model



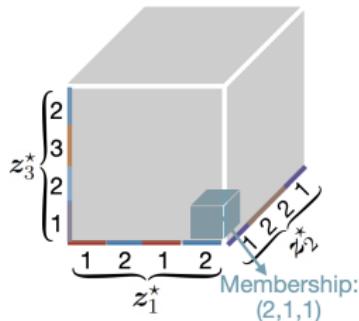
- $z_i^* \in [k_i]^{n_i}$: cluster assignment vector along the i -th mode
 - $z_{i,j}^* = \ell$ if the j th index falls within cluster ℓ

Tensor block model



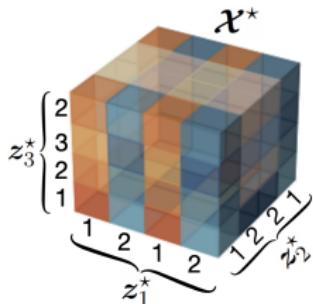
- $z_i^* \in [k_i]^{n_i}$: cluster assignment vector along the i -th mode
 - $z_{i,j}^* = \ell$ if the j th index falls within cluster ℓ
 - The membership of (i_1, i_2, i_3) is $(z_{1,i_1}^*, z_{2,i_2}^*, z_{3,i_3}^*)$

Tensor block model



- $z_i^* \in [k_i]^{n_i}$: cluster assignment vector along the i -th mode
- $\mathcal{S}^* \in \mathbb{R}^{k_1 \times k_2 \times k_3}$: block/clustering mean

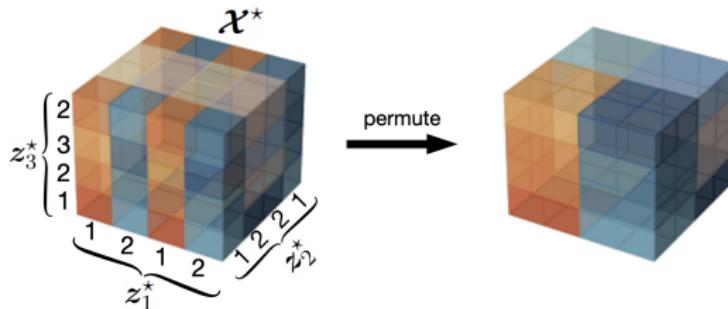
Tensor block model



- $z_i^* \in [k_i]^{n_i}$: cluster assignment vector along the i -th mode
- $\mathcal{S}^* \in \mathbb{R}^{k_1 \times k_2 \times k_3}$: block/clustering mean
- **Truth:** for all $(i_1, i_2, i_3) \in [n_1] \times [n_2] \times [n_3]$,

$$X_{i_1, i_2, i_3}^* = S_{z_{1,i_1}^*, z_{2,i_2}^*, z_{3,i_3}^*}^*.$$

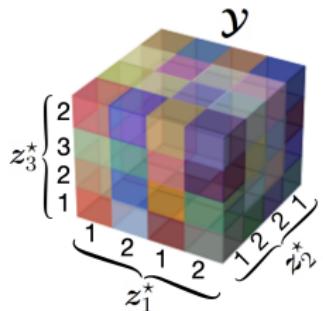
Tensor block model



- $z_i^* \in [k_i]^{n_i}$: cluster assignment vector along the i -th mode
- $\mathcal{S}^* \in \mathbb{R}^{k_1 \times k_2 \times k_3}$: block/clustering mean
- **Truth:** for all $(i_1, i_2, i_3) \in [n_1] \times [n_2] \times [n_3]$,

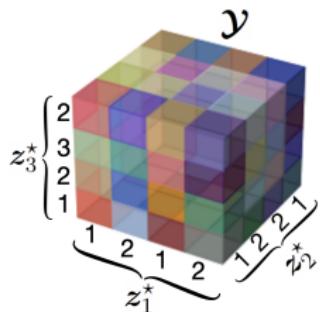
$$X_{i_1, i_2, i_3}^* = S_{z_{1,i_1}^*, z_{2,i_2}^*, z_{3,i_3}^*}^*.$$

Tensor block model



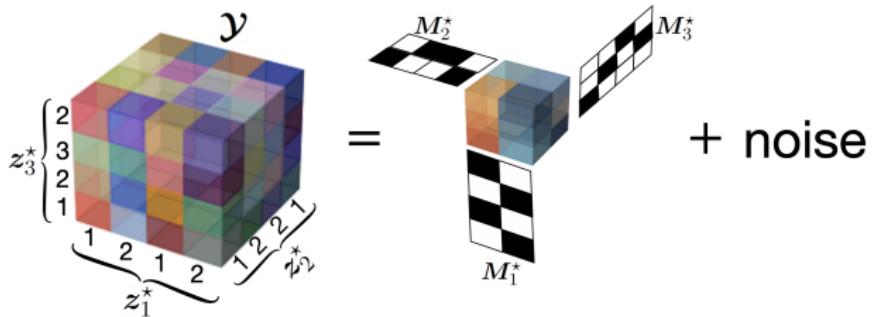
- **Noisy observations:** $\mathcal{Y} = \mathcal{X}^* + \underbrace{\mathcal{E}}_{\text{zero-mean ind. noise}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$

Tensor block model



- **Noisy observations:** $\mathcal{Y} = \mathcal{X}^* + \underbrace{\mathcal{E}}_{\text{zero-mean ind. noise}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$
- **Goal:** recover z_i^* , $i \in [3]$ from \mathcal{Y}

Tensor block model



Equivalently, we observe

$$\mathcal{Y} = \mathcal{X}^* + \underbrace{\mathcal{E}}_{\text{zero-mean ind. noise}} \in \mathbb{R}^{n_1 \times n_2 \times n_3},$$

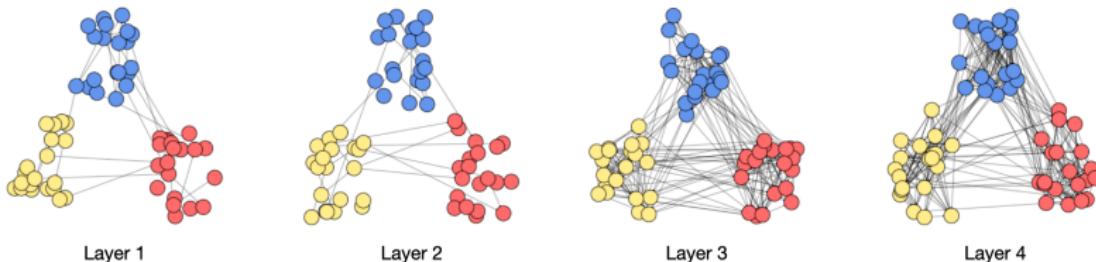
where $\mathcal{X}^* = \mathcal{S}^* \times_1 M_1^* \times_2 M_2^* \times_3 M_3^*$, where $M_i^* \in \{0, 1\}^{n_i \times k_i}$ s.t.

$$(M_i^*)_{j,\ell} = \begin{cases} 1, & \text{if } z_{i,j}^* = \ell, \\ 0, & \text{else.} \end{cases}$$

An important case: stochastic tensor block model

Stochastic tensor block model (STBM)

- generalization of bipartite stochastic block model
- each entry of \mathcal{S}^* is connection probability
- $Y_{i,j,\ell} = \begin{cases} 1, & \text{with prob. } S_{z_{1,i}^*, z_{2,j}^*, z_{3,\ell}^*}^*, \text{ if } (i, j, \ell) \text{ are connected} \\ 0, & \text{o.w.} \end{cases}$



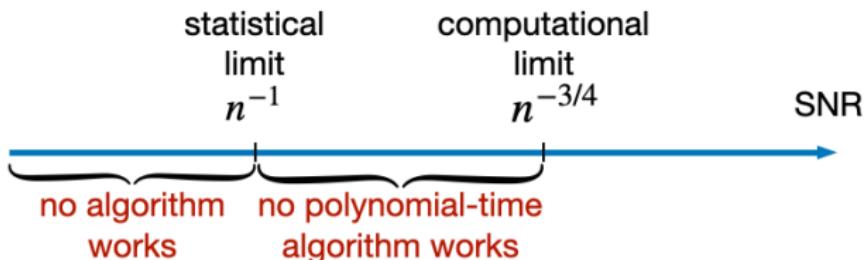
A natural approach

Least-square estimator (Wang and Zeng '19):

$$(\hat{\mathcal{S}}, \hat{z}_1, \hat{z}_2, \hat{z}_3) := \arg \min_{\substack{\mathcal{S} \in [k_1] \times [k_2] \times [k_3] \\ z_i \in [k_i]^{n_i}}} \sum_{i,j,\ell} (Y_{i,j,\ell} - S_{z_1,i, z_2,j, z_3,\ell})^2 \quad (1)$$

- statistically accurate, **computationally intractable!**

Statistical-computational gap



$\text{SNR} := \Delta_{\min}/\omega_{\max}$, where

- signal strength:

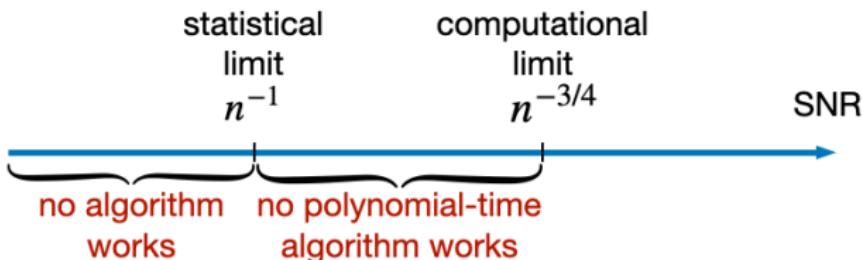
minimum separation distance along mode-1

$$\Delta_1 := \overbrace{\min_{1 \leq j_1 \neq j_2 \leq k_1} \|S_{j_1,:,:}^* - S_{j_2,:,:}^*\|_2}^{\text{minimum separation distance along mode-1}}$$

$$\Delta_{\min} := \min_{1 \leq i \leq 3} \Delta_i$$

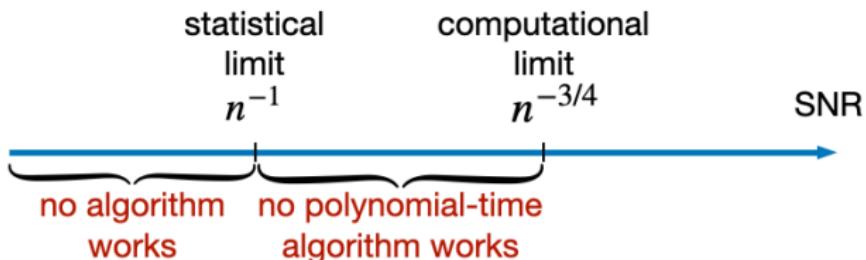
- noise level: ω_{\max}

Statistical-computational gap



- Han et al. '22: HSC + HLloyd
 - spectral clustering + iterative refinements

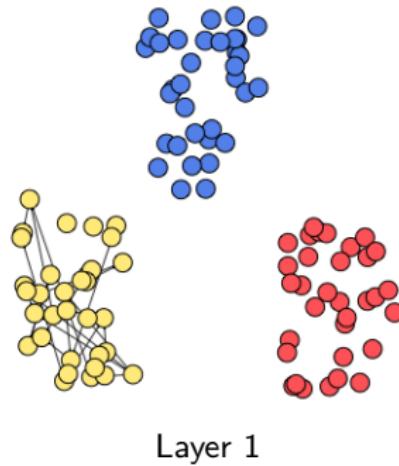
Statistical-computational gap



- Han et al. '22: HSC + HLlloyd
 - spectral clustering + iterative refinements
- Under **i.i.d. sub-Gaussian noise**, HSC + HLlloyd achieves exact clustering if SNR exceeds $n^{-3/4}$ (up to log factors)

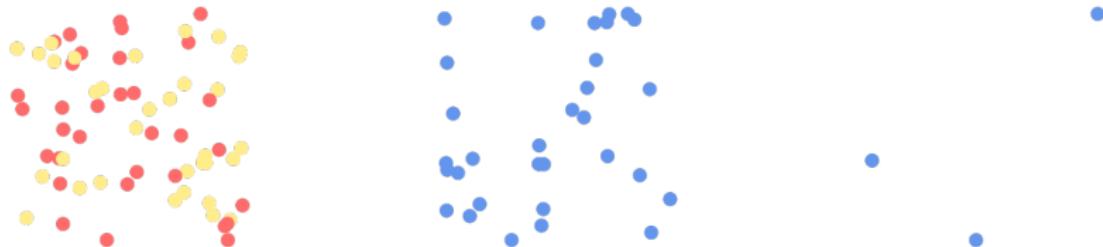
A toy example: STBM

- 100-layer networks, with 100 nodes
- 3 clusters for layers, 3 clusters for nodes
- Nodes within the yellow/red/blue community are sparsely connected.



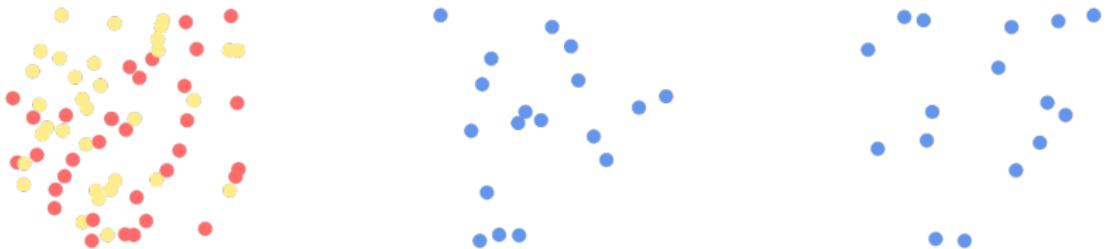
A toy example: STBM

- 100-layer networks, with 100 nodes
- 3 clusters for layers, 3 clusters for nodes
- Nodes within the yellow/red/blue community are sparsely connected.



A toy example: STBM

- 100-layer networks, with 100 nodes
- 3 clusters for layers, 3 clusters for nodes
- Nodes within the yellow/red/blue community are sparsely connected.



HSC + HLloyd

A common scenario: heteroskedastic noise

- noise variances $\{\mathbb{E}[E_{i,j,\ell}^2]\}$ are location-varying
unknown a priori
- discrete-valued observations: multi-tissue gene expression data, multilayer network data

A common scenario: heteroskedastic noise

- noise variances $\{\mathbb{E}[E_{i,j,\ell}^2]\}$ are location-varying
unknown a priori
- discrete-valued observations: multi-tissue gene expression data, multilayer network data
- fail dramatically in the face of **heteroskedastic** noise!

*How to deal with heteroskedastic noise
under essential SNR conditions?*

Spectral clustering

- Step 1: estimate “important” subspace of $\mathbf{X}_i^* = \mathcal{M}_i(\mathcal{X}^*) \Rightarrow \mathbf{U}_i$.



- Step 2: apply approximate k -means on the rows of

$$\widehat{\mathbf{B}}_1 = \mathcal{M}_1(\mathcal{Y} \underbrace{\times_1 \mathbf{U}_1 \mathbf{U}_1^\top}_{\text{denoising}} \quad \underbrace{\times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3}_{\text{dimension reduction \& denoising}}) \quad (\text{resp. } \widehat{\mathbf{B}}_2, \widehat{\mathbf{B}}_3)$$

- can be efficiently done by using k -means++!

Spectral clustering

- Step 1: estimate “important” subspace of $\mathbf{X}_i^* = \mathcal{M}_i(\mathcal{X}^*) \Rightarrow \mathbf{U}_i$.



Key challenges:

- unbalanced dimensionality
- heteroskedastic noise

SVD can be highly sub-optimal \implies HSC fails drastically!

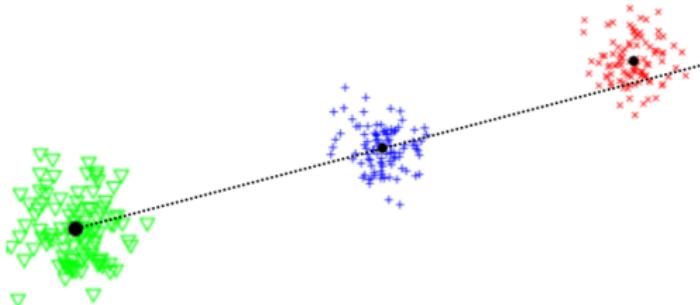
Spectral clustering

- Step 1: estimate “important” subspace of $\mathbf{X}_i^* = \mathcal{M}_i(\mathcal{X}^*) \Rightarrow U_i$.



Key challenges:

- avoid unnecessary assumptions on condition number of \mathcal{S}^*



$$\text{condi. number } \kappa := \frac{\max_i \sigma_1(\mathcal{M}_i(\mathcal{S}^*))}{\min_i \sigma_{k_i}(\mathcal{M}_i(\mathcal{S}^*))}$$

Road map

Subspace estimation problem

- .Key challenges: heteroskedastic noise
unbalanced dimensionality
not well-conditioning

.Propose Deflated HeteroPCA! 😎

More reliable solution to PCA problem!



Spectral tensor clustering

.Still need assumptions on least singular value

.Thresholded Deflated HeteroPCA + k -means!

A detour: a subspace estimation / model

$$Y = U^\star \Sigma^\star V^{\star\top} + E$$

The diagram shows a large green matrix \$Y\$ on the left, which is decomposed into four smaller matrices: \$U^\star\$ (blue), \$\Sigma^\star\$ (green), \$V^{\star\top}\$ (purple), and \$E\$ (grey). Above \$Y\$, curly braces indicate its dimensions: \$n_1\$ vertically and \$n_2\$ horizontally. The matrices \$U^\star\$, \$\Sigma^\star\$, and \$V^{\star\top}\$ are arranged horizontally, while \$E\$ is positioned to their right.

- **Ground truth:** rank- r matrix X^* with SVD ($r \ll \min\{n_1, n_2\}$)

$$\mathbf{X}^{\star} = \mathbf{U}^{\star}\boldsymbol{\Sigma}^{\star}\mathbf{V}^{\star\top} = \sum_{i=1}^r \sigma_i^{\star} \mathbf{u}_i^{\star} \mathbf{v}_i^{\star\top} \in \mathbb{R}^{n_1 \times n_2}$$

where $U^* \in \mathbb{R}^{n_1 \times r}$, $\Sigma^* = \text{diag}\{\sigma_1^*, \dots, \sigma_r^*\}$, $V^* \in \mathbb{R}^{n_2 \times r}$

A detour: a subspace estimation / model

$$Y = U^* \Sigma^* V^{*\top} + E$$

- **Ground truth:** rank- r matrix X^* with SVD ($r \ll \min\{n_1, n_2\}$)

$$\mathbf{X}^{\star} = \mathbf{U}^{\star}\boldsymbol{\Sigma}^{\star}{\mathbf{V}^{\star}}^{\top} = \sum_{i=1}^r \sigma_i^{\star} \mathbf{u}_i^{\star} {\mathbf{v}_i^{\star}}^{\top} \in \mathbb{R}^{n_1 \times n_2}$$

- **Noisy observations:** $Y = X^* + \underbrace{E}_{\text{zero-mean ind. noise}}$

A detour: a subspace estimation / model

$$Y = U^\star \Sigma^\star V^{\star\top} + E$$

The diagram shows a large green matrix \$Y\$ on the left, which is decomposed into four smaller matrices: \$U^\star\$ (blue), \$\Sigma^\star\$ (green), \$V^{\star\top}\$ (purple), and \$E\$ (grey). The dimensions of \$Y\$ are indicated by curly braces: \$n_1\$ vertically and \$n_2\$ horizontally.

- **Ground truth:** rank- r matrix X^* with SVD ($r \ll \min\{n_1, n_2\}$)

$$\mathbf{X}^{\star} = \mathbf{U}^{\star}\boldsymbol{\Sigma}^{\star}\mathbf{V}^{\star\top} = \sum_{i=1}^r \sigma_i^{\star} \mathbf{u}_i^{\star} \mathbf{v}_i^{\star\top} \in \mathbb{R}^{n_1 \times n_2}$$

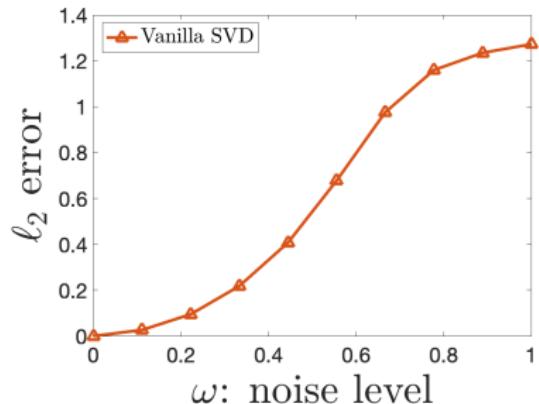
- **Noisy observations:** $Y = X^* + \underbrace{E}_{\text{zero-mean ind. noise}}$
 - **Goal:** estimate column subspace $U^* \in \mathbb{R}^{n_1 \times r}$ based on Y

Review of popular methods

$$n_1 = 100, n_2 = 10,000$$

$$r = 2, \kappa := \sigma_1^*/\sigma_r^* = 2$$

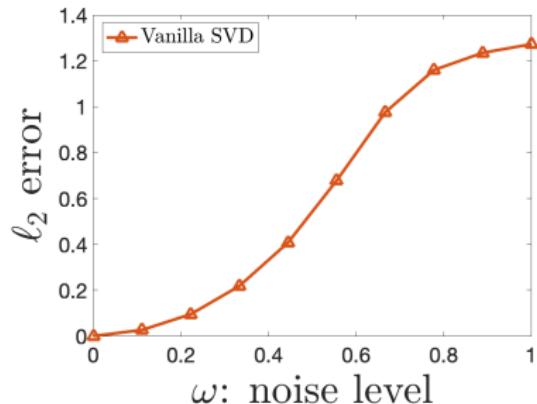
$$\omega_i \stackrel{\text{ind.}}{\sim} \text{Unif}([0, \omega]), E_{i,j} \stackrel{\text{ind.}}{\sim} \mathcal{N}(0, \omega_i^2)$$



vanilla SVD: $\mathbf{U} \leftarrow$ rank- r left singular subspace of $\mathbf{Y} = \mathbf{X}^* + \mathbf{E}$

Review of popular methods

$$\begin{aligned}n_1 &= 100, n_2 = 10,000 \\r &= 2, \kappa := \sigma_1^*/\sigma_r^* = 2 \\ \omega_i &\stackrel{\text{ind.}}{\sim} \text{Unif}([0, \omega]), E_{i,j} \stackrel{\text{ind.}}{\sim} \mathcal{N}(0, \omega_i^2)\end{aligned}$$



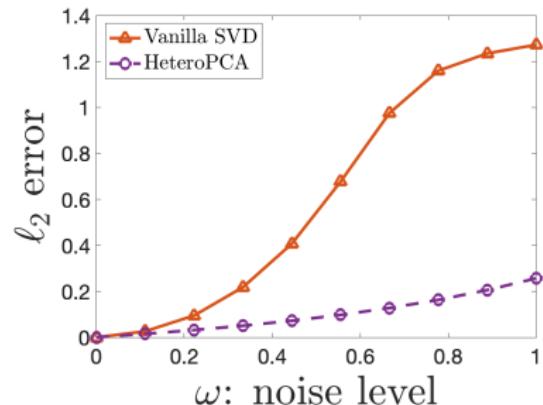
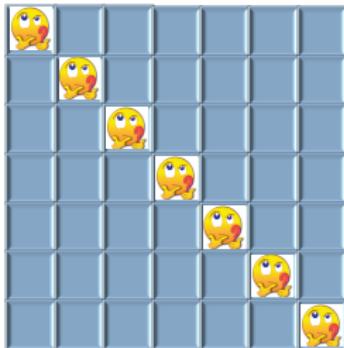
vanilla SVD: $\mathbf{U} \leftarrow \text{rank-}r \text{ left singular subspace of } \mathbf{Y} = \mathbf{X}^* + \mathbf{E}$

- often sub-optimal due to large bias in diagonal entries:

$$\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = \underbrace{\mathbf{X}^*\mathbf{X}^{*\top}}_{\checkmark} + \underbrace{\text{diag}\left\{\left[\sum_j \mathbb{E}[E_{i,j}^{*2}]\right]_{1 \leq i \leq n_1}\right\}}_{\text{potentially large diagonal matrix!}}$$

Review of popular methods

$$n_1 = 100, n_2 = 10,000$$
$$r = 2, \kappa := \sigma_1^*/\sigma_r^* = 2$$
$$\omega_i \stackrel{\text{ind.}}{\sim} \text{Unif}([0, \omega]), E_{i,j} \stackrel{\text{ind.}}{\sim} \mathcal{N}(0, \omega_i^2)$$

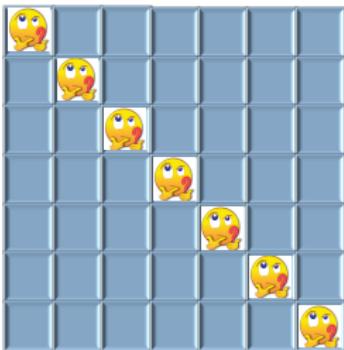
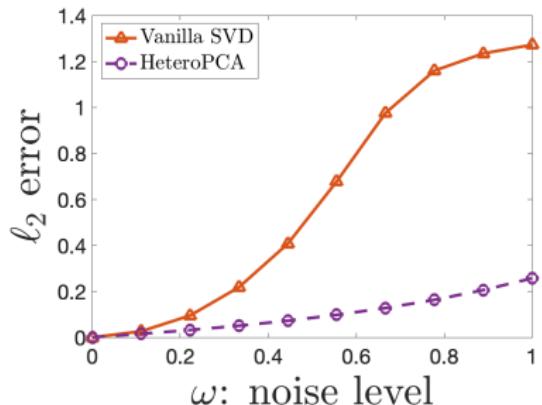


HeteroPCA (Zhang, Cai, Wu '22)

- remove $\text{diag}(\mathbf{Y}\mathbf{Y}^\top)$ & compute top- r eigen-space
- iteratively estimate $\text{diag}(\mathbf{X}\mathbf{X}^\top)$
- compute top- r eigen-space

Review of popular methods

$$n_1 = 100, n_2 = 10,000$$
$$r = 2, \kappa := \sigma_1^*/\sigma_r^* = 2$$
$$\omega_i \stackrel{\text{ind.}}{\sim} \text{Unif}([0, \omega]), E_{i,j} \stackrel{\text{ind.}}{\sim} \mathcal{N}(0, \omega_i^2)$$

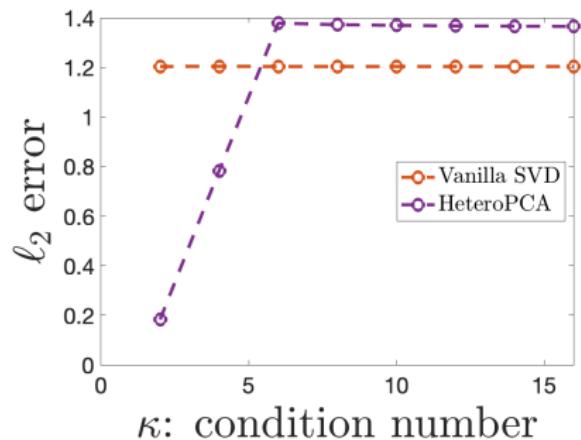


HeteroPCA (Zhang, Cai, Wu '22)

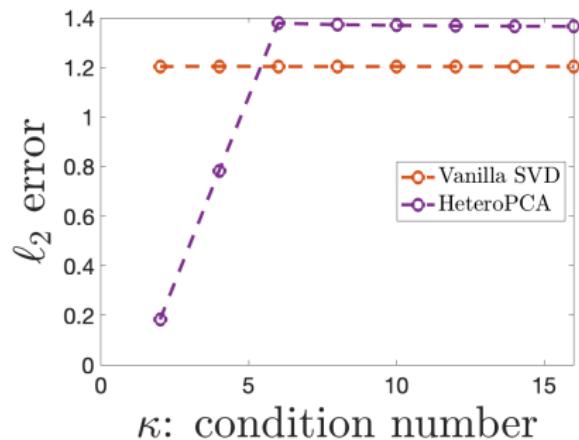
- **initialize:** $\mathbf{G}^0 = \mathcal{P}_{\text{off-diag}}(\mathbf{Y}\mathbf{Y}^\top)$
- for $t = 0, 1, \dots$
 $(\mathbf{U}^t, \Lambda^t) = \text{eigs}(\mathbf{G}^t, r)$
 $\mathbf{G}^{t+1} = \mathbf{G}^0 + \mathcal{P}_{\text{diag}}(\mathbf{U}^t \Lambda^t \mathbf{U}^{t\top})$

A curious phenomenon: curse of ill-conditioning

Somewhat surprising numerical example: $r = 2, n_1 = 200, n_2 = 40,000$



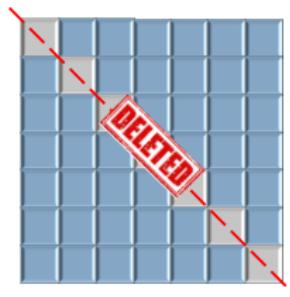
Somewhat surprising numerical example: $r = 2, n_1 = 200, n_2 = 40,000$



Previous methods degrade as condition number of X^* increases!

but this actually makes problem info-theoretically easier ...

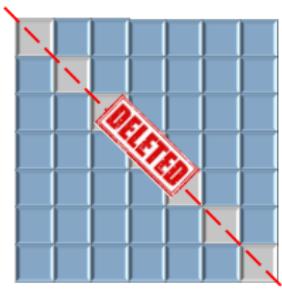
Diagonsis: influences of diagonal deletion



$$\mathbb{E} \left[\mathcal{P}_{\text{off-diag}}(\mathbf{Y}\mathbf{Y}^\top) \right] = \mathbf{X}^* \mathbf{X}^{*\top} - \underbrace{\mathcal{P}_{\text{diag}}(\mathbf{X}^* \mathbf{X}^{*\top})}_{\text{ideally negligible compared to } \sigma_r^{*2}}$$

- ideally, we hope diagonal deletion has negligible influences

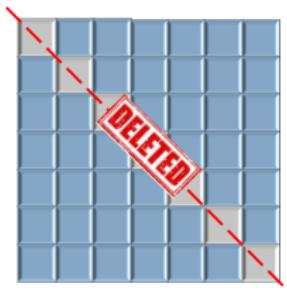
Diagonals: influences of diagonal deletion



$$\mathbb{E}[\mathcal{P}_{\text{off-diag}}(\mathbf{Y}\mathbf{Y}^\top)] = \mathbf{X}^* \mathbf{X}^{*\top} - \underbrace{\mathcal{P}_{\text{diag}}(\mathbf{X}^* \mathbf{X}^{*\top})}_{\text{but larger than } \sigma_r^{*2} \text{ if } \sigma_1^*/\sigma_r^* \text{ is too large}}$$

- ideally, we hope diagonal deletion has negligible influences
- non-negligible for ill-conditioned case though ...

Diagnosing: influences of diagonal deletion



$$\mathbb{E}[\mathcal{P}_{\text{off-diag}}(\mathbf{Y}\mathbf{Y}^\top)] = \mathbf{X}^* \mathbf{X}^{*\top} - \underbrace{\mathcal{P}_{\text{diag}}(\mathbf{X}^* \mathbf{X}^{*\top})}_{\text{but larger than } \sigma_r^{*2} \text{ if } \sigma_1^*/\sigma_r^* \text{ is too large}}$$

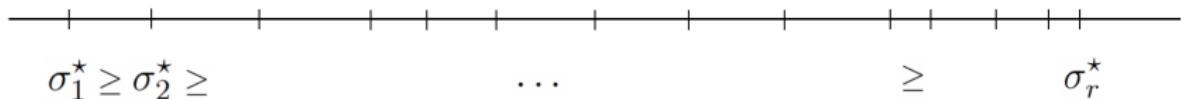
- ideally, we hope diagonal deletion has negligible influences
- non-negligible for ill-conditioned case though ...

HeteroPCA becomes ineffective in the presence of ill-conditioning!

*Can we break the curse of ill-conditioning while
accommodating widest SNR range?*

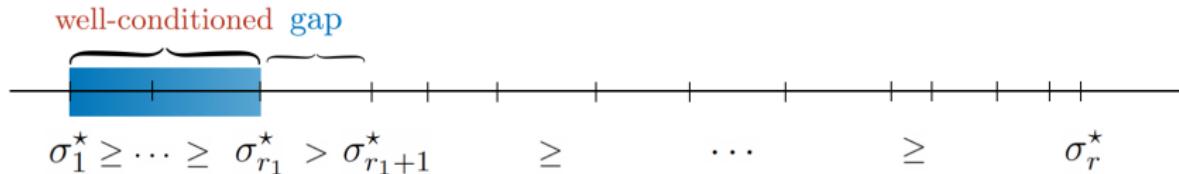
revisit HeteroPCA theory: works well if

- X^* is well-conditioned
- least singular value σ_r^* (or spectral gap) is not buried by noise



revisit HeteroPCA theory: works well if

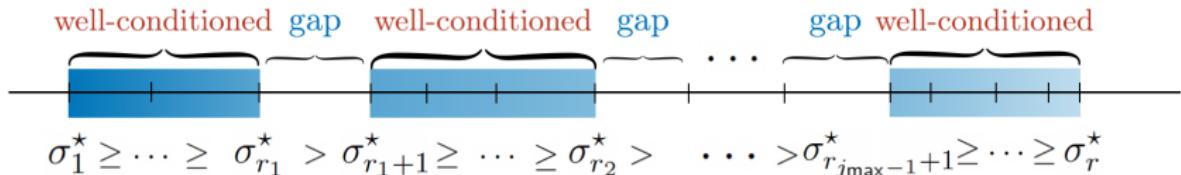
- X^* is well-conditioned
- least singular value σ_r^* (or spectral gap) is not buried by noise



solution:

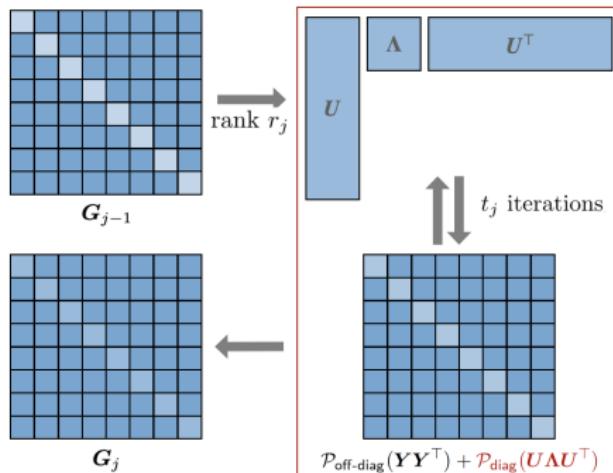
- divide eigenvalues into well-conditioned & well-separated subblocks
- estimate subblocks sequentially

Proposed algorithm: deflated-HeteroPCA



- sequentially choose ranks $r_0 = 0 < r_1 < \dots < r_{j_{\max}} = r$ s.t.
 - $\sigma_{r_{j-1}+1}^* / \sigma_{r_j}^*$ is small
 - sufficient gap between $\sigma_{r_j}^*$ and $\sigma_{r_j+1}^*$

Proposed algorithm: deflated-HeteroPCA



- sequentially choose ranks $r_0 = 0 < r_1 < \dots < r_{j_{\max}} = r$ s.t.
 - $\sigma_{r_{j-1}+1}^*/\sigma_{r_j}^*$ is small
 - sufficient gap between $\sigma_{r_j}^*$ and $\sigma_{r_j+1}^*$
- invoke HeteroPCA($\underbrace{G_{j-1}}_{\text{input}}, \underbrace{r_j}_{\text{rank}}$) to impute diagonals & obtain G_k

Proposed algorithm: deflated-HeteroPCA

- **Initialize:** $\mathbf{G}^0 = \mathcal{P}_{\text{off-diag}}(\mathbf{Y}\mathbf{Y}^\top)$, $k = 0$, $r_0 = 0$
- **Sequential updates:** while $r_k < r$

$$j = j + 1$$

select r_j in a data-driven manner

$$(\mathbf{G}_j, \mathbf{U}_j) = \text{HeteroPCA}(\underbrace{\mathbf{G}_{j-1}}_{\text{input}}, \underbrace{r_j}_{\text{rank}})$$

- **Output:** $\mathbf{U} := \mathbf{U}_j \longrightarrow$ estimate of \mathbf{U}^*

Proposed algorithm: deflated-HeteroPCA

- **Initialize:** $\mathbf{G}^0 = \mathcal{P}_{\text{off-diag}}(\mathbf{Y}\mathbf{Y}^\top)$, $k = 0$, $r_0 = 0$
- **Sequential updates:** while $r_k < r$

$$j = j + 1$$

select r_j in a data-driven manner

$$(\mathbf{G}_j, \mathbf{U}_j) = \text{HeteroPCA}(\underbrace{\mathbf{G}_{j-1}}_{\text{input}}, \underbrace{r_j}_{\text{rank}})$$

- **Output:** $\mathbf{U} := \mathbf{U}_j \longrightarrow$ estimate of \mathbf{U}^*

Select $r_j = \begin{cases} \max \mathcal{R}_j, & \text{if } \mathcal{R}_j \neq \emptyset, \\ r, & \text{otherwise.} \end{cases}$ Here,

$$\mathcal{R}_j := \{r' : r_{j-1} < r' \leq r, \underbrace{\sigma_{r_{j-1}+1}(\mathbf{G}_{j-1}) / \sigma_{r'}(\mathbf{G}_{j-1})}_{\text{well-conditioned}} \leq 4 \& \underbrace{\sigma_{r'}(\mathbf{G}_{j-1}) - \sigma_{r'+1}(\mathbf{G}_{j-1})}_{\text{gap}} \geq \sigma_{r'}(\mathbf{G}_{j-1}) / r\}.$$

Performance of Deflated-HeteroPCA

Zhou, Chen '23a (informal):

Under essential assumptions on σ_r^* and regularity assumptions,
Deflated-HeteroPCA

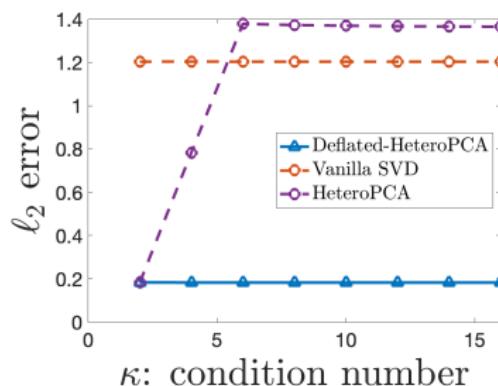
- can handle heteroskedastic noise
- enjoys near-optimal and condition-number-free guarantees

Performance of Deflated-HeteroPCA

Zhou, Chen '23a (informal):

Under essential assumptions on σ_r^* and regularity assumptions,
Deflated-HeteroPCA

- can handle **heteroskedastic noise**
- enjoys **near-optimal** and **condition-number-free** guarantees



Back to tensor clustering...

Subspace estimation problem

- .Key challenges: heteroskedastic noise
unbalanced dimensionality
not well-conditioning
- .Propose Deflated HeteroPCA! 😎
More reliable solution to PCA problem!



Spectral tensor clustering

- .Still need assumptions on least singular value
- .Thresholded Deflated HeteroPCA + k -means!

Back to tensor clustering...

Subspace estimation problem

- .Key challenges: heteroskedastic noise
unbalanced dimensionality
not well-conditioning

.Propose Deflated HeteroPCA! 

More reliable solution to PCA problem!

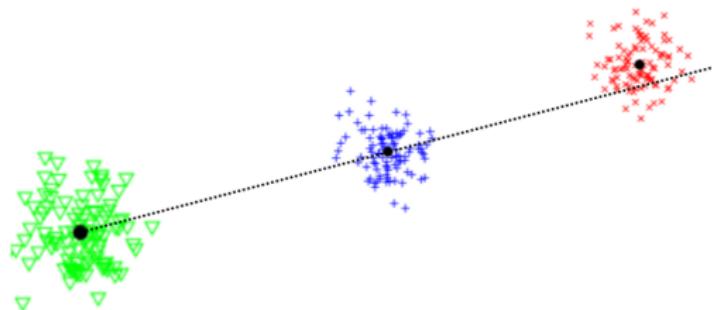


Spectral tensor clustering

- .Still need assumptions on least singular value
- .Thresholded Deflated HeteroPCA + k -means!

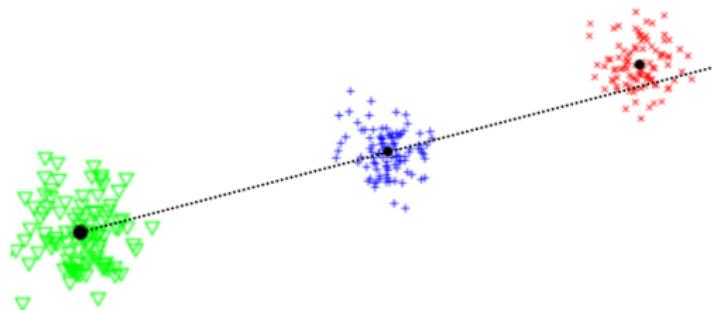
Back to tensor clustering...

- Least singular value assumption is **unnecessary** for clustering!



Back to tensor clustering...

- Least singular value assumption is **unnecessary** for clustering!



- Only large singular values matter! Add a thresholding procedure!

Thresholded Deflated-HeteroPCA(\mathbf{Y}, r, τ)

- **Initialize:** $\mathbf{G}^0 = \mathcal{P}_{\text{off-diag}}(\mathbf{Y}\mathbf{Y}^\top), j = 0, r_0 = 0$
- **Sequential updates:** while $r_j < r$ and $\sigma_{r_j+1}(\mathbf{G}_j) >$  
 $j = j + 1$
select r_j in a data-driven manner
 $(\mathbf{G}_j, \mathbf{U}_j) = \text{HeteroPCA}(\mathbf{G}_{j-1}, r_j)$
- **Output:** $\mathbf{U} := \mathbf{U}_j \longrightarrow$ estimate of \mathbf{U}^*

Proposed algorithm: High-order HeteroClustering

- Step 1: estimate “important” subspaces of $\mathbf{X}_i^* = \mathcal{M}_i(\mathcal{X}^*)$



- Obtain \mathbf{U}_i via Thresholded Deflated-HeteroPCA($\mathbf{X}_i^*, k_i, \tau$)
- Step 2: apply approximate k -means on the rows of
$$\widehat{\mathbf{B}}_1 = \mathcal{M}_1(\underbrace{\mathcal{Y} \times_1 \mathbf{U}_1 \mathbf{U}_1^\top}_{\text{denoising}} \quad \underbrace{\times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3}_{\text{dimension reduction \& denoising}}) \quad (\text{resp. } \widehat{\mathbf{B}}_2, \widehat{\mathbf{B}}_3)$$
- Optional update: HLlloyd (Han et al. '22)

Assumptions (ignoring log factors)

- dimension $n_1 \asymp n_2 \asymp n_3 \asymp n$
- **heteroskedasticity:** \mathcal{E} has indep. zero-mean entries obeying
 - $\text{Var}[E_{i,j,\ell}] \leq \omega_{\max}^2$
 - $|E_{i,j,\ell}| \lesssim \omega_{\max} n^{3/4}$ with high prob.
 - **examples:** sub-Gaussian, centered Poisson, STBM

Assumptions (ignoring log factors)

- dimension $n_1 \asymp n_2 \asymp n_3 \asymp n$
- **heteroskedasticity:** \mathcal{E} has indep. zero-mean entries obeying
 - $\text{Var}[E_{i,j,\ell}] \leq \omega_{\max}^2$
 - $|E_{i,j,\ell}| \lesssim \omega_{\max} n^{3/4}$ with high prob.
 - **examples:** sub-Gaussian, centered Poisson, STBM
- **signal-to-noise ratio (SNR):**

$$\frac{\Delta_{\min}}{\omega_{\max}} \gtrsim n^{-3/4}$$

- match the computational limit (Han et al. '22)

Assumptions (ignoring log factors)

- dimension $n_1 \asymp n_2 \asymp n_3 \asymp n$
- **heteroskedasticity:** \mathcal{E} has indep. zero-mean entries obeying
 - $\text{Var}[E_{i,j,\ell}] \leq \omega_{\max}^2$
 - $|E_{i,j,\ell}| \lesssim \omega_{\max} n^{3/4}$ with high prob.
 - **examples:** sub-Gaussian, centered Poisson, STBM
- **signal-to-noise ratio (SNR):**

$$\frac{\Delta_{\min}}{\omega_{\max}} \gtrsim n^{-3/4}$$

- match the computational limit (Han et al. '22)
- balanced cluster sizes: $|\{j \in [n_i] : (z_i^*)_j = \ell\}| \asymp n_i/k_i$
- number of clusters $k_i = O(1)$

Theoretical guarantees

Theorem 1 (Zhou, Chen '23)

Assume that either (1) or (2) is satisfied:

- (1). noise is not too spiky (e.g., $\text{Var}[E_{i,j,\ell}] \asymp \omega_{\max}^2$);
- (2). The observation model is the stochastic tensor block model.

W.h.p., HHC and HHC + HLloyd achieve exact clustering.

Theoretical guarantees

Theorem 1 (Zhou, Chen '23)

Assume that either (1) or (2) is satisfied:

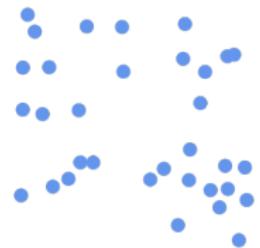
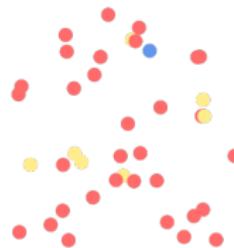
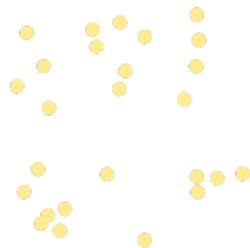
- (1). noise is not too spiky (e.g., $\text{Var}[E_{i,j,\ell}] \asymp \omega_{\max}^2$);
- (2). The observation model is the stochastic tensor block model.

W.h.p., HHC and HHC + HLloyd achieve exact clustering.

- (almost) necessary SNR condition among poly-time algorithms
- handle heteroskedastic noise & no superfluous assumptions on \mathcal{S}^*

Numerical experiments

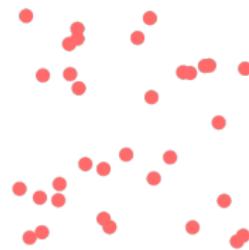
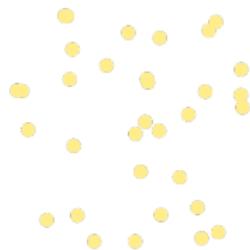
The toy example



HHC

Numerical experiments

The toy example



HHC + HLloyd

Real data example: the flight route network

- OpenFlights Airports Database: global flight information*
- top 50 airports based on the number of flights → 39 airlines†
- data $\mathcal{Y} \in \{0, 1\}^{39 \times 50 \times 50}$:

$$Y_{i,j,\ell} = \begin{cases} 1, & \text{if airline } i \text{ operates a flight route b/w airports } j, \ell, \\ 0, & \text{otherwise.} \end{cases}$$



*original data: <https://openflights.org/data>

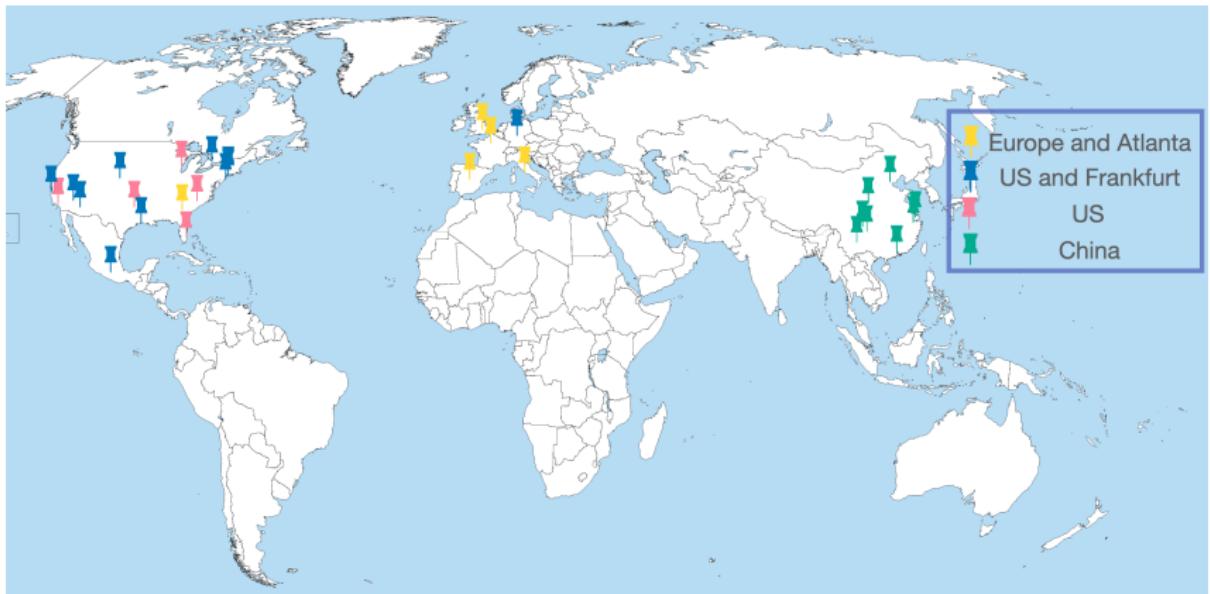
†processed data: https://github.com/RungangHLloyd/blob/master/experiment/flight_route.RData

Real data example: the flight route network

Select the clustering sizes based on BIC: $(k_1, k_2, k_3) = (5, 5, 5)$

Real data example: the flight route network

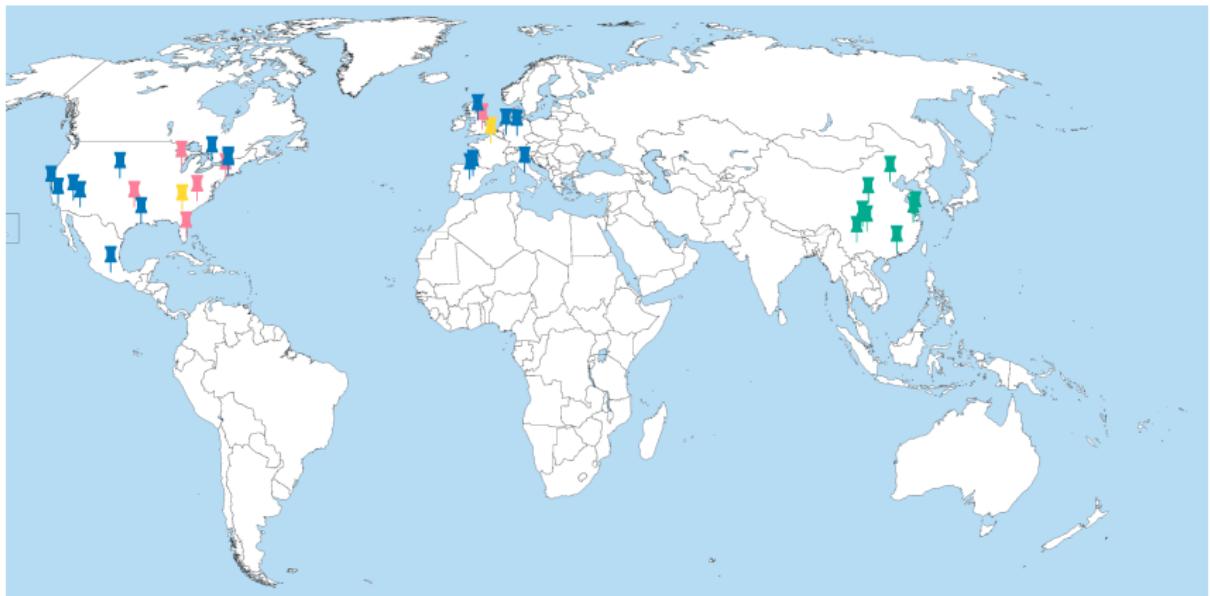
Select the clustering sizes based on BIC: $(k_1, k_2, k_3) = (5, 5, 5)$



Ours

Real data example: the flight route network

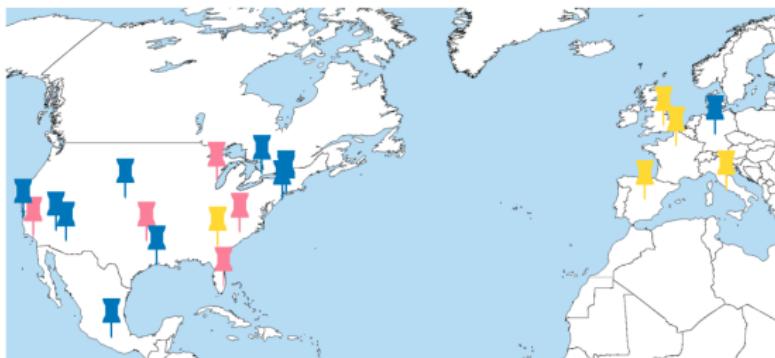
Select the clustering sizes based on BIC: $(k_1, k_2, k_3) = (5, 5, 5)$



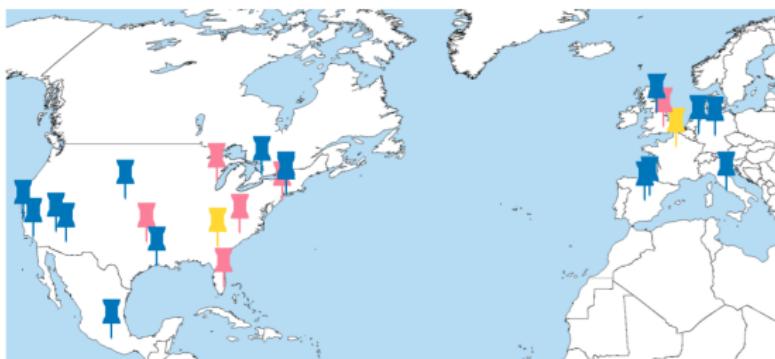
Han et al.'22

Real data example: the flight route network

Ours



Han et al.'22



A glimpse of proof highlights

- need **sharp** and **condition-number-free** guarantees w/o least singular value and condition number restrictions
- **New technique:** bounding an infinite sum of $\ell_{2,\infty}$ norms of error polynomials instead
- careful induction analyses for error polynomials

Summary

Tensor clustering: a novel method that

- can handle heteroskedastic noise
- achieves exact clustering if SNR exceeds computational limit

Summary

Tensor clustering: a novel method that

- can handle **heteroskedastic noise**
- achieves **exact clustering** if SNR exceeds **computational limit**

a detour: new method for subspace estimation

- overcome curse of ill-conditioning & **near-optimal** guarantees

Summary

Tensor clustering: a novel method that

- can handle **heteroskedastic noise**
- achieves **exact clustering** if SNR exceeds **computational limit**

a detour: new method for subspace estimation

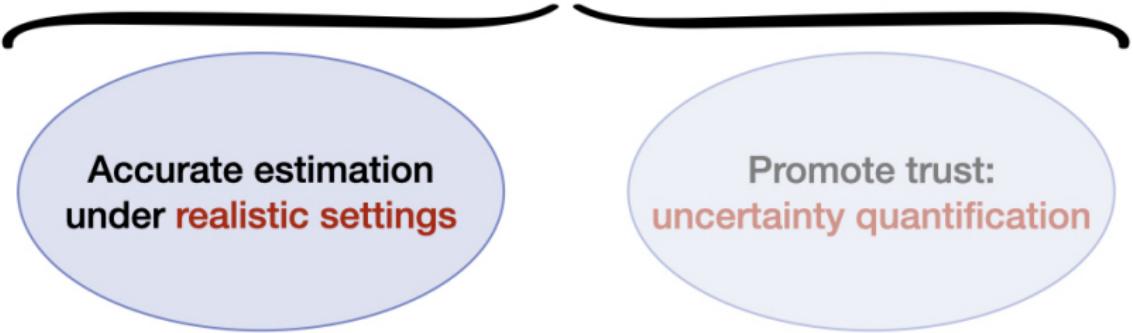
- overcome curse of ill-conditioning & **near-optimal** guarantees

papers:

Y. Zhou, Y. Chen, "Deflated HeteroPCA: Overcoming the Curse of Ill-conditioning in Heteroskedastic PCA," under review at *Annals of Statistics*

Y. Zhou, Y. Chen, "Heteroskedastic Tensor Clustering," under review at *JRSSB*

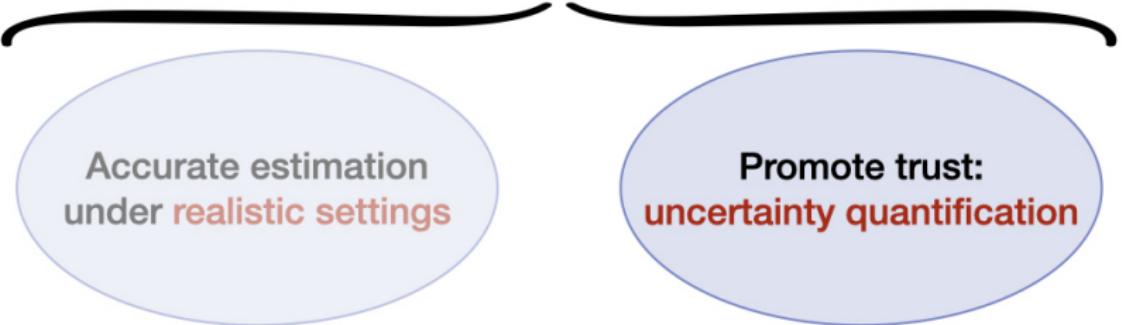
More reliable tensor learning



Accurate estimation
under **realistic settings**

Promote trust:
uncertainty quantification

More reliable tensor learning



```
graph TD; A([Accurate estimation under realistic settings]); B([Promote trust: uncertainty quantification]); A --- C; B --- C;
```

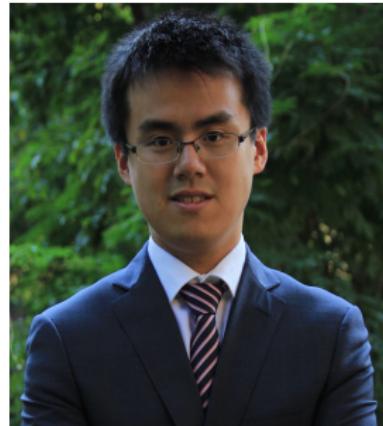
Accurate estimation
under **realistic settings**

Promote trust:
uncertainty quantification

Part 2:
Uncertainty quantification for tensor learning



Dong Xia
HKUST



Anru Zhang
Duke University

Uncertainty Quantification for Tensor Learning?

- Many methods on low-rank tensor estimation
 - tensor PCA
 - tensor clustering
 - tensor regression
 - ...

Uncertainty Quantification for Tensor Learning?

- Many methods on low-rank tensor estimation
 - tensor PCA
 - tensor clustering
 - tensor regression
 - ...
- **The statistical inference or uncertainty quantification** for low-rank tensor models remains largely unexplored!

Uncertainty Quantification for Tensor Learning?

We focus on

- inference of principal components of Tucker low-rank tensors under tensor PCA model

Uncertainty Quantification for Tensor Learning?

We focus on

- **inference of principal components** of Tucker low-rank tensors under tensor PCA model

Covered in the paper but not the talk:

- **entry-wise inference** for rank-1 tensors
- **inference of principal components** of Tucker low-rank tensors under tensor regression model

Uncertainty Quantification for Tensor Learning?

We focus on

- **inference of principal components** of Tucker low-rank tensors under tensor PCA model

Covered in the paper but not the talk:

- **entry-wise inference** for rank-1 tensors
- **inference of principal components** of Tucker low-rank tensors under tensor regression model

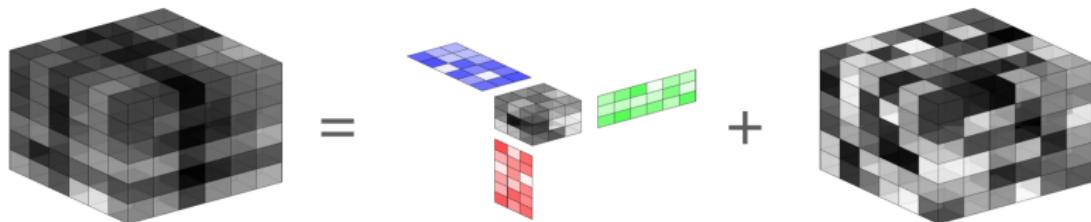
First inference results for Tucker low-rank tensors!

Inference for Tucker low-rank tensor PCA

Tensor PCA model: Observe

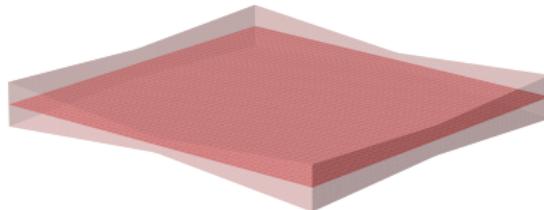
$$\mathcal{A} = \mathcal{T}^* + \mathcal{Z} \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \quad (2)$$

where the signal $\mathcal{T}^* = \mathcal{S}^* \times_1 \mathbf{U}_1^* \times_2 \mathbf{U}_2^* \times_3 \mathbf{U}_3^*$ and the noise \mathcal{Z} contains i.i.d. entries $N(0, \omega^2)$.

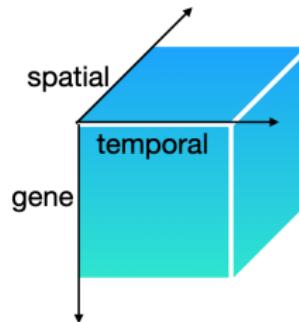


Inference for Tucker low-rank tensor PCA

Goal: Inference for $\{U_i^*\}_{i=1}^3$



- Spatial and temporal patterns of gene regulation during brain development (Liu et al. '17)



- Hidden components in Gaussian mixture models (Anandkumar et al. '14)

Estimation

- Least squares estimator:

$$\hat{\mathcal{T}} = \arg \min_{\text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3)} \|\mathcal{T} - \mathcal{A}\|_{\text{F}}^2$$

$$\iff (\widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3) = \arg \max_{\mathbf{U}_j \in \mathbb{O}_{n_j, r_j}} \|\mathcal{A} \times_1 \mathbf{U}_1^\top \times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top\|_{\text{F}}^2.$$

Highly non-convex!

- High-order orthogonal iteration (HOOI, De Lathauwer et al. '00):
alternating minimization

High-order orthogonal iteration (HOOI)

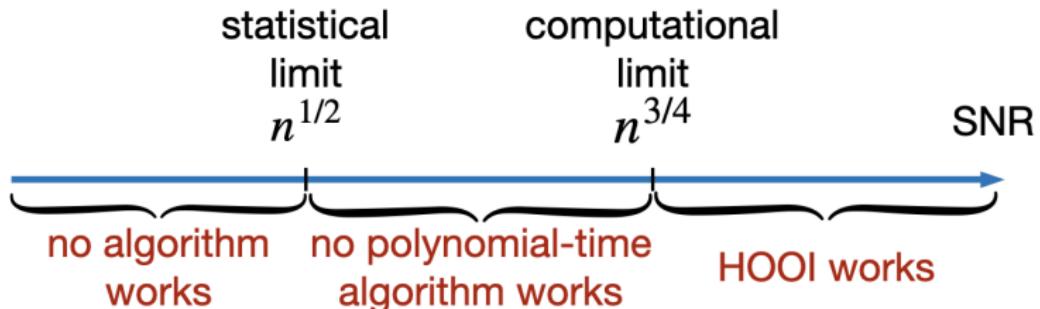
- **Spectral initialization:** $\widehat{\mathbf{U}}_i^{(0)}$, $1 \leq i \leq 3$
- **Sequential updates:** for $t = 0, 1, \dots, t_{\max}$

$$\widehat{\mathbf{U}}_1^{(t)} = \text{SVD}_{r_1} \left(\mathcal{M}_1 \left(\mathcal{A} \times_2 \widehat{\mathbf{U}}_2^{(t-1)\top} \times_3 \widehat{\mathbf{U}}_3^{(t-1)\top} \right) \right).$$

Update $\widehat{\mathbf{U}}_2^{(t)}$ and $\widehat{\mathbf{U}}_3^{(t)}$ similarly

- **Output:** $\widehat{\mathbf{U}}_i := \widehat{\mathbf{U}}_i^{(t_{\max})} \rightarrow \text{estimate of } \mathbf{U}_i^*$
 $\widehat{\mathcal{S}} := \mathcal{A} \times_1 \widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{U}}_2^\top \times_3 \widehat{\mathbf{U}}_3^\top \rightarrow \text{estimate of } \mathcal{S}^*$

Statistical-computational gap



signal strength $\lambda_{\min}^* := \lambda_{\min}(\mathcal{T}^*) = \min_j \sigma_{r_j}(\mathcal{M}_j(\mathcal{T}^*))$

noise level $\omega :=$ standard deviation of entries of \mathcal{Z}

SNR := λ_{\min}^*/ω

Inference for Tucker Low-rank Tensor PCA

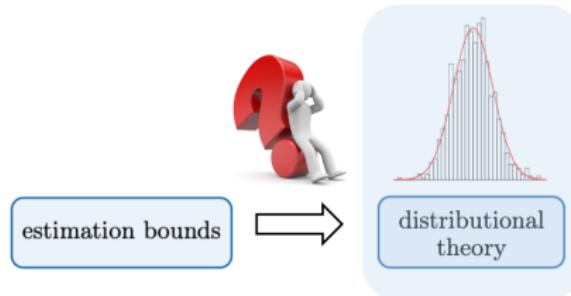
- How to make inference?

Inference for Tucker Low-rank Tensor PCA

- How to make inference?
- Most existing statistical inference methods are for **global minimizers of convex problems**
 - MLE
 - sparse linear regression
 - matrix completion
 - matrix regression

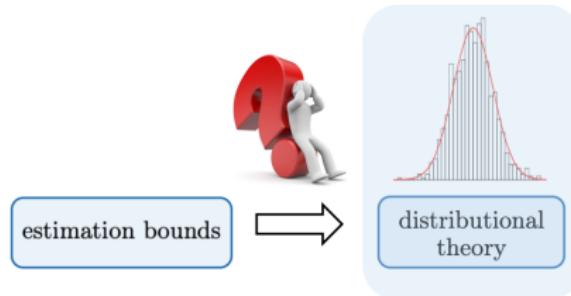
Inference for Tucker Low-rank Tensor PCA

- How to make inference?
- Most existing statistical inference methods are for **global minimizers of convex problems**
- **Challenge:** distribution theory for subspace estimates from non-convex iterations under essential SNR assumption



Inference for Tucker Low-rank Tensor PCA

- How to make inference?
- Most existing statistical inference methods are for **global minimizers of convex problems**
- **Challenge:** distribution theory for subspace estimates from non-convex iterations under essential SNR assumption



- A slightly stronger SNR condition enables us to make inference!

Two-iteration algorithm for inference

- **Input:** Rate-optimal estimators $\widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3$
- **Sequential updates:** for $t = 0, 1$

$$\widehat{\mathbf{U}}_1 = \text{SVD}_{r_1} \left(\mathcal{M}_1 \left(\mathcal{A} \times_2 \widehat{\mathbf{U}}_2^\top \times_3 \widehat{\mathbf{U}}_3^\top \right) \right)$$

$$\widehat{\mathbf{U}}_2 = \text{SVD}_{r_2} \left(\mathcal{M}_2 \left(\mathcal{A} \times_3 \widehat{\mathbf{U}}_3^\top \times_1 \widehat{\mathbf{U}}_1^\top \right) \right)$$

$$\widehat{\mathbf{U}}_3 = \text{SVD}_{r_3} \left(\mathcal{M}_3 \left(\mathcal{A} \times_1 \widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{U}}_2^\top \right) \right)$$

- **Output:** Estimates $\widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3, \widehat{\mathcal{G}} := \mathcal{A} \times_1 \widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{U}}_2^\top \times_3 \widehat{\mathbf{U}}_3^\top$

Theoretical guarantees

Theorem 2 (Xia, Zhang, Zhou '22)

Suppose $n_j \asymp n$ and $\lambda_{\min}^*/\omega \gg n^{3/4}$. Under regularity assumptions,

$$\|\sin \Theta(\widehat{\mathbf{U}}_j, \mathbf{U}_j^*)\|_{\text{F}}^2 \sim N(\text{mean}, \text{variance}) + \text{negligible terms}.$$

- mean = $n_j \omega^2 \|\Lambda_j^{\star -1}\|_{\text{F}}^2$
- variance = $\sqrt{2n_j} \omega^2 \|\Lambda_j^{\star -2}\|_{\text{F}}$
- Λ_j^* : diagonal matrix containing the singular values of $\mathcal{M}_j(\mathcal{T}^*)$.

$$\|\sin \Theta(\widehat{\mathbf{U}}_j, \mathbf{U}_j^*)\|_{\text{F}} = \sqrt{r_j - \sum_{i=1}^{r_j} \sigma_i^2 (\mathbf{U}_j^{\star \top} \widehat{\mathbf{U}}_j)}.$$

Theoretical guarantees

Theorem 2 (Xia, Zhang, Zhou '22)

Suppose $n_j \asymp n$ and $\lambda_{\min}^*/\omega \gg n^{3/4}$. Under regularity assumptions,

$$\|\sin \Theta(\widehat{\mathbf{U}}_j, \mathbf{U}_j^*)\|_{\text{F}}^2 \sim N(\text{mean}, \text{variance}) + \text{negligible terms}.$$

- mean = $n_j \omega^2 \|\Lambda_j^{\star -1}\|_{\text{F}}^2$
 - variance = $\sqrt{2n_j} \omega^2 \|\Lambda_j^{\star -2}\|_{\text{F}}$
 - Λ_j^* : diagonal matrix containing the singular values of $\mathcal{M}_j(\mathcal{T}^*)$.
-
- SNR condition $\lambda_{\min}^*/\omega \gg n^{3/4}$ is slightly stronger than the one for estimation ($\lambda_{\min}^*/\omega \gtrsim n^{3/4}$)

Theoretical guarantees

Theorem 2 (Xia, Zhang, Zhou '22)

Suppose $n_j \asymp n$ and $\lambda_{\min}^*/\omega \gg n^{3/4}$. Under regularity assumptions,

$$\|\sin \Theta(\widehat{\mathbf{U}}_j, \mathbf{U}_j^*)\|_{\text{F}}^2 \sim N(\widehat{\text{mean}}, \widehat{\text{variance}}) + \text{negligible terms.}$$

- $\widehat{\text{mean}} = n_j \widehat{\omega}^2 \|\widehat{\Lambda}_j^{-1}\|_{\text{F}}^2$
- $\widehat{\text{variance}} = \sqrt{2n_j} \widehat{\omega}^2 \|\widehat{\Lambda}_j^{-2}\|_{\text{F}}$
- $\widehat{\Lambda}_j$ and $\widehat{\omega}$: data-driven estimates for Λ_j^* and ω

Theoretical guarantees

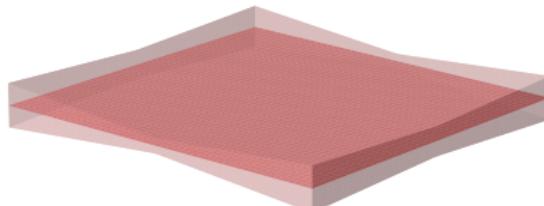
Theorem 2 (Xia, Zhang, Zhou '22)

Suppose $n_j \asymp n$ and $\lambda_{\min}^*/\omega \gg n^{3/4}$. Under regularity assumptions,

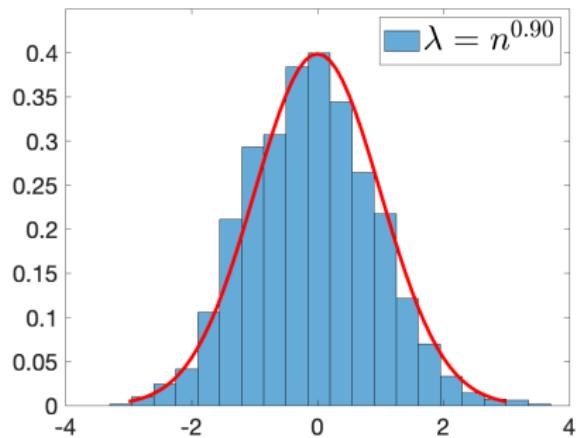
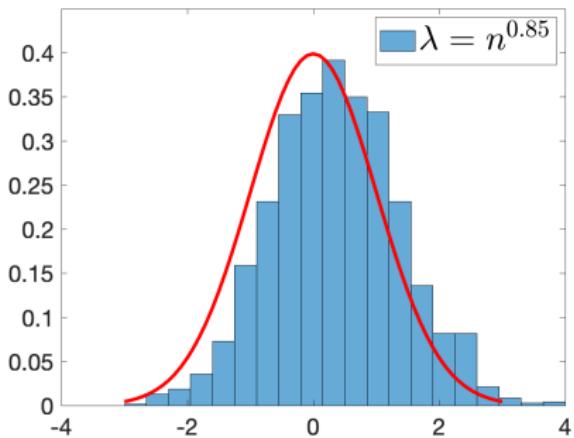
$$\|\sin \Theta(\widehat{\mathbf{U}}_j, \mathbf{U}_j^*)\|_{\text{F}}^2 \sim N(\widehat{\text{mean}}, \widehat{\text{variance}}) + \text{negligible terms}.$$

($1 - \alpha$)-level confidence region for \mathbf{U}_j^* : Let $z_\alpha = \Phi^{-1}(1 - \alpha)$,

$$\text{CR}_\alpha(\widehat{\mathbf{U}}_j) := \left\{ \mathbf{V} \in \mathbb{O}_{n_j, r_j} : \|\sin \Theta(\widehat{\mathbf{U}}_j, \mathbf{V})\|_{\text{F}}^2 \leq \widehat{\text{mean}} + z_\alpha \widehat{\text{variance}} \right\}.$$



Numerical simulations



Normal approximation of $\frac{\|\sin \Theta(\widehat{\mathbf{U}}_j, \mathbf{U}_j^\star)\|_F^2 - n_j \widehat{\omega}^2 \|\widehat{\boldsymbol{\Lambda}}_j^{-1}\|_F^2}{\sqrt{2n_j \widehat{\omega}^2 \|\widehat{\boldsymbol{\Lambda}}_j^{-2}\|_F}}$

$n_1 = n_2 = n_3 = n = 200$, $r = 3$, and $\omega = 1$

Summary

- Two-iteration algorithm for inference
- The inference of principle components can be efficiently done when a computationally feasible optimal estimate is achievable!

paper:

D. Xia, A. Zhang, Y. Zhou, "Inference for Low-rank Tensors – No Need to Debias," *Annals of Statistics*, 2022

Highlight of my research

For modern statistical problems:

- design **efficient and statistically accurate** algorithms
- establish **theoretical foundations**

High-order tensor SVD

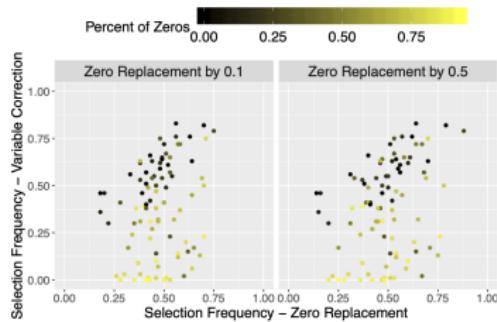
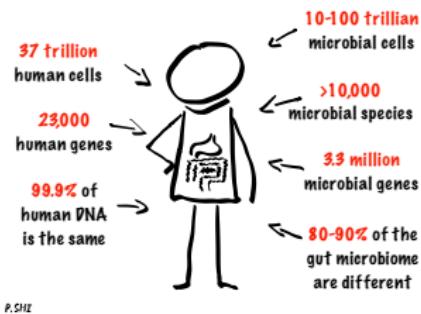
$$\mathcal{X}_{i_1 \dots i_d} = G_1 \times \mathcal{G}_2 \times \dots \times \mathcal{G}_{d-1} \times G_d$$

Diagram illustrating the decomposition of a high-order tensor \mathcal{X} into a product of matrices $G_1, \mathcal{G}_2, \dots, \mathcal{G}_{d-1}, G_d$. The dimensions of the factors are indicated below each matrix:
- G_1 : Dimension $i_1 = 1$ (represented by a 1x3 matrix with entries black, yellow, red).
- \mathcal{G}_2 : Dimension $i_2 = 2$ (represented by a 2x4 matrix with colored blocks: orange, red, yellow, pink).
- \dots
- \mathcal{G}_{d-1} : Dimension $i_{d-1} = 1$ (represented by a 1x4 matrix with colored blocks: yellow, red, orange, yellow).
- G_d : Dimension $i_d = 3$ (represented by a 3x1 matrix with colored blocks: dark red, grey, dark red).

- Novel method: tensor-train orthogonal iteration (TTOI)
- Theory: *minimax optimal* estimation error

Y. Zhou, A. Zhang, L. Zheng, Y. Wang, "Optimal High-Order Tensor SVD via Tensor-Train Orthogonal Iteration," *IEEE Trans. on Information Theory*, 2022

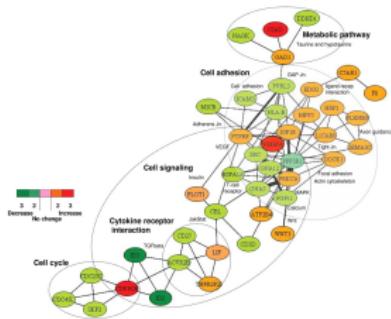
Microbial compositional data



- Goal: Identifying microbial taxa or genes associated with clinical phenotypes
- Propose **log-error-in-variable model** & new method
- Theory: **optimal guarantee** for estimation error

P. Shi, Y. Zhou, A. Zhang, "High-dimensional Log-Error-in-Variable Regression with Applications to Microbial Compositional Data Analysis," *Biometrika*, 2022

Double sparse regression



- Theory: sparse group Lasso achieves **optimal** estimation guarantees
 - Multi-objective regularizers can help!

T. Cai, A. Zhang, Y. Zhou, "Sparse Group Lasso: Optimal Sample Complexity, Convergence Rate, and Statistical Inference," *IEEE Trans. on Information Theory*, 2022

Thank you!

Assumptions (ignoring log factors)

- **heteroskedasticity:** $E'_{i,j} s$ are indep. obeying
 - $\mathbb{E}[E_{i,j}] = 0, \quad \text{Var}[E_{i,j}] \leq \omega^2$
 - $|E_{i,j}| \lesssim \omega \min \left\{ (n_1 n_2)^{1/4}, \sqrt{n_2} \right\}$ with high prob.
 - examples: sub-Gaussian, centered Poisson, centered Bernoulli

Assumptions (ignoring log factors)

- **heteroskedasticity:** $E'_{i,j}$ s are indep. obeying
 - $\mathbb{E}[E_{i,j}] = 0$, $\text{Var}[E_{i,j}] \leq \omega^2$
 - $|E_{i,j}| \lesssim \omega \min\left\{(n_1 n_2)^{1/4}, \sqrt{n_2}\right\}$ with high prob.
 - examples: sub-Gaussian, centered Poisson, centered Bernoulli
- **signal-to-noise ratio (SNR):**
$$\frac{\sigma_r^*}{\omega} \gtrsim (n_1 n_2)^{1/4} + n_1^{1/2}$$
 - necessary for existence of consistent estimators (Cai et al. '21)

Assumptions (ignoring log factors)

- **heteroskedasticity:** $E'_{i,j} s$ are indep. obeying
 - $\mathbb{E}[E_{i,j}] = 0, \quad \text{Var}[E_{i,j}] \leq \omega^2$
 - $|E_{i,j}| \lesssim \omega \min\left\{(n_1 n_2)^{1/4}, \sqrt{n_2}\right\}$ with high prob.
 - examples: sub-Gaussian, centered Poisson, centered Bernoulli
- **signal-to-noise ratio (SNR):**
$$\frac{\sigma_r^*}{\omega} \gtrsim (n_1 n_2)^{1/4} + n_1^{1/2}$$
 - necessary for existence of consistent estimators (Cai et al. '21)
- rank $r = O(1)$
- incoherence $\mu := \max\left\{\frac{n_1}{r} \|\mathbf{U}^*\|_{2,\infty}^2, \frac{n_2}{r} \|\mathbf{V}^*\|_{2,\infty}^2\right\} = O(1)$

Theoretical guarantees

Theorem 3 (Zhou, Chen '23)

With high prob., Deflated-HeteroPCA yields

$$\|UR_U - U^*\| \lesssim \zeta_{\text{op}}$$

for some rotation matrix R_U , where $\zeta_{\text{op}} = \frac{\sqrt{n_1 n_2} \omega^2}{\sigma_r^{*2}} + \frac{\sqrt{n_1} \omega}{\sigma_r^*}$

- match minimax lower bounds in Zhang et al. '22 & Cai et al. '21
- condition-number-free

Theoretical guarantees

Theorem 3 (Zhou, Chen '23)

With high prob., Deflated-HeteroPCA yields

$$\|U\mathbf{R}_U - \mathbf{U}^*\| \lesssim \zeta_{\text{op}}$$

$$\|U\mathbf{R}_U - \mathbf{U}^*\|_{2,\infty} \lesssim \frac{1}{\sqrt{n_1}} \zeta_{\text{op}} \quad (\text{fine-grained } \ell_{2,\infty})$$

for some rotation matrix \mathbf{R}_U , where $\zeta_{\text{op}} = \frac{\sqrt{n_1 n_2} \omega^2}{\sigma_r^{*2}} + \frac{\sqrt{n_1} \omega}{\sigma_r^*}$

- match minimax lower bounds in Zhang et al. '22 & Cai et al. '21
- condition-number-free

Estimates for Λ_j^* and ω^2

$\widehat{\Lambda}_j$: diagonal matrix containing the top r_j singular values of

$$\mathcal{M}_j(\mathcal{A} \times_{j+1} \widehat{\mathbf{U}}_{j+1}^\top \times_{j+2} \widehat{\mathbf{U}}_{j+2}^\top),$$

$$\widehat{\omega} = \left\| \mathcal{A} - \underbrace{\mathcal{A} \times_1 \widehat{\mathbf{U}}_1 \widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{U}}_2 \widehat{\mathbf{U}}_2^\top \times_3 \widehat{\mathbf{U}}_3 \widehat{\mathbf{U}}_3^\top}_{\approx \mathcal{T}} \right\|_{\text{F}} / \sqrt{n_1 n_2 n_3}.$$

Theorem 4 (Xia, Zhang, Zhou '22)

$$\frac{\|\sin \Theta(\widehat{\mathbf{U}}_j, \mathbf{U}_j^*)\|_{\text{F}}^2 - n_j \widehat{\omega}^2 \|\widehat{\Lambda}_j^{-1}\|_{\text{F}}^2}{\sqrt{2n_j} \widehat{\omega}^2 \|\widehat{\Lambda}_j^{-2}\|_{\text{F}}} \sim N(0, 1) + \text{negligible terms.}$$

Entry-wise Inference for Rank-1 Tensors

Model: Observe

$$\mathcal{A} = \mathcal{T}^* + \mathcal{Z}, \quad \mathcal{T}^* = \lambda^* \cdot \mathbf{u}^* \otimes \mathbf{v}^* \otimes \mathbf{w}^*.$$

Here, $\mathbf{u}^* \in \mathbb{S}^{n_1-1}$, $\mathbf{v}^* \in \mathbb{S}^{n_2-1}$, $\mathbf{w}^* \in \mathbb{S}^{n_3-1}$, the singular value $\lambda^* > 0$, and $\mathcal{Z} \stackrel{i.i.d.}{\sim} N(0, \omega^2)$.

Goal: Inference for $T_{i,j,k}^*$

Entry-wise Inference for Rank-1 Tensors

Theorem 5

Suppose $\lambda^*/\omega \gg n^{3/4}$. Apply HOOI with $t_{\max} \geq C_1 \log n$. If $|u_i^*|, |v_j^*|, |w_k^*| \ll \min\{\lambda^*/(n\omega), 1\}$, then

$$\left(\frac{\lambda}{\omega}(\hat{u}_i - u_i^*), \frac{\lambda}{\omega}(\hat{v}_j - v_j^*), \frac{\lambda}{\omega}(\hat{w}_k - w_k^*) \right)^\top \xrightarrow{\text{d}} N(0, I_3) \quad \text{as } p \rightarrow \infty.$$

If $\omega/\lambda^* \ll |u_i^*|, |v_j^*|, |w_k^*| \ll \min\{\lambda^*/(n\omega), 1/\sqrt{\log n}\}$, then

$$\frac{\hat{T}_{ijk} - \mathcal{T}_{ijk}}{\hat{\omega} \sqrt{\hat{u}_i^2 \hat{v}_j^2 + \hat{v}_j^2 \hat{w}_k^2 + \hat{w}_k^2 \hat{u}_i^2}} \xrightarrow{\text{d}} N(0, 1) \quad \text{as } n \rightarrow \infty.$$

Entry-wise Inference for Rank-1 Tensors

- upper bound conditions on $|u_i^*|, |v_j^*|, |w_k^*|$ are weaker than the incoherence condition
- lower bound condition is essential for asymptotic normality: if $u_i^* = v_j^* = w_k^* = 0$, then

$$\frac{\lambda^2 \widehat{T}_{ijk}}{\omega^3} \xrightarrow{d} G_1 G_2 G_3 \text{ as } n \rightarrow \infty, \quad (G_1, G_2, G_3)^\top \sim N(0, I_3).$$

- Asymptotic $(1 - \alpha)$ confidence interval for T_{ijk} :
 $[\widehat{T}_{ijk} - z_{\alpha/2} \sigma \sqrt{\widehat{u}_i^2 \widehat{v}_j^2 + \widehat{v}_j^2 \widehat{w}_k^2 + \widehat{w}_k^2 \widehat{u}_i^2}, \widehat{T}_{ijk} + z_{\alpha/2} \sigma \sqrt{\widehat{u}_i^2 \widehat{v}_j^2 + \widehat{v}_j^2 \widehat{w}_k^2 + \widehat{w}_k^2 \widehat{u}_i^2}]$.