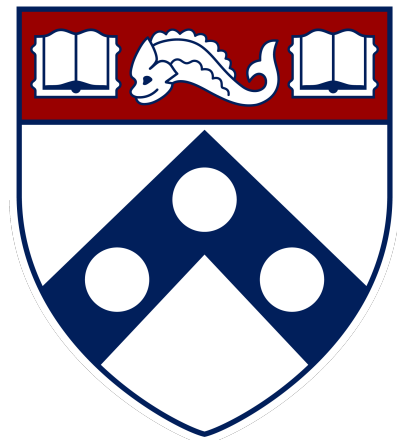
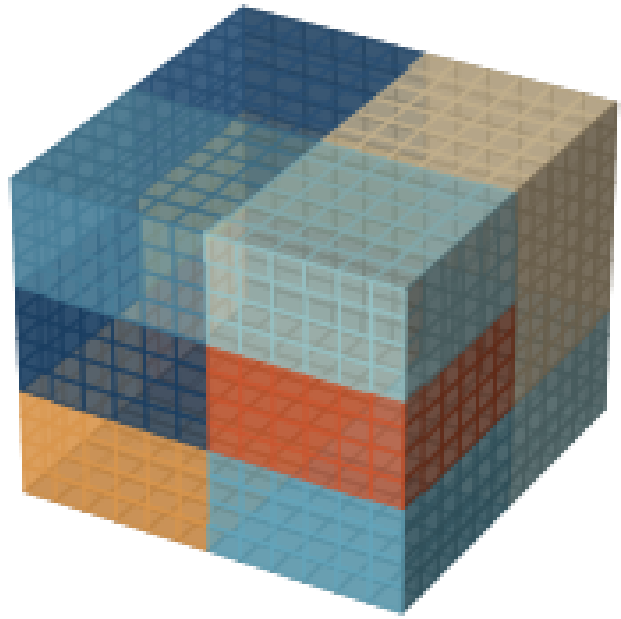
Towards more reliable tensor learning

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



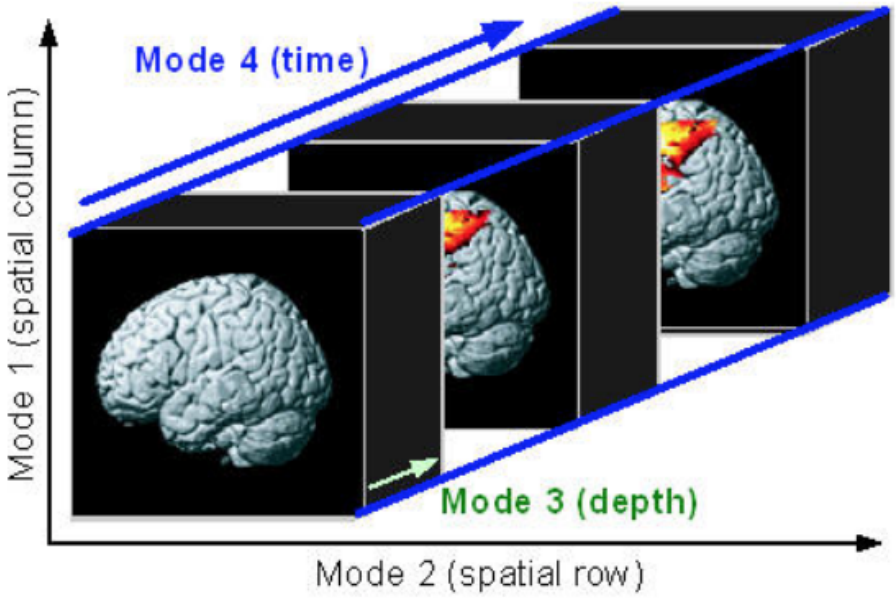
Yuchen Zhou, Wharton Statistics & Data Science

Tensors: high-order arrays



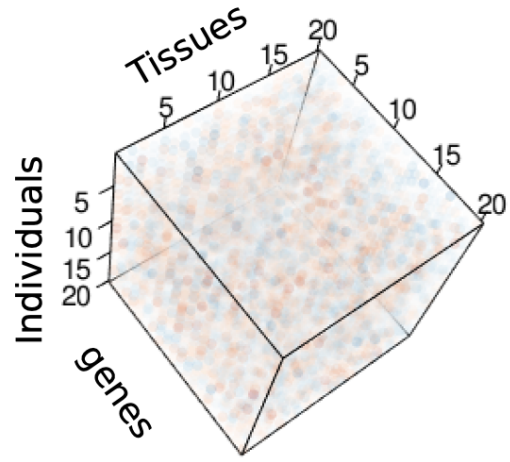
- $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$: order-$d$ tensor
- vectors and matrices are order-1 and order-2 tensors

Tensors are everywhere



fMRI

fig. credit: Lu et al. '13



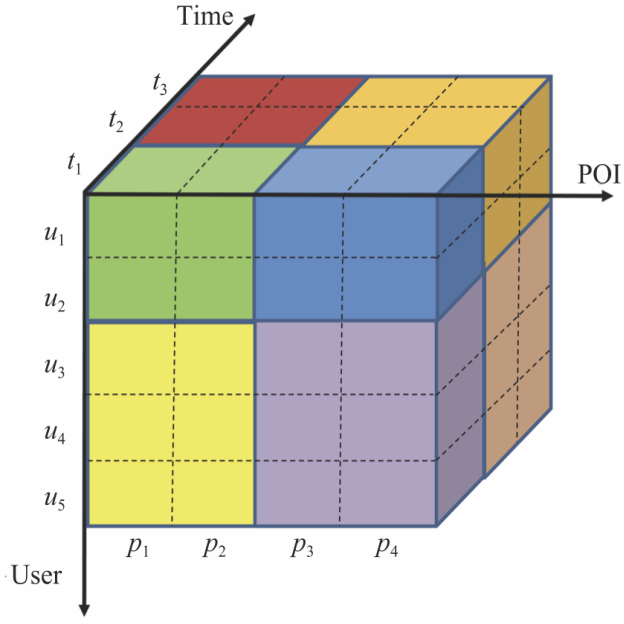
Multi-tissue gene expression

fig. credit: Wang et al. '19

Tensors are everywhere

Multilayer network analysis

fig. credit: Domenico et al. '15



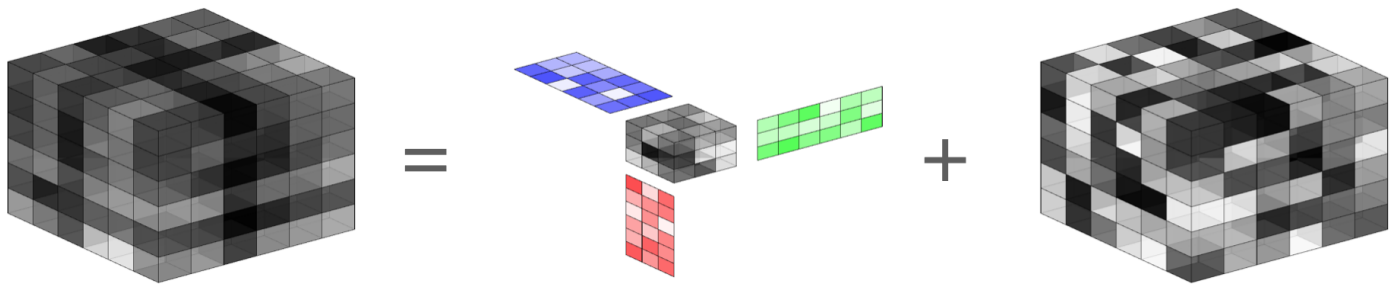
Recommender system

fig. credit: Luan et al. '17

Various tensor problems

— *Montanari and Richard '14, Zhang and Xia '18, Chen '19*

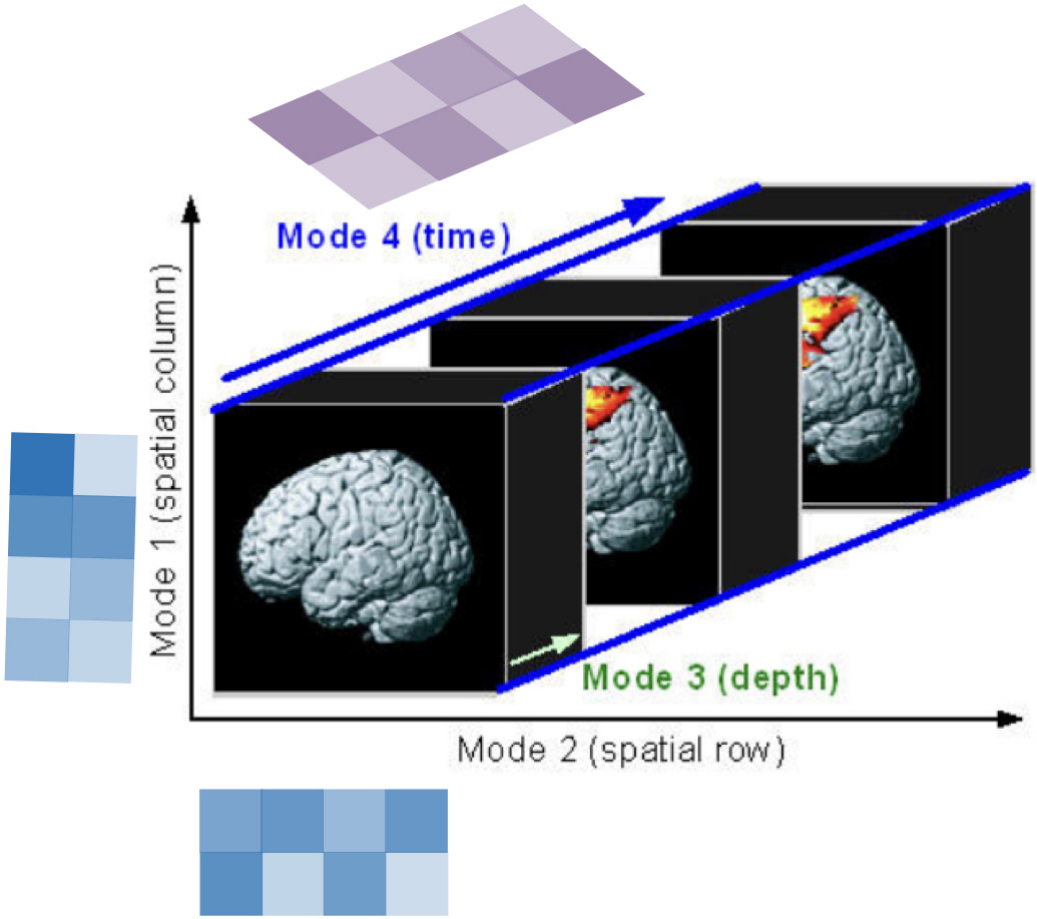
Tensor SVD/PCA



Various tensor problems

— *Montanari and Richard '14, Zhang and Xia '18, Chen '19*

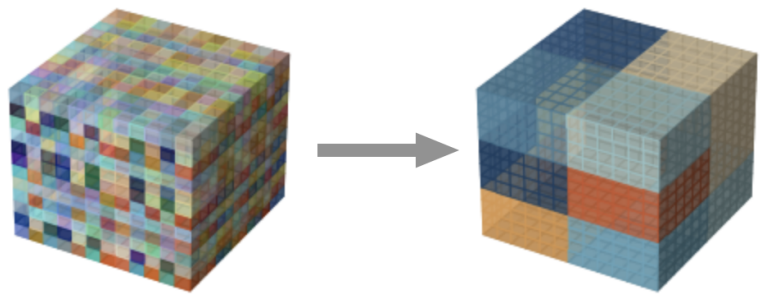
Tensor SVD/PCA



Various tensor problems

— *Wang and Zeng '19, Chi et al. '20, Han et al. '22*

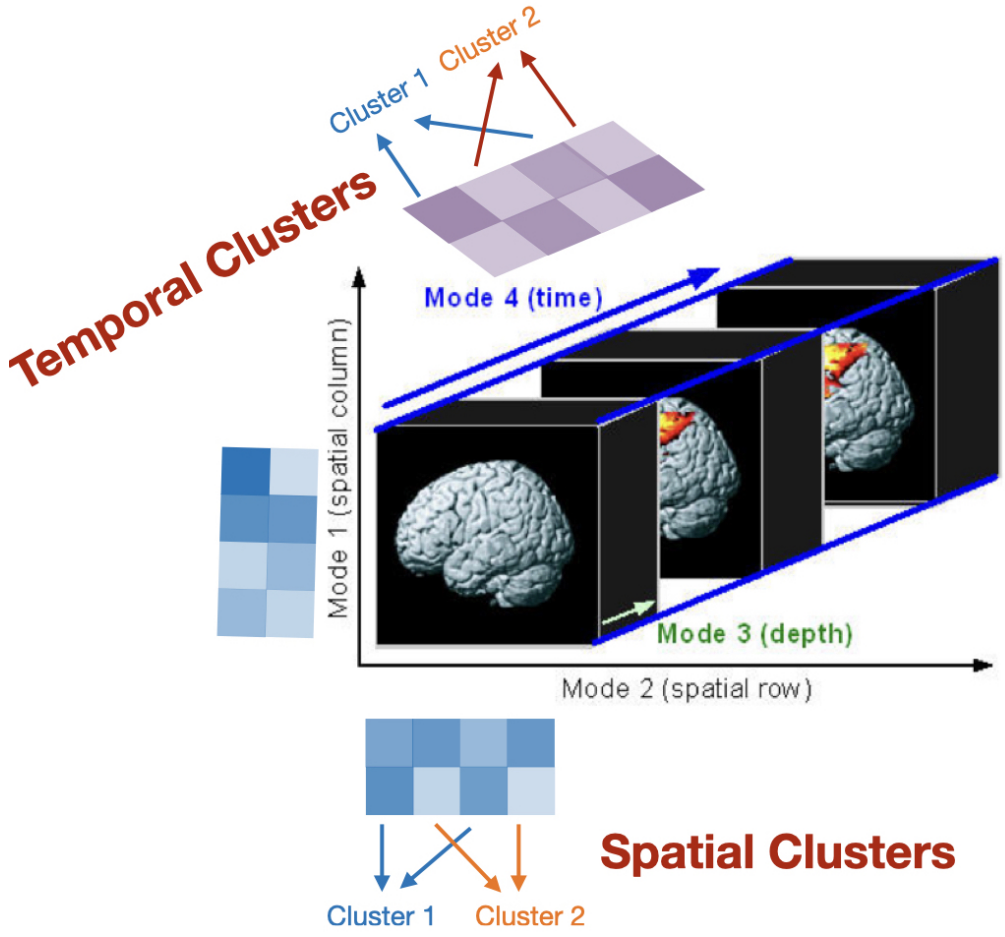
Tensor clustering



Various tensor problems

— *Wang and Zeng '19, Chi et al. '20, Han et al. '22*

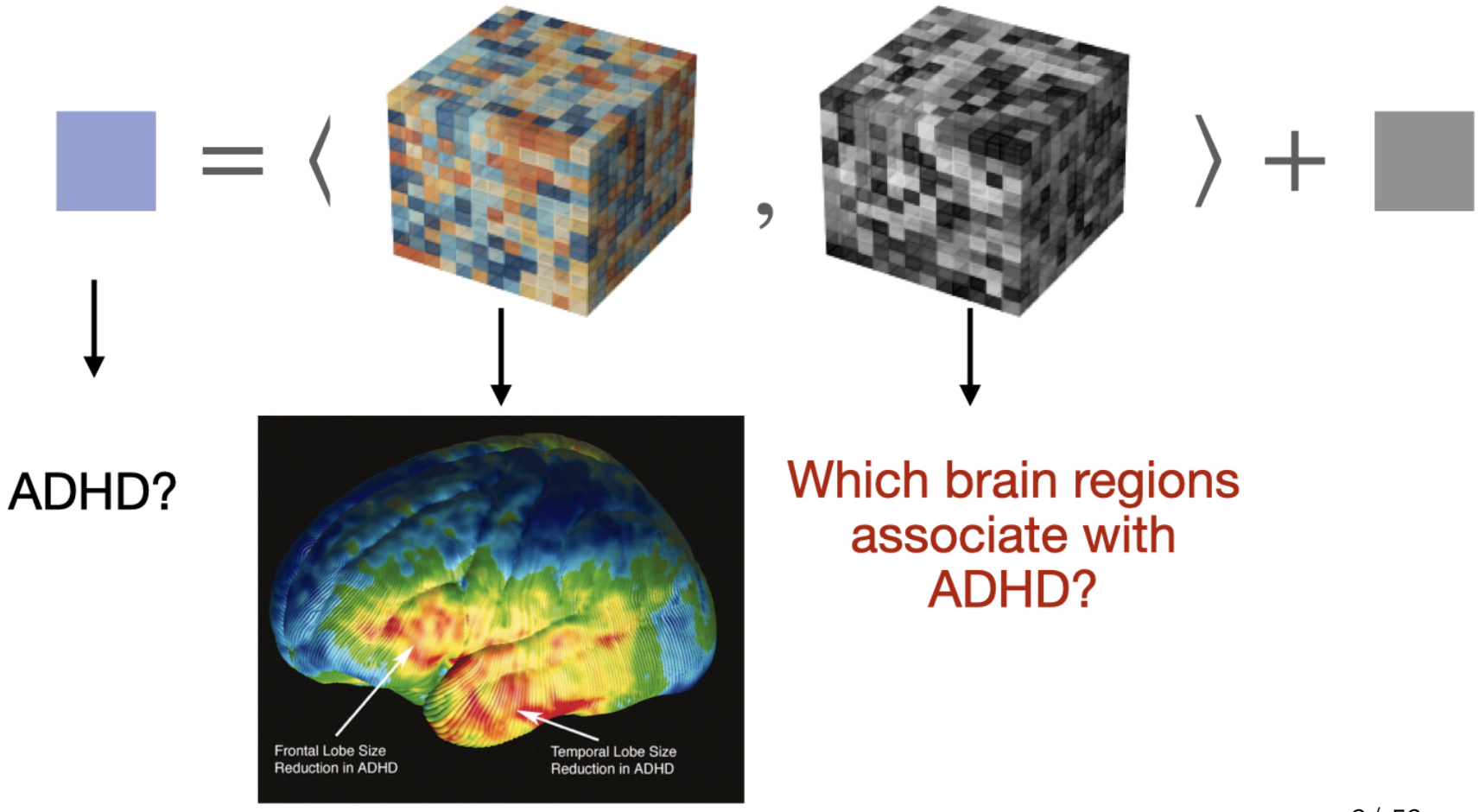
Tensor clustering



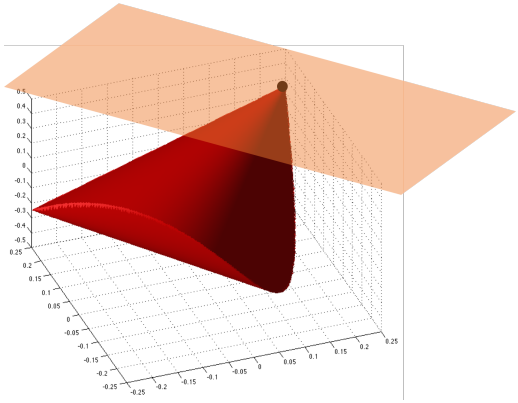
Various tensor problems

— *Zhou et al. '13, Raskutti et al. '19, Zhang et al. '20*

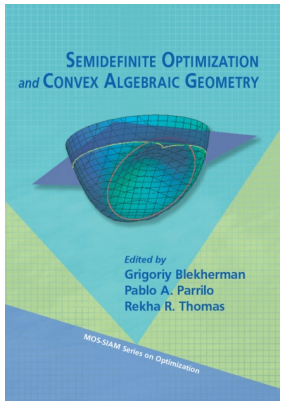
Tensor regression



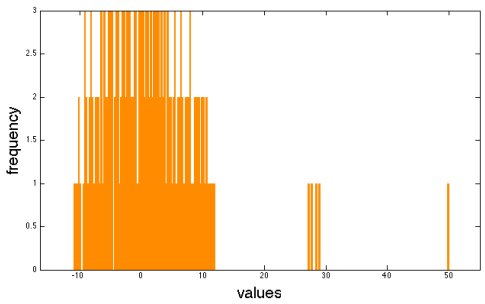
Prior arts



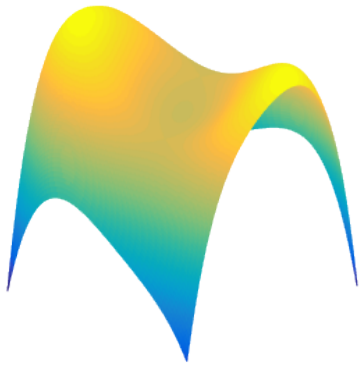
convex relaxation



sum-of-squares



spectral methods



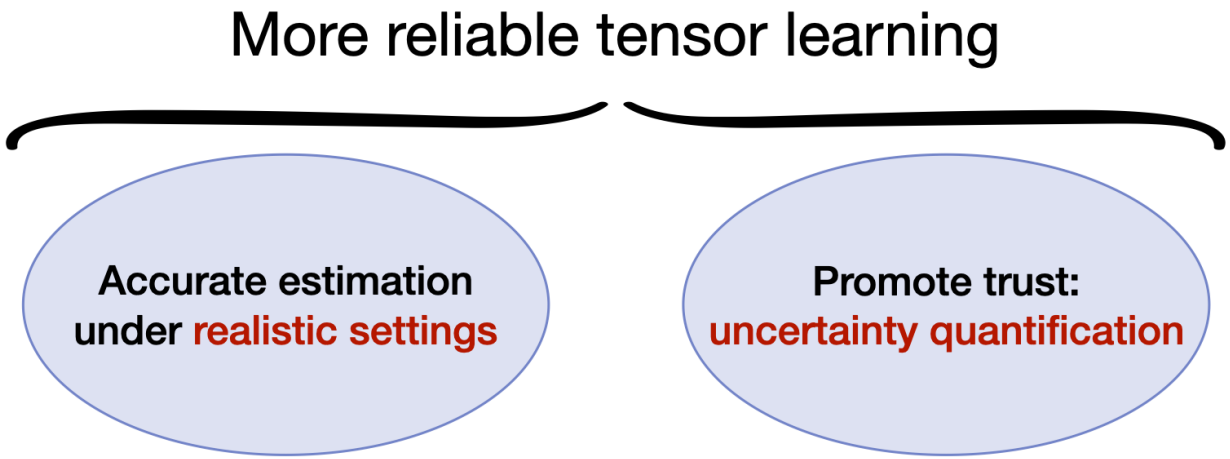
nonconvex optimization

- 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



Can we make tensor learning more reliable and realistic?

In this talk...

- **Heteroskedastic** tensor clustering
  - also discuss heteroskedastic PCA

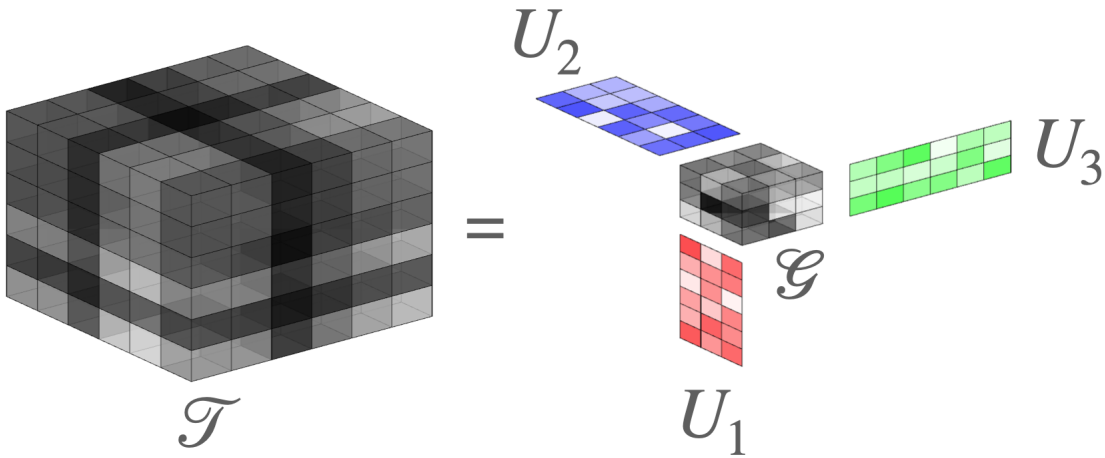
- **Uncertainty quantification** for tensor learning

Tucker Low-rank Tensors

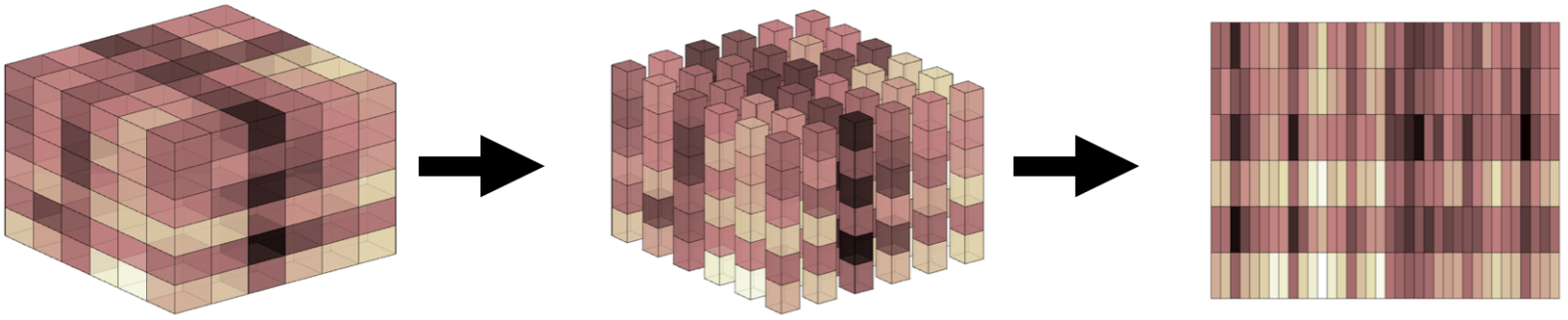
- Low-rank structures are commonly assumed!
- $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ has Tucker rank (r_1, r_2, r_3) if

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

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

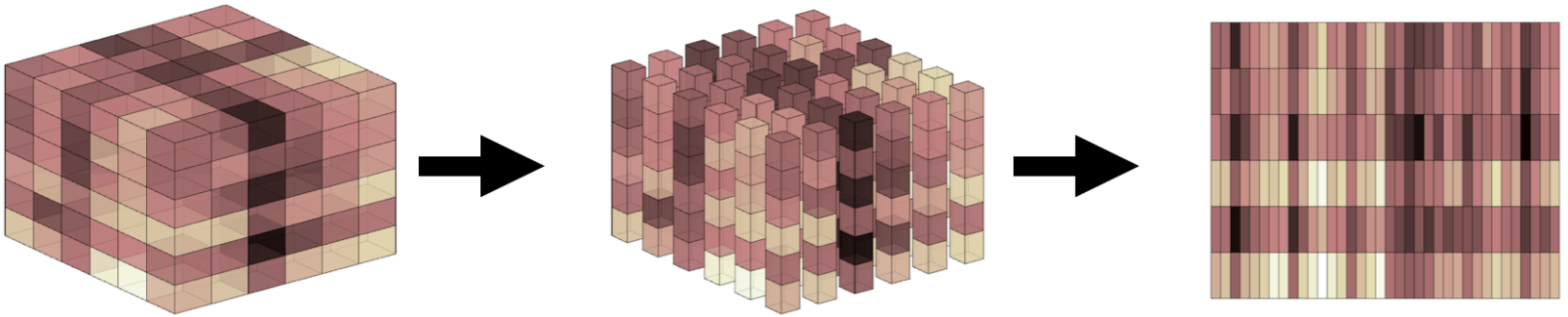


Tucker Low-rank Tensors



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

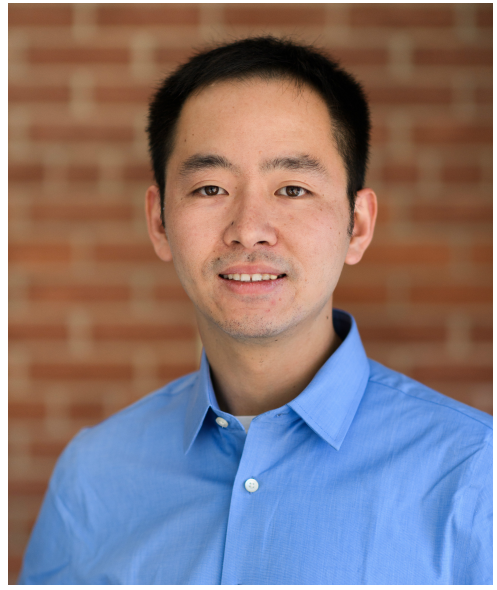
Tucker Low-rank Tensors



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

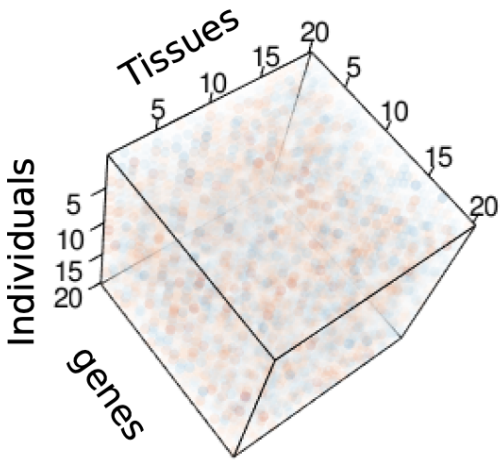
$$\mathsf{rank}(\boldsymbol{\mathcal{T}}) = (r_1, r_2, r_3) \iff \mathsf{rank}(\mathcal{M}_j(\boldsymbol{\mathcal{T}})) = r_j, \quad \forall 1 \leq j \leq 3.$$

Part 1: Heteroskedastic tensor clustering



Yuxin Chen
Wharton Statistics & Data Science

Multiway clustering structure



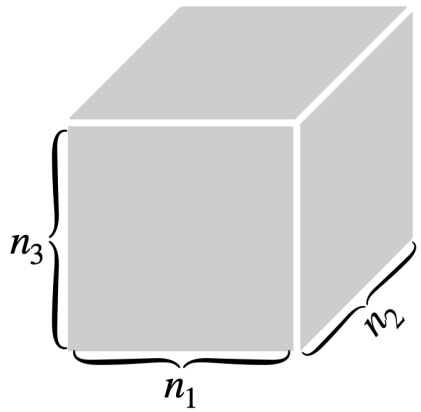
Multi-tissue gene expression

fig. credit: Wang et al. '19

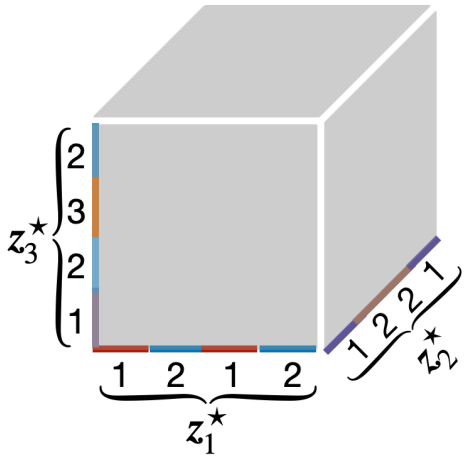
Multilayer network analysis

fig. credit: Domenico et al. '15

Tensor block model

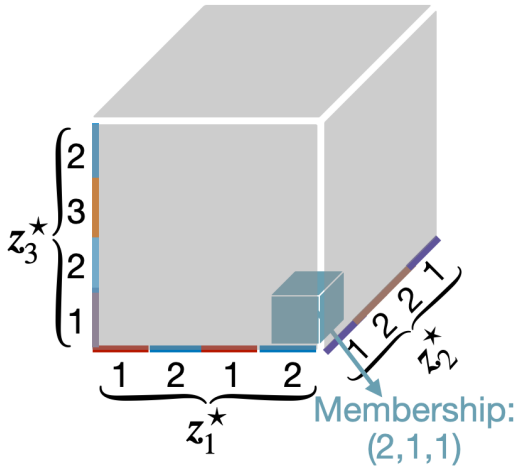


Tensor block model



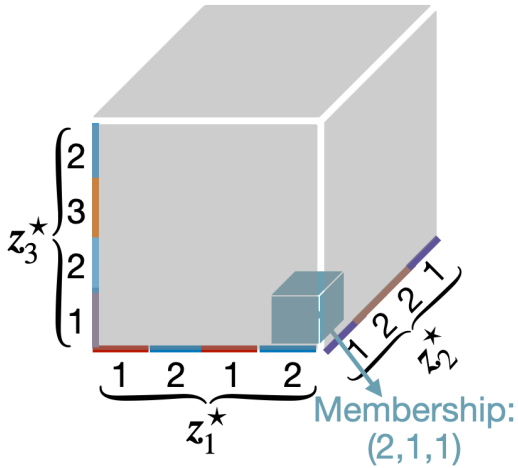
- $\boldsymbol{z}_i^\star \in [k_i]^{n_i}$: cluster assignment vector along the i-th mode
  - $z_{i,j}^\star = \ell$ if the jth index falls within cluster ℓ

Tensor block model



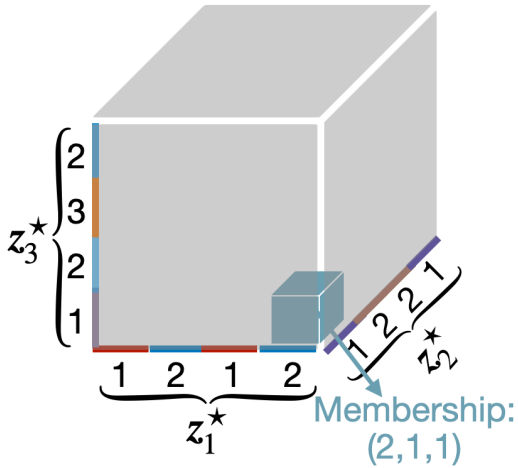
- $\boldsymbol{z}_i^\star \in [k_i]^{n_i}$: cluster assignment vector along the i-th mode
  - $z_{i,j}^\star = \ell$ if the jth index falls within cluster ℓ
  - The membership of (i_1, i_2, i_3) is $(z_{1,i_1}^\star, z_{2,i_2}^\star, z_{3,i_3}^\star)$

Tensor block model



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

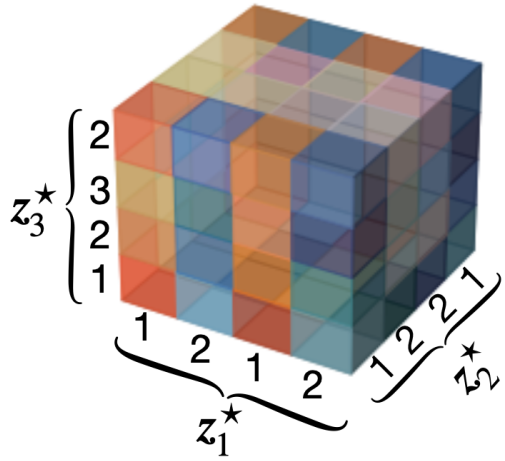
Tensor block model



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

$$Y_{i_1,i_2,i_3} = S^\star_{z^\star_{1,i_1}, z^\star_{2,i_2}, z^\star_{3,i_3}} + \underbrace{E_{i_1,i_2,i_3}}_{\text{zero-mean ind. noise}} .$$

Tensor block model

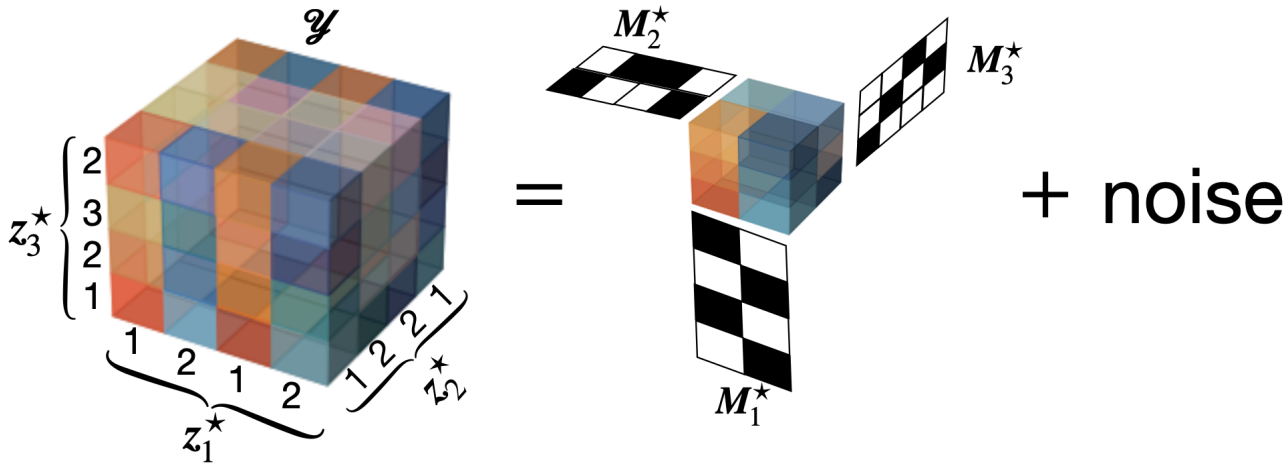


- $\boldsymbol{z}_i^\star \in [k_i]^{n_i}$: cluster assignment vector along the i-th mode
- $\boldsymbol{\mathcal{S}}^\star \in \mathbb{R}^{k_1 \times k_2 \times k_3}$: block/clustering mean
- **Noisy observations:** for all $(i_1, i_2, i_3) \in [n_1] \times [n_2] \times [n_3]$,

$$Y_{i_1,i_2,i_3} = S^\star_{z^\star_{1,i_1}, z^\star_{2,i_2}, z^\star_{3,i_3}} + \underbrace{E_{i_1,i_2,i_3}}_{\text{zero-mean ind. noise}} .$$

- **Goal**: recover $\boldsymbol{z}_i^\star$, $i \in [3]$ from $\boldsymbol{\mathcal{Y}}$

Tensor block model



Equivalently, we observe

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

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

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

- **Goal**: recover $M_i^\star$, $i \in [3]$ from $\mathcal{Y}$

Two special cases

- **Sub-Gaussian tensor block models**
  - entries of $\mathcal{E}$ are indep. zero-mean sub-Gaussian
  - generalization of (sub-)Gaussian mixture model

Two special cases

- **Sub-Gaussian tensor block models**
  - entries of $\mathcal{E}$ are indep. zero-mean sub-Gaussian
  - generalization of (sub-)Gaussian mixture model

- **Stochastic tensor block models**
  - generalization of bipartite stochastic block model
  - each entry of $\mathcal{S}^\star$ is connection probability
  - $Y_{i,j,\ell} = \begin{cases} 1, & \text{with prob. } S^\star_{z^\star_{1,i}, z^\star_{2,j}, z^\star_{3,\ell}}, \\ 0, & \text{o.w.} \end{cases}$ represents if there is a hyper-edge connecting (i, j, ℓ)

Two methods in the literature

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

$$\left(\widehat{\boldsymbol{\mathcal{S}}}, \widehat{\boldsymbol{z}}_1, \widehat{\boldsymbol{z}}_2, \widehat{\boldsymbol{z}}_3\right) := \underset{\substack{\boldsymbol{\mathcal{S}} \in [k_1]\times[k_2]\times[k_3] \\ \boldsymbol{z}_i \in [k_i]^{n_i}}}{\arg\min} \sum_{i,j,\ell} \left(Y_{i,j,\ell} - S_{z_{1,i}, z_{2,j}, z_{3,\ell}}\right)^2 \qquad (1)$$

- statistically accurate, computationally intractable!

Two methods in the literature

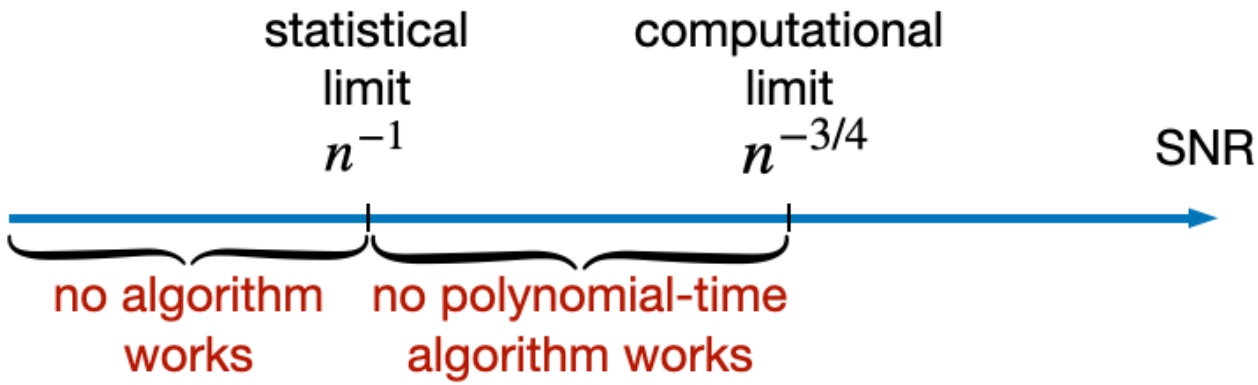
High-order Spectral Clustering + High-order Lloyd (Han et al. '22)

- estimate top singular subspaces along each mode (based on vanilla SVD)

- project tensor data onto estimated subspaces & approximate k-means procedure to cluster nodes

- High-order Lloyd (HLloyd): iterative further refinements

Statistical-computational gap



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

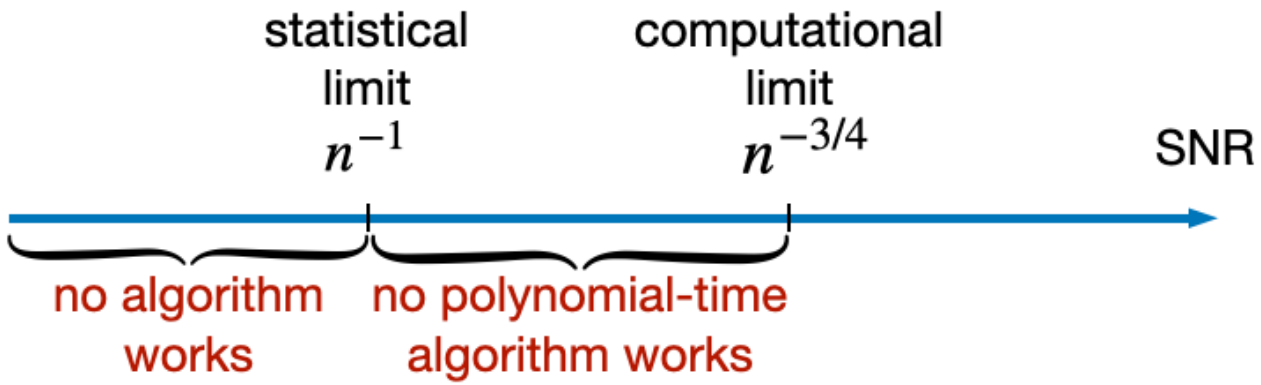
- signal strength:

$$\Delta_1 := \overbrace{\min_{1 \le j_1 \ne j_2 \le k_1} \left\| S^\star_{j_1,:,:} - S^\star_{j_2,:,:} \right\|_2}^{\text{minimum seperation distance along mode}-1}$$

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

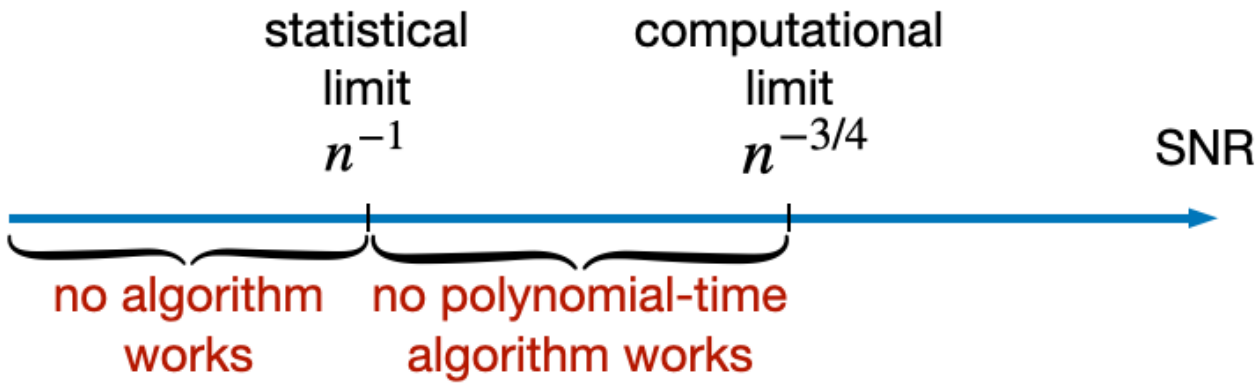
- noise level: $\omega_{\max}$

Statistical-computational gap



- Under i.i.d. sub-Gaussian noise, HSC + HLloyd achieves exact clustering if SNR exceeds $n^{-3/4}$ (up to log factors)

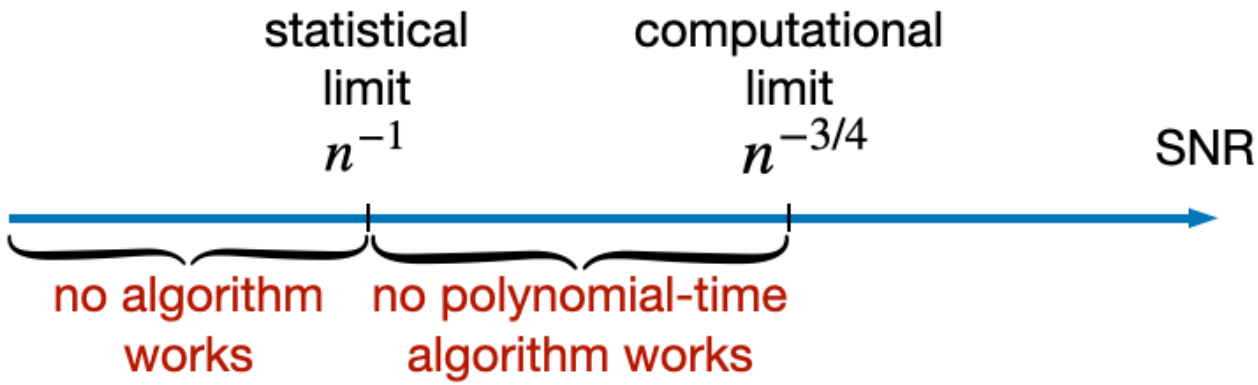
Statistical-computational gap



- Under i.i.d. sub-Gaussian noise, HSC + HLloyd achieves exact clustering if SNR exceeds $n^{-3/4}$ (up to log factors)

- A common senario in practice: **heteroskedastic data**
  - noise variances $\underbrace{\left\{\mathbb{E}[E_{i,j,\ell}^2]\right\}}_{\text{unknown } \textit{a priori}}$ are location-varying
  - discrete-valued observations: multi-tissue gene expression data, multilayer network data

Statistical-computational gap



- Under i.i.d. sub-Gaussian noise, HSC + HLloyd achieves exact clustering if SNR exceeds $n^{-3/4}$ (up to log factors)

- A common senario in practice: **heteroskedastic data**
  - noise variances $\underbrace{\left\{\mathbb{E}[E_{i,j,\ell}^2]\right\}}_{\text{unknown } \textit{a priori}}$ are location-varying
  - discrete-valued observations: multi-tissue gene expression data, multilayer network data

- fail dramatically in the face of **heteroskedastic** noise!

Can we achieve exact clustering with heteroskedastic noise under essential conditions?

Spectral clustering

- Step 1: estimate "important" subspaces of $\boldsymbol{X}_i^\star = \mathcal{M}_i(\boldsymbol{\mathcal{X}}^\star)$

- Step 2: apply approximate k-means to obtain cluster nodes

Spectral clustering

- Step 1: estimate "important" subspaces of $\boldsymbol{X}_i^\star = \mathcal{M}_i(\boldsymbol{\mathcal{X}}^\star)$

  - key challenge: how to handle heteroskedastic noise?

- Step 2: apply approximate k-means to obtain cluster nodes

Road map

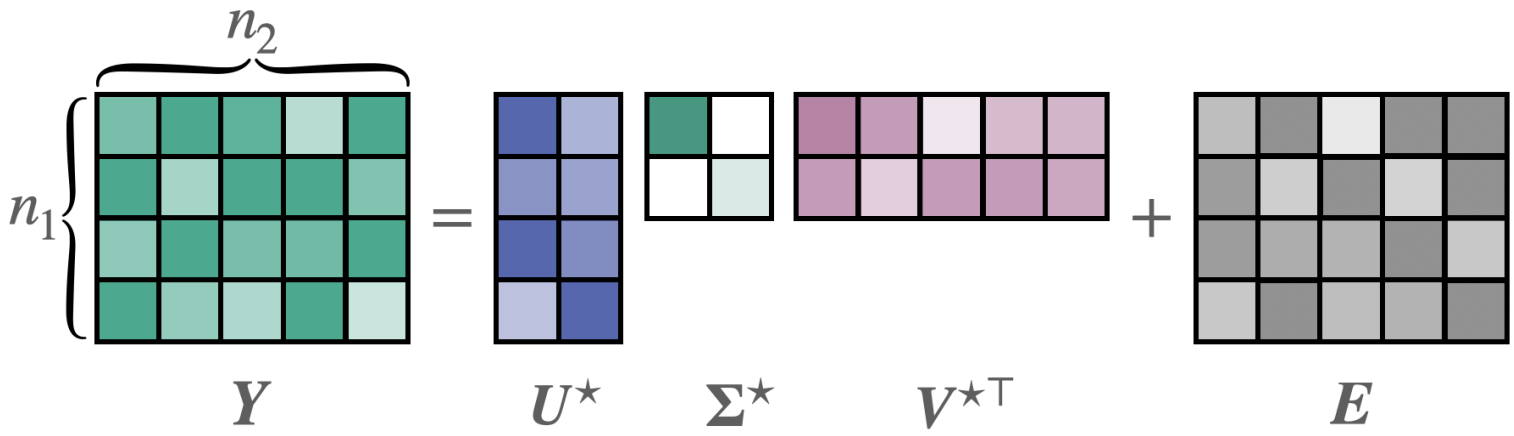
Subspace estimation problem

- Key challenges: heteroskedastic noise, …
- 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

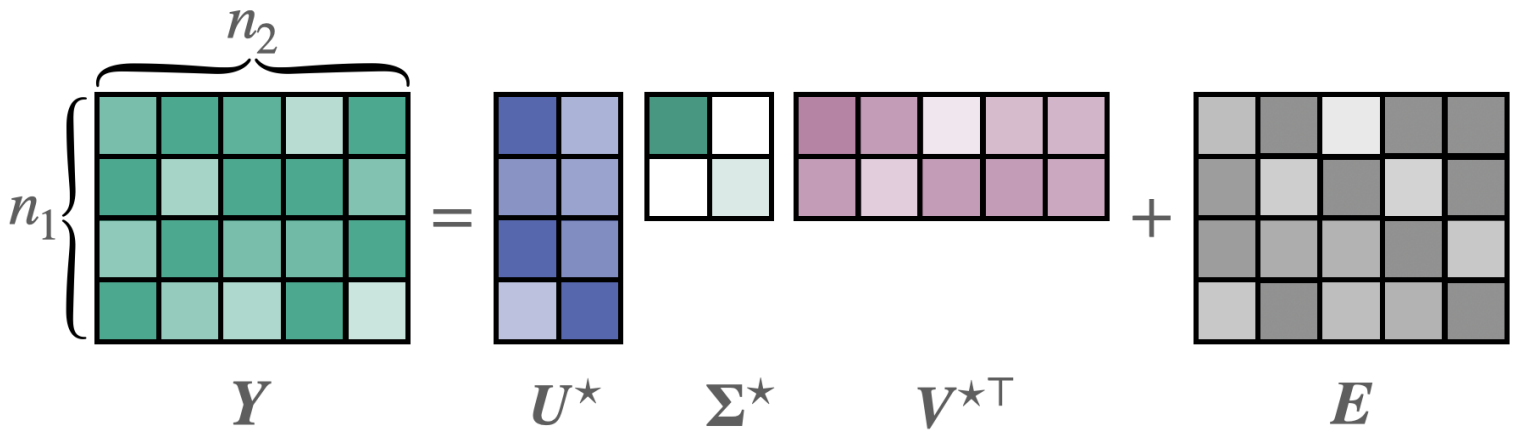


- **Ground truth:** rank-r matrix $\boldsymbol{X}^\star$ with SVD ($r \ll \min\{n_1, n_2\}$)

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

where $\boldsymbol{U}^\star \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{\Sigma}^\star = \mathsf{diag}\{\sigma_1^\star, \cdots, \sigma_r^\star\}$, $\boldsymbol{V}^\star \in \mathbb{R}^{n_2 \times r}$

A detour: a subspace estimation / model



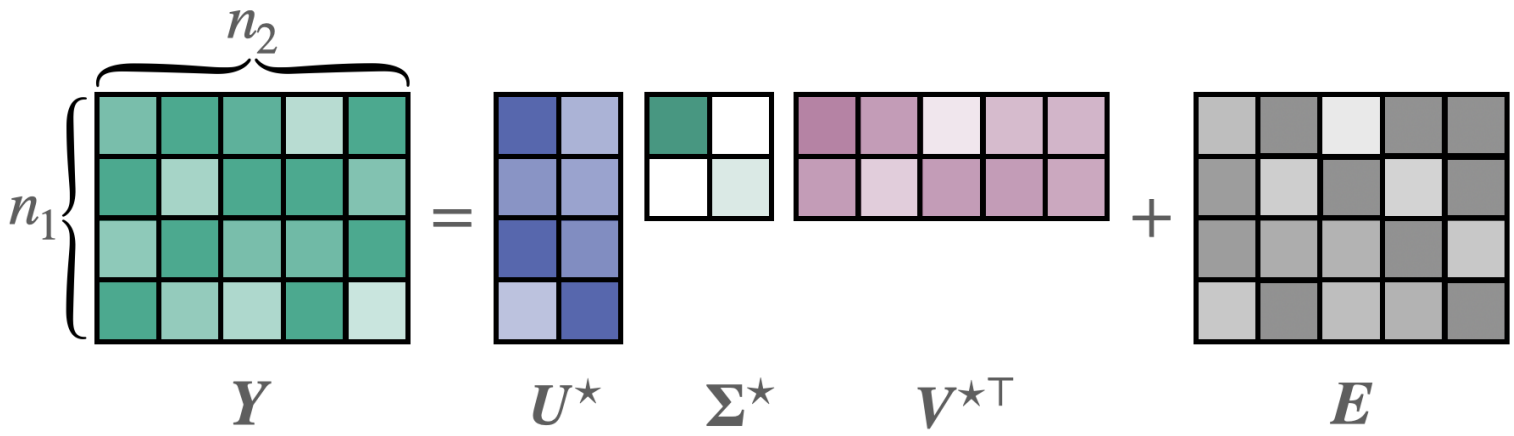
- **Ground truth:** rank-r matrix $\boldsymbol{X}^\star$ with SVD ($r \ll \min\{n_1, n_2\}$)

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

- **Noisy observations:** $\boldsymbol{Y} = \boldsymbol{X}^\star + \underbrace{\boldsymbol{E}}$

zero-mean ind. noise

A detour: a subspace estimation / model



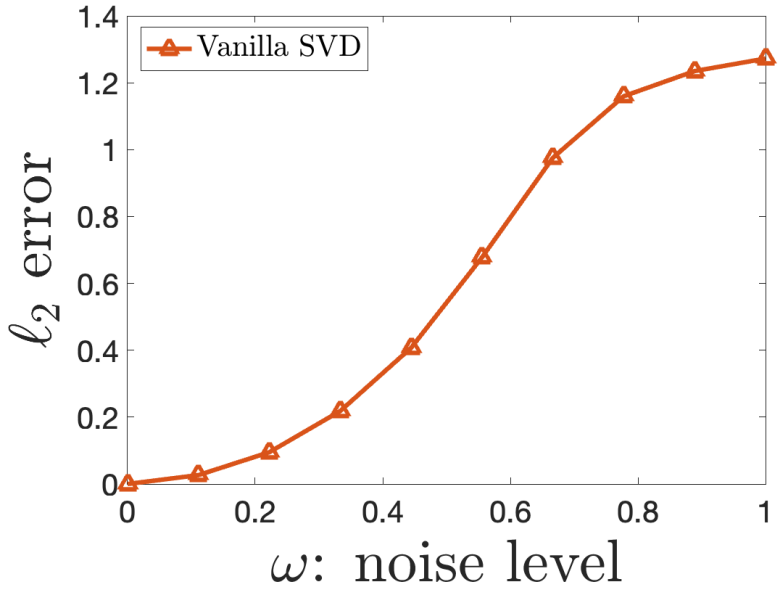
- **Ground truth:** rank-r matrix $\boldsymbol{X}^\star$ with SVD ($r \ll \min\{n_1, n_2\}$)

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

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

Review of popular methods

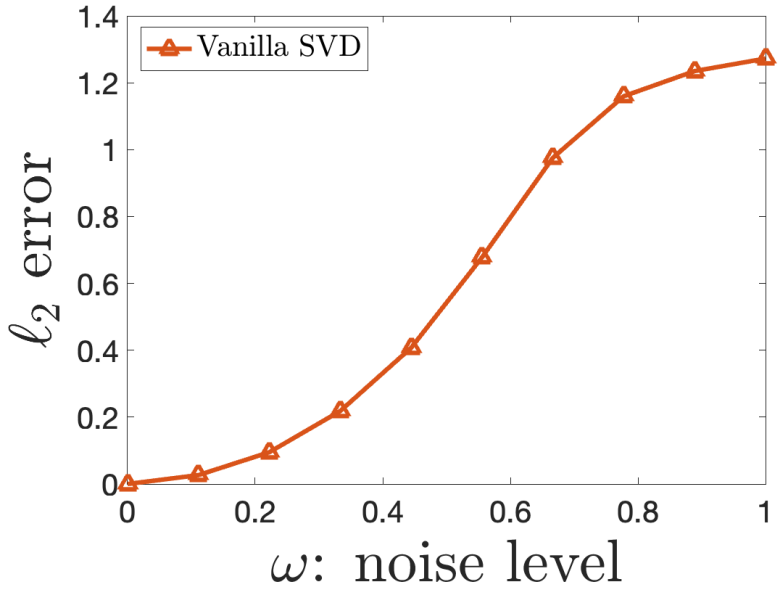
$$n_1 = 100, n_2 = 10,000$$
$$r = 2, \kappa := \sigma_1^\star / \sigma_r^\star = 2$$
$$\omega_i \overset{\text{ind.}}{\sim} \mathsf{Unif}([0, \omega]),$$
$$E_{i,j} \overset{\text{ind.}}{\sim} \mathcal{N}(0, \omega_i^2)$$



vanilla SVD: $U \longleftarrow$ rank-r left singular subspace of $\boldsymbol{Y} = \boldsymbol{X}^\star + \boldsymbol{E}$

Review of popular methods

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



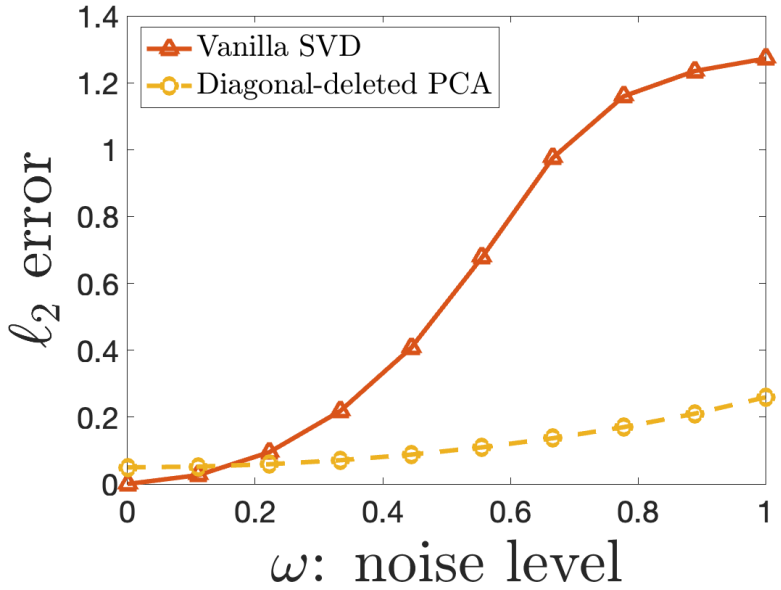
vanilla SVD: $U \longleftarrow$ rank-r left singular subspace of $\boldsymbol{Y} = \boldsymbol{X}^\star + \boldsymbol{E}$

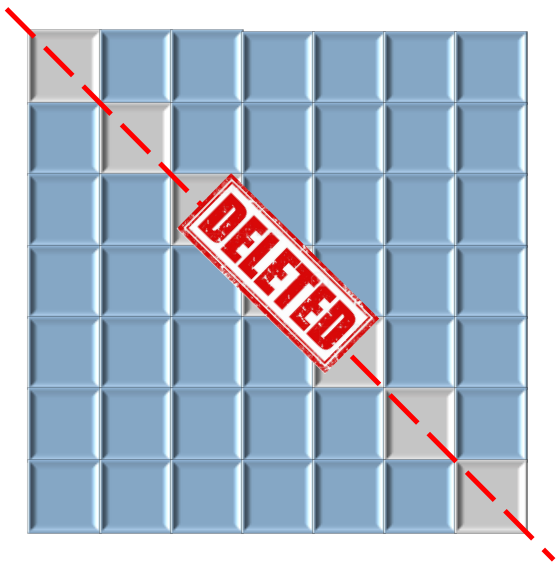
- often sub-optimal due to large bias in diagonal entries:

$$\mathbb{E}\big[\boldsymbol{Y}\boldsymbol{Y}^\top\big] = \underbrace{\boldsymbol{X}^\star \boldsymbol{X}^{\star\top}}_{\checkmark} + \underbrace{\mathsf{diag}\Big\{\Big[\sum\nolimits_j \mathbb{E}[E_{i,j}^{\star 2}]\Big]_{1\le i\le n_1}\Big\}}_{\text{potentially large diagonal matrix!}}$$

Review of popular methods

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



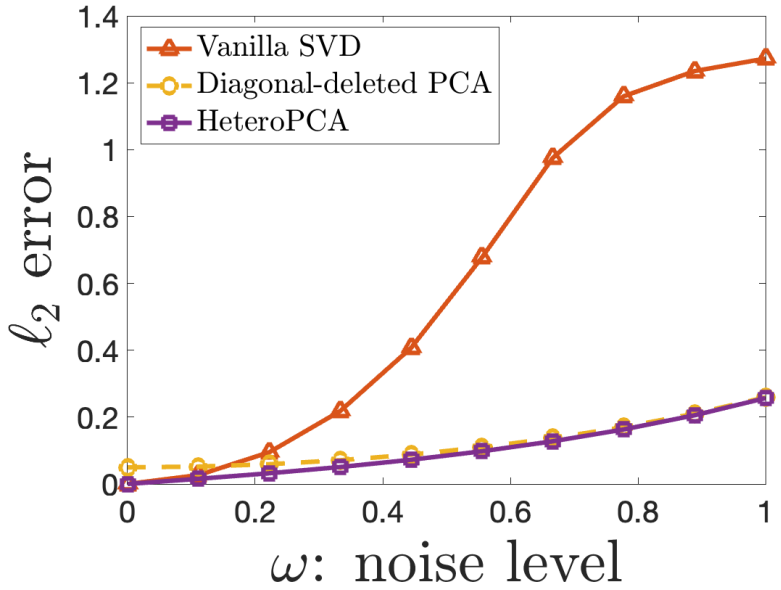


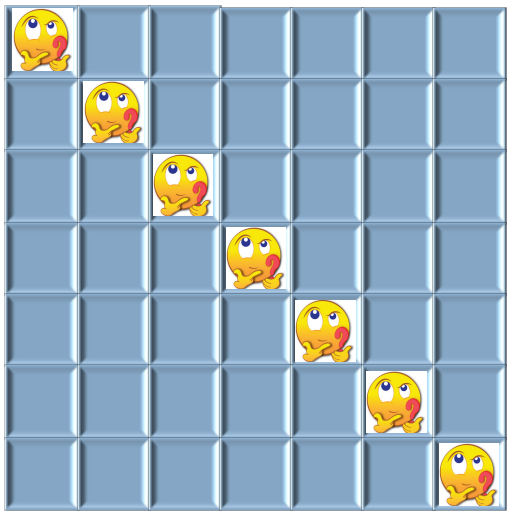
diagonal-deleted PCA:

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

Review of popular methods

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



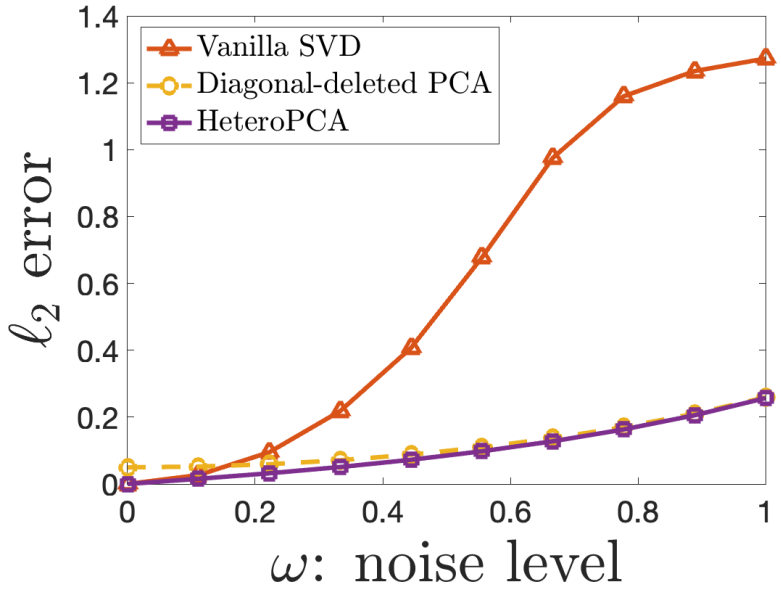


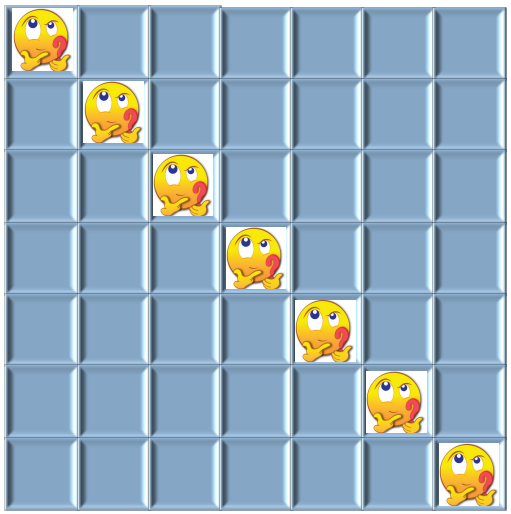
HeteroPCA (Zhang, Cai, Wu '22)

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

Review of popular methods

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





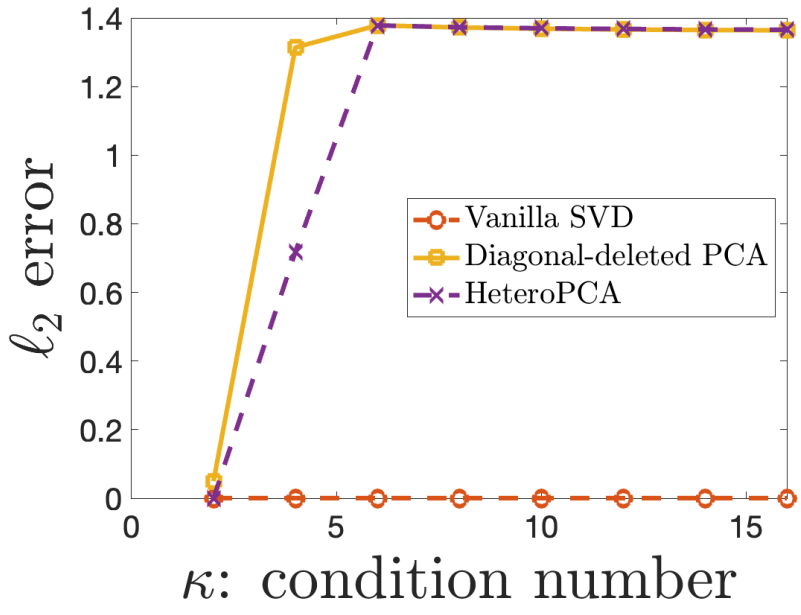
HeteroPCA (Zhang, Cai, Wu '22)

- **initialize:** $\boldsymbol{G}^0 = \mathcal{P}_{\text{off-diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top)$
- for $t = 0, 1, \ldots$

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

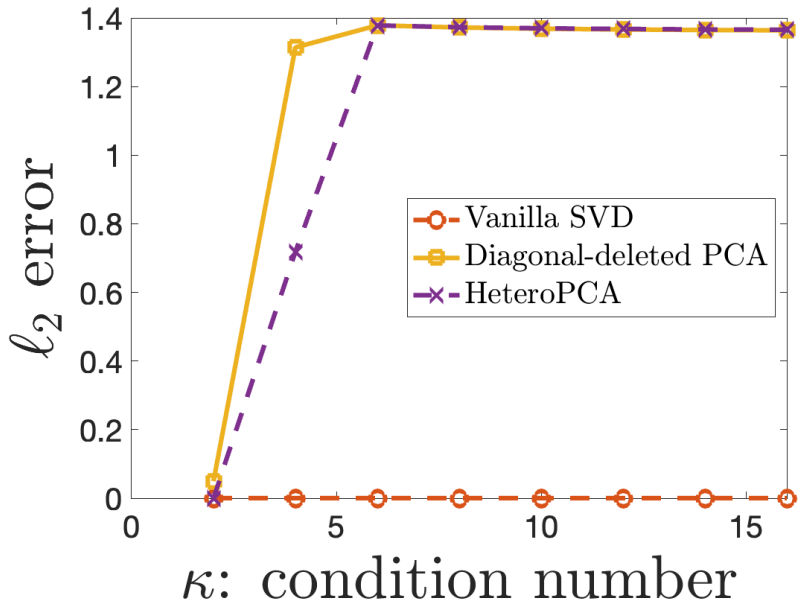
A curious phenomenon: curse of ill-conditioning

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

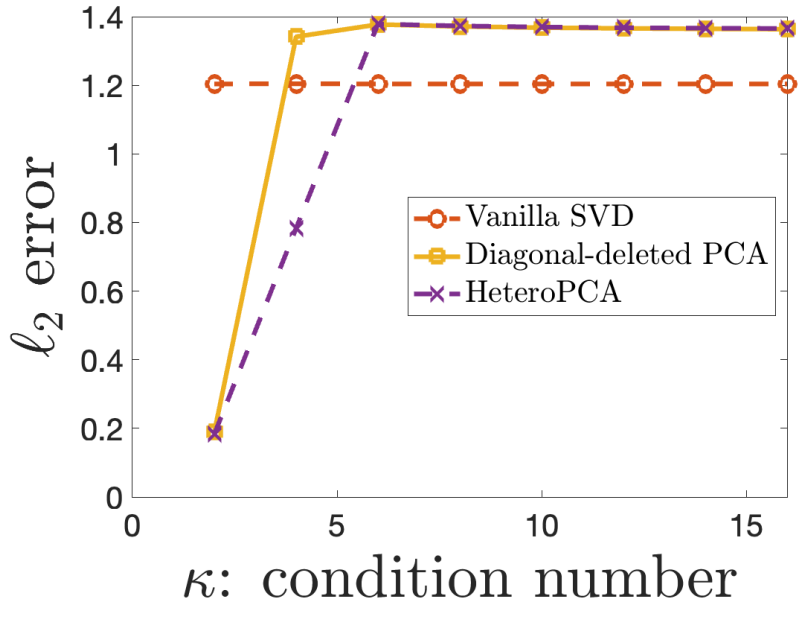


noiseless case

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

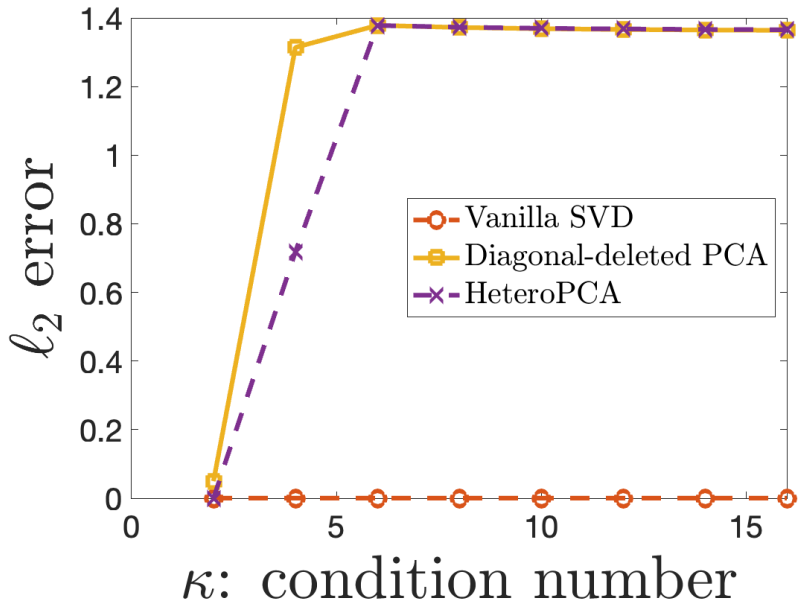


noiseless case

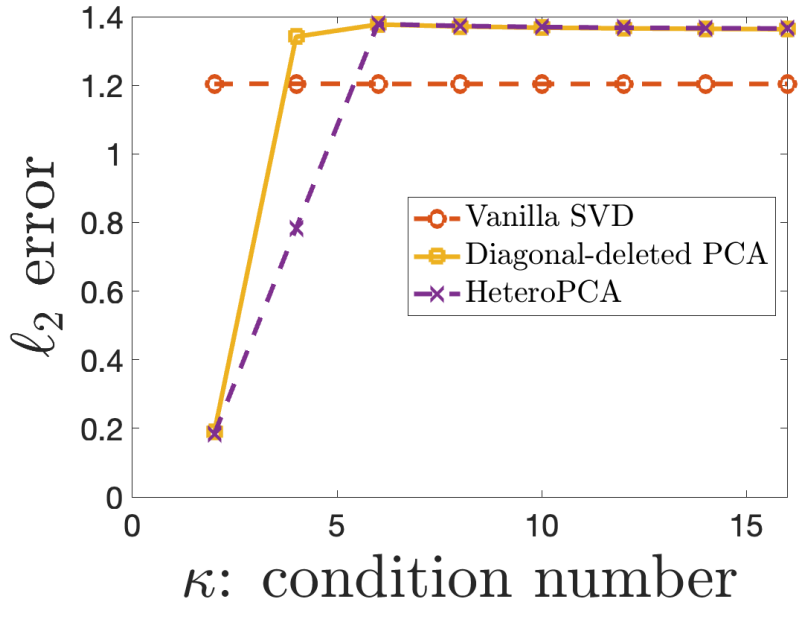


noisy case

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



noiseless case

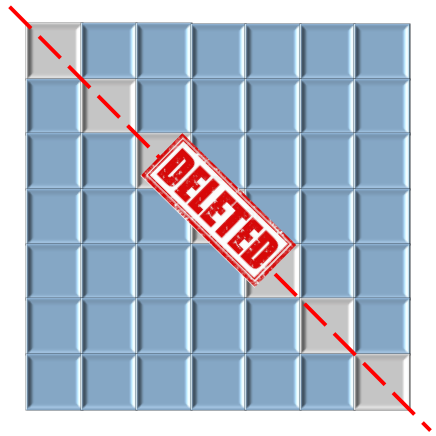


noisy case

Previous methods degrade as *condition number of $\boldsymbol{X}^\star$ increases*!

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

Diagonsis: influences of diagonal deletion

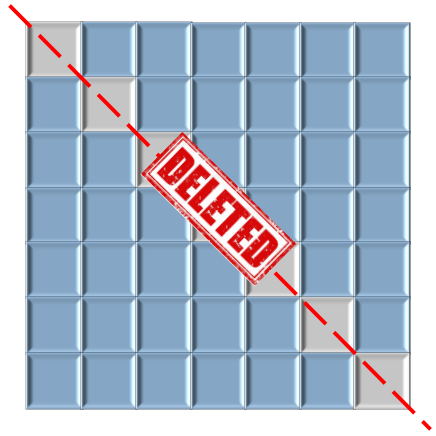


$$\mathbb{E}\Big[\mathcal{P}_{\text{off-diag}}\big(\boldsymbol{Y}\boldsymbol{Y}^\top\big)\Big] = \boldsymbol{X}^\star\boldsymbol{X}^{\star\top} - \underbrace{\mathcal{P}_{\text{diag}}\big(\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\big)}$$

ideally negligible compared to $\sigma_r^{\star 2}$

- ideally, we hope digonal deletion has negligible influences

Diagonsis: influences of diagonal deletion

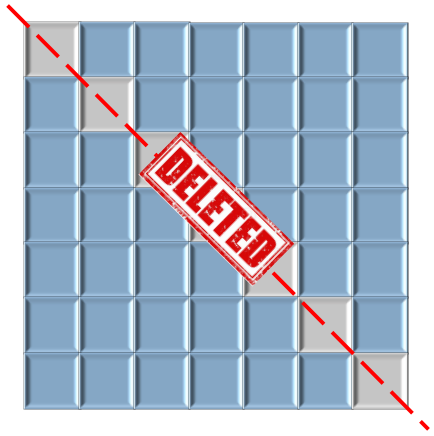


$$\mathbb{E}\Big[\mathcal{P}_{\text{off-diag}}\big(\boldsymbol{Y}\boldsymbol{Y}^\top\big)\Big] = \boldsymbol{X}^\star\boldsymbol{X}^{\star\top} - \underbrace{\mathcal{P}_{\text{diag}}\big(\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\big)}$$

but larger than $\sigma_r^{\star 2}$ if $\sigma_1^\star/\sigma_r^\star$ is too large

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

Diagonsis: influences of diagonal deletion



$$\mathbb{E}\Big[\mathcal{P}_{\text{off-diag}}\big(\boldsymbol{Y}\boldsymbol{Y}^\top\big)\Big] = \boldsymbol{X}^\star\boldsymbol{X}^{\star\top} - \underbrace{\mathcal{P}_{\text{diag}}\big(\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\big)}$$

but larger than $\sigma_r^{\star 2}$ if $\sigma_1^\star/\sigma_r^\star$ is too large

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

Both diagonal-deleted PCA & HeteroPCA become ineffective

initialized by diagonal-deleted PCA

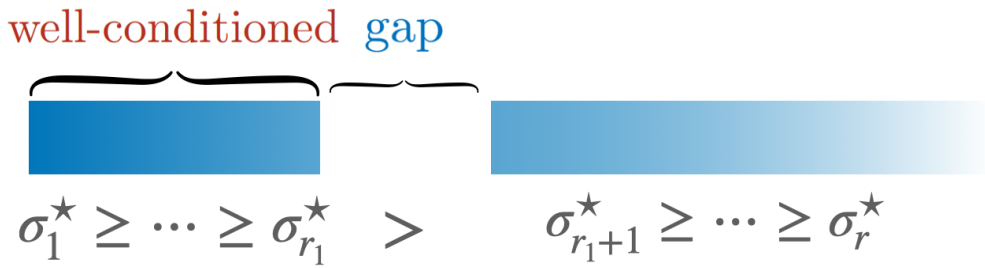
in the presence of ill-conditioning!

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

— *Zhang et al. '18, Yan et al. '21*

revisit HeteroPCA theory: works well if

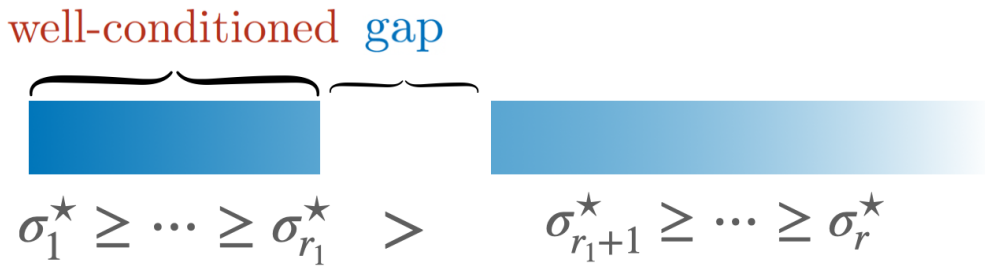
- $\boldsymbol{X}^\star$ is well-conditioned
- least singular value $\sigma_r^\star$ (or spectral gap) is not buried by noise



— *Zhang et al. '18, Yan et al. '21*

revisit HeteroPCA theory: works well if

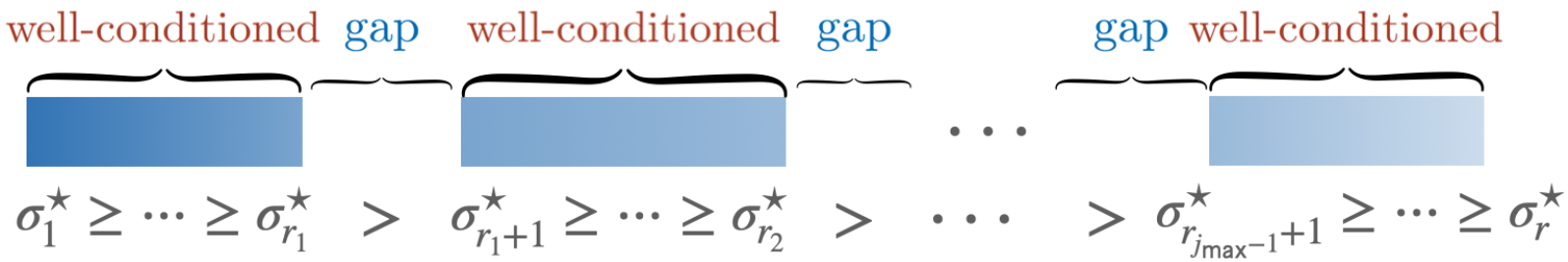
- $\boldsymbol{X}^\star$ is well-conditioned
- least singular value $\sigma_r^\star$ (or spectral gap) is not buried by noise



solution:

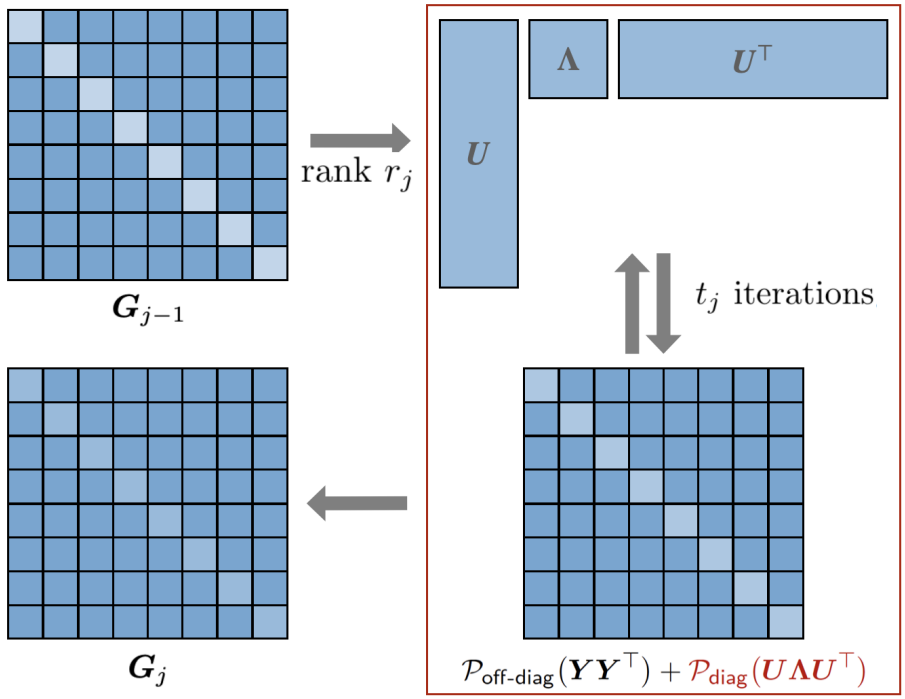
- divide eigenvalues into well-conditioned & well-separated subblocks
- estimate subblocks sequentially

Proposed algorithm: deflated-HeteroPCA



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

Proposed algorithm: deflated-HeteroPCA



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

- invoke HeteroPCA($\underbrace{\boldsymbol{G}_{j-1}}_{\text{input}}, \underbrace{r_j}_{\text{rank}}$) to impute diagonals & obtain $\boldsymbol{G}_k$

Proposed algorithm: deflated-HeteroPCA

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

$$
\begin{aligned}
&j = j + 1 \\
&\text{select } r_j \text{ in a data-driven manner} \\
&(\boldsymbol{G}_j, \boldsymbol{U}_j) = \text{HeteroPCA}(\underbrace{\boldsymbol{G}_{j-1}}_{\text{input}}, \underbrace{r_j}_{\text{rank}})
\end{aligned}
$$

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

Proposed algorithm: deflated-HeteroPCA

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

$$j = j + 1$$

select r_j in a data-driven manner

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

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

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

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

Assumptions (ignoring log factors)

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

Assumptions (ignoring log factors)

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

$$\frac{\sigma_r^\star}{\omega} \gtrsim (n_1 n_2)^{1/4} + n_1^{1/2}$$

  - necessary for existence of consistent estimators (Cai et al. '21)

Assumptions (ignoring log factors)

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

$$\frac{\sigma_r^\star}{\omega} \gtrsim (n_1 n_2)^{1/4} + n_1^{1/2}$$

  - necessary for existence of consistent estimators (Cai et al. '21)
- rank $r = O(1)$
- incoherence $\mu := \max\{ \frac{n_1}{r} \left\| \boldsymbol{U}^\star \right\|_{2,\infty}^2, \frac{n_2}{r} \left\| \boldsymbol{V}^\star \right\|_{2,\infty}^2 \} = O(1)$

Theoretical guarantees

Theorem 1 (Zhou, Chen '23)

With high prob., Deflated-HeteroPCA yields

$$\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^{\star}\| \lesssim \zeta_{\mathsf{op}}$$

for some rotation matrix $\boldsymbol{R}_{\boldsymbol{U}}$*, where* $\zeta_{\mathsf{op}} = \frac{\sqrt{n_1 n_2}\,\omega^2}{\sigma_r^{\star 2}} + \frac{\sqrt{n_1}\,\omega}{\sigma_r^{\star}}$

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

Theoretical guarantees

Theorem 1 (Zhou, Chen '23)

With high prob., Deflated-HeteroPCA yields

$$\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^{\star}\| \lesssim \zeta_{\mathsf{op}}$$

$$\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^{\star}\|_{2,\infty} \lesssim \frac{1}{\sqrt{n_1}}\zeta_{\mathsf{op}} \quad \textit{(fine-grained } \ell_{2,\infty}\textit{)}$$

for some rotation matrix $\boldsymbol{R}_{\boldsymbol{U}}$*, where* $\zeta_{\mathsf{op}} = \frac{\sqrt{n_1 n_2}\,\omega^2}{\sigma_r^{\star 2}} + \frac{\sqrt{n_1}\,\omega}{\sigma_r^{\star}}$

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

Come back to tensor clustering...

Subspace estimation problem

- Key challenges: heteroskedastic noise, 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!

Come back to tensor clustering...

Subspace estimation problem

- Key challenges: heteroskedastic noise, not well-conditioning
- Propose Deflated HeteroPCA!

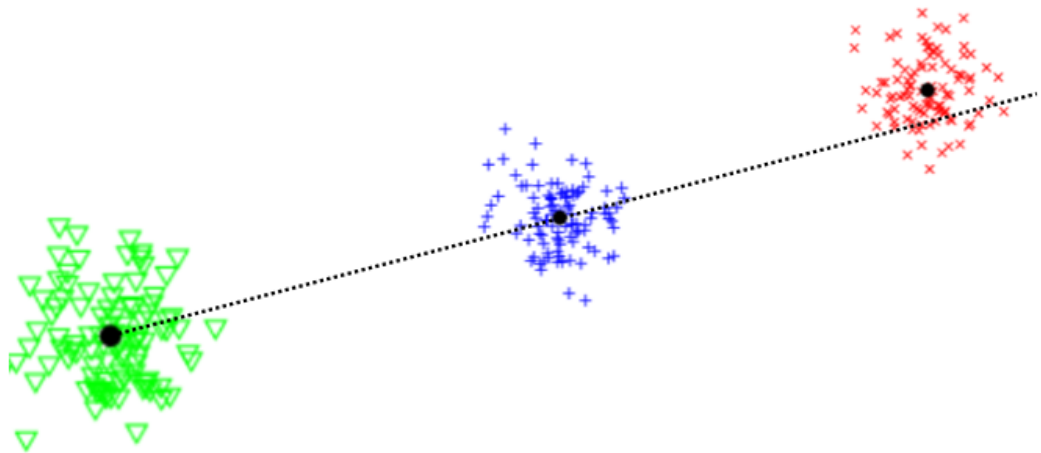
More reliable solution to PCA problem!

Spectral tensor clustering

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

Come back to tensor clustering...

- Least singular value assumption is unnecessary for clustering!



- Add a thresholding procedure!

Thresholded Deflated-HeteroPCA$(\boldsymbol{Y}, r, \tau)$

- **Initialize:** $\boldsymbol{G}^0 = \mathcal{P}_{\text{off-diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top), j = 0, r_0 = 0$
- **Sequential updates:** while $r_j < r$ and $\sigma_{r_j+1}(\boldsymbol{G}_j) > \underbrace{\tau}_{\text{threshold}}$

  $j = j + 1$

  select r_j in a data-driven manner

  $(\boldsymbol{G}_j, \boldsymbol{U}_j) = \text{HeteroPCA}(\underbrace{\boldsymbol{G}_{j-1}}_{\text{input}}, \underbrace{r_j}_{\text{rank}})$
- **Output:** $U := \boldsymbol{U}_j \quad \longrightarrow \quad$ estimate of $\boldsymbol{U}^\star$

Thresholded Deflated-HeteroPCA$(\boldsymbol{Y}, r, \tau)$

- **Initialize:** $\boldsymbol{G}^0 = \mathcal{P}_{\text{off-diag}}(\boldsymbol{Y}\boldsymbol{Y}^\top), j = 0, r_0 = 0$
- **Sequential updates:** while $r_j < r$ and $\sigma_{r_j+1}(\boldsymbol{G}_j) > \underbrace{\tau}_{\text{threshold}}$

$$j = j + 1$$

select r_j in a data-driven manner

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

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

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

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

Proposed algorithm: High-order HeteroClustering

- Step 1: estimate "important" subspaces of $\boldsymbol{X}_i^\star = \mathcal{M}_i(\boldsymbol{\mathcal{X}}^\star)$

  - Apply Thresholded Deflated-HeteroPCA$(\boldsymbol{X}_i^\star, k_i, \tau)$ to obtain subspace estimate $\boldsymbol{U}_i \in \mathbb{R}^{n_i \times r_i}$

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

$$\widehat{\boldsymbol{B}}_i = \mathcal{M}_i(\boldsymbol{\mathcal{Y}} \underbrace{\times_i \boldsymbol{U}_i \boldsymbol{U}_i^\top}_{\text{denoising}} \underbrace{\times_{i+1} \boldsymbol{U}_{i+1} \times_{i+2} \boldsymbol{U}_{i+2}}_{\text{dimension reduction \& denoising}}) \in \mathbb{R}^{n_i \times r_{i+1} r_{i+2}}$$

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

Assumptions (ignoring log factors)

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

Assumptions (ignoring log factors)

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

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

  - 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
  - $\mathsf{Var}[E_{i,j,\ell}] \le \omega_{\mathsf{max}}^2$
  - $|E_{i,j,\ell}| \lesssim \omega_{\mathsf{max}} n^{3/4}$ with high prob.
  - examples: sub-Gaussian, centered Poisson, stochastic tensor block model
- **signal-to-noise ratio (SNR)**:

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

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

Theoretical guarantees

Theorem 2 (Zhou, Chen '23)

Suppose that $\tau \asymp n^{3/4}\omega_{\max}^2$. *Then with high prob., High-order HeteroClustering (HHC) yields*

$$\mathsf{MCR}(\widehat{\boldsymbol{z}}_i, \boldsymbol{z}_i^\star) = 0, \qquad \forall 1 \le i \le 3.$$

Here, $\mathsf{MCR}(\boldsymbol{z}, \widehat{\boldsymbol{z}}) := \inf_{\text{permutation } \phi:[k]\to[k]} \frac{1}{n}\sum_{j=1}^n \mathbb{1}\{\widehat{z}_j \neq \phi(z_j)\}$.

Theoretical guarantees

Theorem 2 (Zhou, Chen '23)

Suppose that $\tau \asymp n^{3/4}\omega_{\max}^2$. *Then with high prob., High-order HeteroClustering (HHC) yields*

$$\mathrm{MCR}(\widehat{\boldsymbol{z}}_i, \boldsymbol{z}_i^\star) = 0, \qquad \forall 1 \le i \le 3.$$

Here, $\mathrm{MCR}(\boldsymbol{z}, \widehat{\boldsymbol{z}}) := \inf_{\text{permutation } \phi:[k]\to[k]} \frac{1}{n}\sum_{j=1}^n \mathbb{1}\{\widehat{z}_j \neq \phi(z_j)\}$.

- exact clustering under (almost) necessary SNR condition among poly-time algorithms
- handle heteroskedastic noise, no superfluous assumptions on $\boldsymbol{\mathcal{S}}^\star$

Theoretical guarantees

Theorem 2 (Zhou, Chen '23)

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

(1). *noise is not too spiky (e.g.,* $\mathsf{Var}\left[E_{i,j,\ell}\right] \asymp \omega_{\mathsf{max}}^2$*);*

(2). *The observation model is the stochastic tensor block model.*

Then we can choose τ *in a data-driven manner, w.h.p., HHC yields*

$$\mathsf{MCR}(\widehat{\boldsymbol{z}}_i, \boldsymbol{z}_i^\star) = 0, \qquad \forall 1 \le i \le 3.$$

Theoretical guarantees

Theorem 2 (Zhou, Chen '23)

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

(1). *noise is not too spiky (e.g.,* $\mathsf{Var}\left[E_{i,j,\ell}\right] \asymp \omega_{\max}^2$*);*

(2). *The observation model is the stochastic tensor block model.*

Then we can choose τ *in a data-driven manner, w.h.p., HHC yields*

$$\mathsf{MCR}(\widehat{\boldsymbol{z}}_i, \boldsymbol{z}_i^\star) = 0, \qquad \forall 1 \le i \le 3.$$

Theorem 2 + Han et al. '22 Theorem 2: HHC + HLloyd can achieve same theoretical guarantees

Comparisons with prior theory

	heteroskedastic noise	no superfluous spectrum assumptions	SNR for exact recovery
HSC Han et al. '22	✗	✓	$n^{-3/4}$
HSC + HLloyd Han et al. '22	✗	✓	$n^{-3/4}$
Successive Projection Agterberg and Zhang '22	✓	$\kappa \lesssim n^{1/8}$	$\kappa n^{-3/4}$
HHC	✓	✓	$n^{-3/4}$
HHC + HLloyd	✓	✓	$n^{-3/4}$

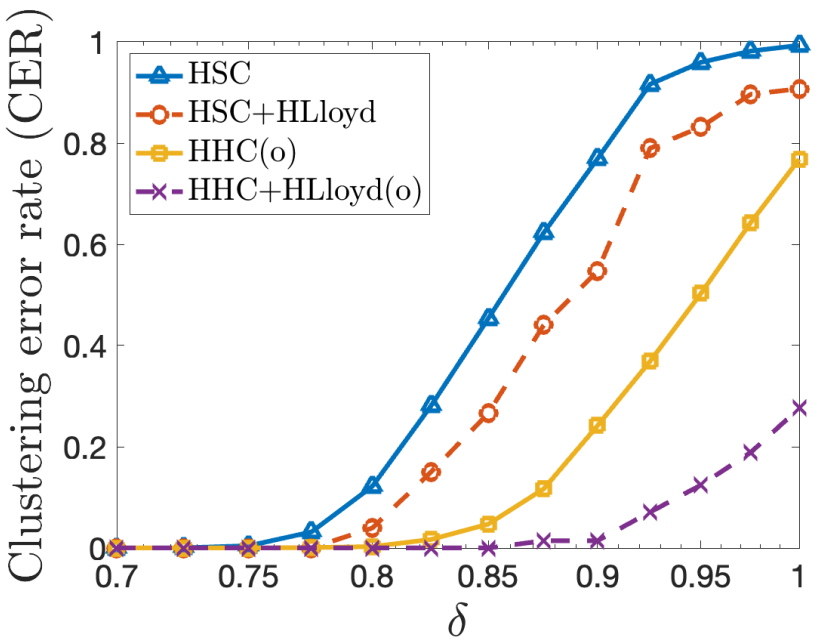
Comparisons with prior theory

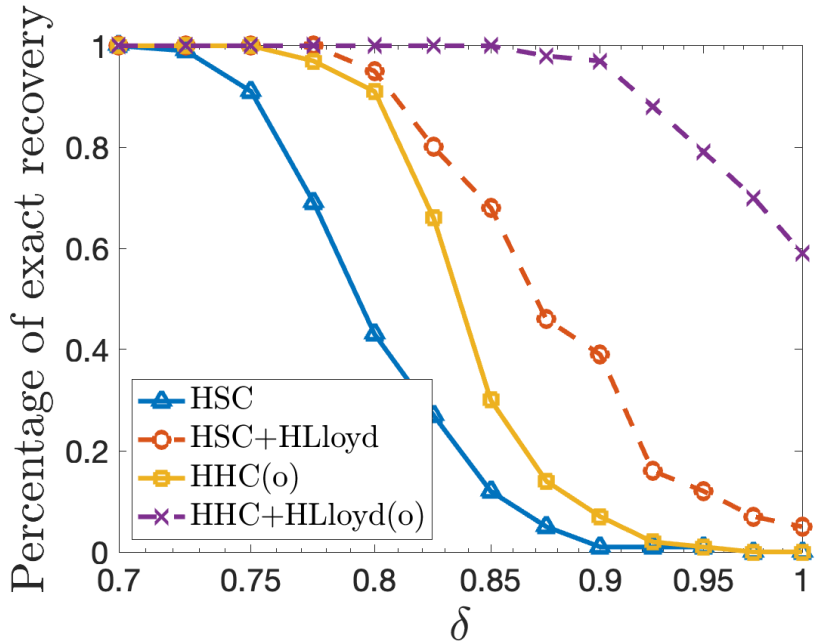
	heteroskedastic noise	no superfluous spectrum assumptions	SNR for exact recovery
HSC Han et al. '22	✗	✓	$n^{-3/4}$
HSC + HLloyd Han et al. '22	✗	✓	$n^{-3/4}$
Successive Projection Agterberg and Zhang '22	✓	$\kappa \lesssim n^{1/8}$	$\kappa n^{-3/4}$
HHC	✓	✓	$n^{-3/4}$
HHC + HLloyd	✓	✓	$n^{-3/4}$

$\kappa :=$ condition number of $\boldsymbol{\mathcal{S}}^\star$

Numerical experiments

Sub-Gaussian tensor block models

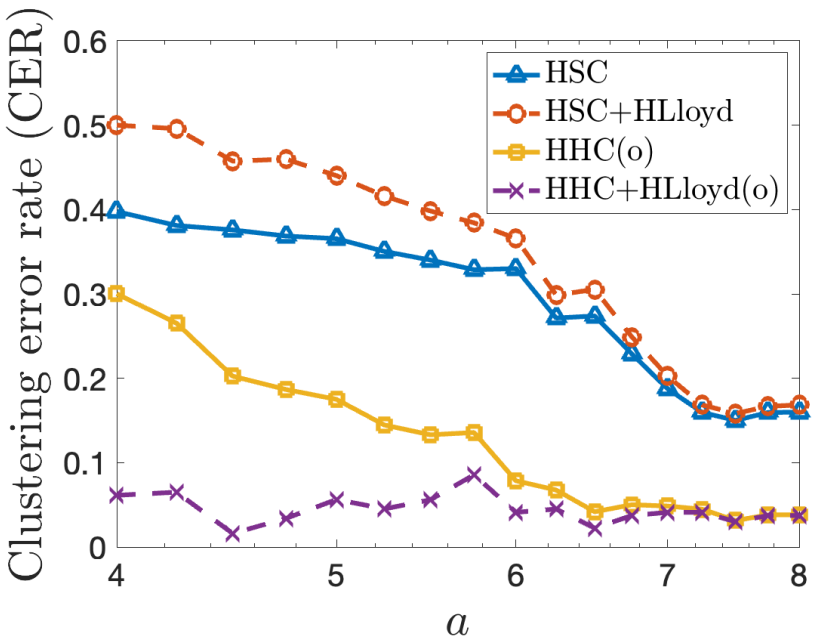


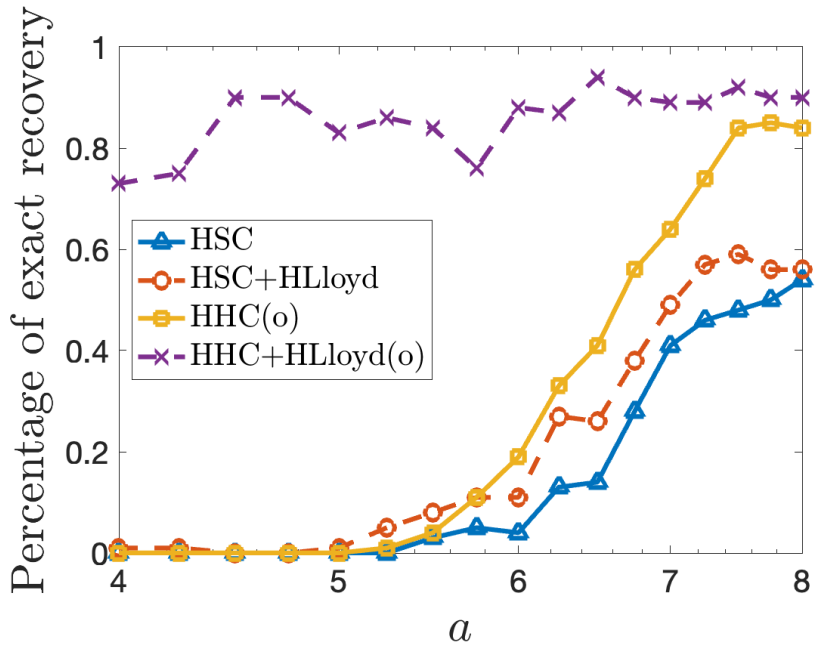


$$n = 100, k = 3, \Delta_{\mathsf{min}} = 40n^{-\delta}, \omega_{\mathsf{max}} = 1$$

Numerical experiments

Stochastic tensor block models





- $n = 100, k = 3$
- $$S^{\star}_{i_1,i_2,i_3} = \begin{cases} 10a \cdot n^{-3/2} \left(1 - \frac{i_1 - 1}{2(k-1)}\right), & i_1 = i_2 = i_3, \\ 0.1a \cdot n^{-3/2}, & \text{otherwise.} \end{cases}$$

Real data example: the flight route network

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

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

*original data: `https://openflights.org/data`

†processed data: `https://github.com/Rungang/HLloyd/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: $(k_1, k_2, k_3) = (5, 5, 5)$
- Airport clustering results using HHC + HLloyd (our method)
  1. Mixture
  2. USA
  3. Europe: LHR (London), MAD (Madrid), CDG (Paris), FCO (Rome), USA: ATL (Atlanta)
  4. China
  5. North America + FRA (Germany)
- Airport clustering results using HSC + HLloyd (Han et al. '22)
  1. Mixture
  2. USA + LHR (UK)
  3. CDG (France), ATL (USA)
  4. China
  5. North America + Europe

Summary

a novel method called High-order HeteroClustering that

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

Summary

a novel method called High-order HeteroClustering that

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

subspace estimation: new algorithm called Deflated-HeteroPCA that

- breaks curse of ill-conditioning & achieves near-optimal statistical guarantees (ℓ_2 and $\ell_{2,\infty}$)
- can also lead to remarkable improvement for tensor PCA

Summary

a novel method called High-order HeteroClustering that

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

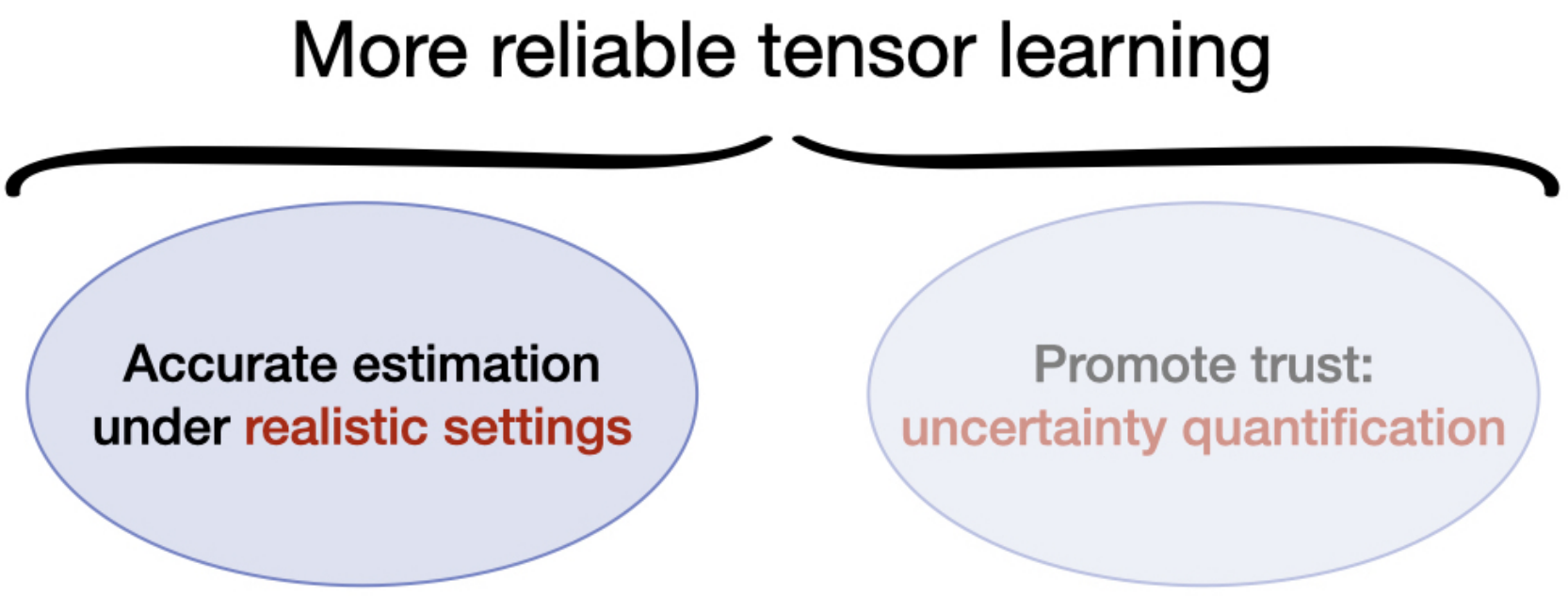
subspace estimation: new algorithm called Deflated-HeteroPCA that

- breaks curse of ill-conditioning & achieves near-optimal statistical guarantees (ℓ_2 and $\ell_{2,\infty}$)
- can also lead to remarkable improvement for tensor PCA

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

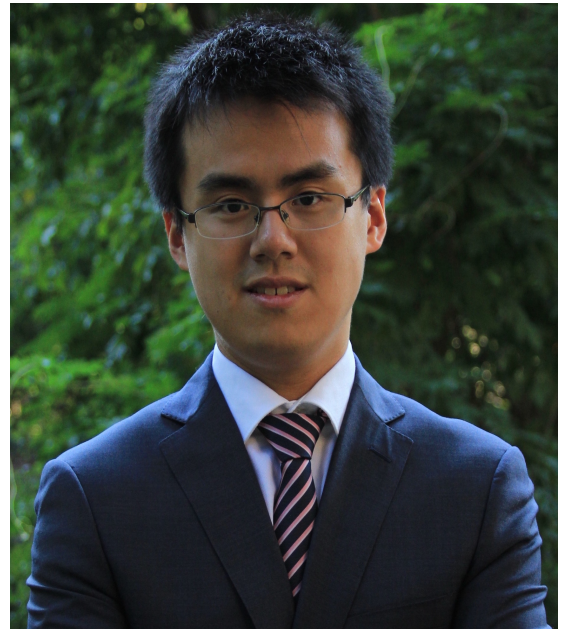
More reliable tensor learning

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 principle components of Tucker low-rank tensors under tensor PCA model

Uncertainty Quantification for Tensor Learning?

We focus on

- inference of principle 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 principle components of Tucker low-rank tensors under tensor regression model

Uncertainty Quantification for Tensor Learning?

We focus on

- inference of principle 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 principle components of Tucker low-rank tensors under tensor regression model

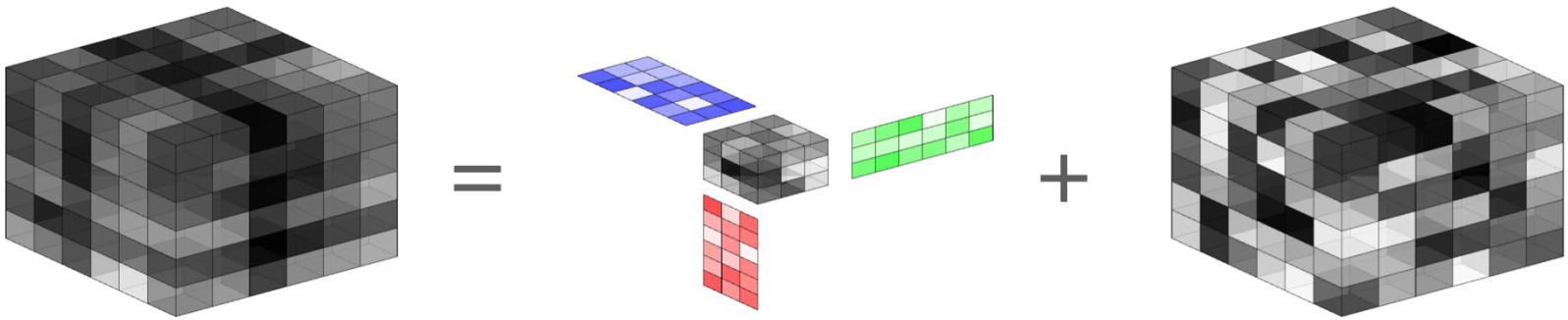
First inference results for Tucker low-rank tensors!

Inference for Tucker low-rank tensor PCA

Tensor PCA model: Observe

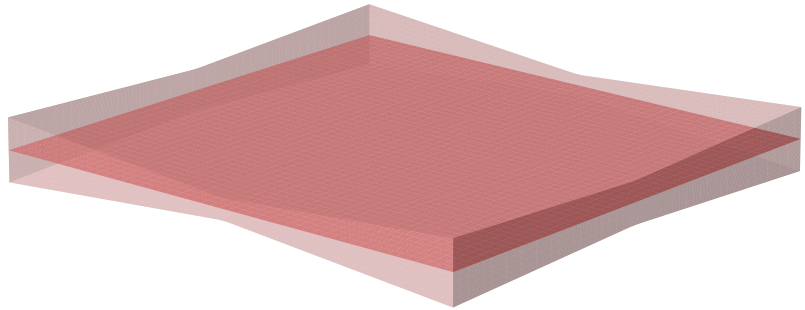
$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{T}}^\star + \boldsymbol{\mathcal{Z}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \tag{1}$$

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

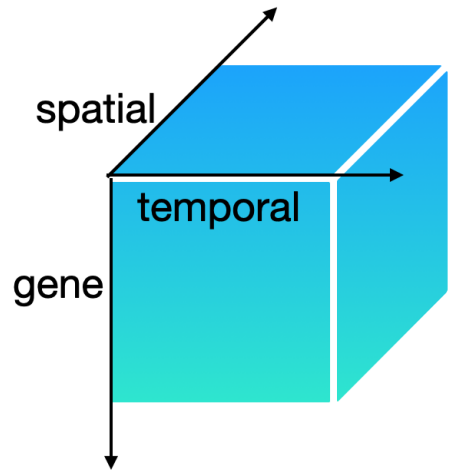


Inference for Tucker low-rank tensor PCA

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)

Review: estimation

Estimation

- Least squares estimator:

$$\widehat{\boldsymbol{\mathcal{T}}} = \underset{\mathsf{rank}(\boldsymbol{\mathcal{T}}) \leq (r_1, r_2, r_3)}{\arg\min} \|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}^2$$

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

Highly non-convex!

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

High-order orthogonal iteration (HOOI)

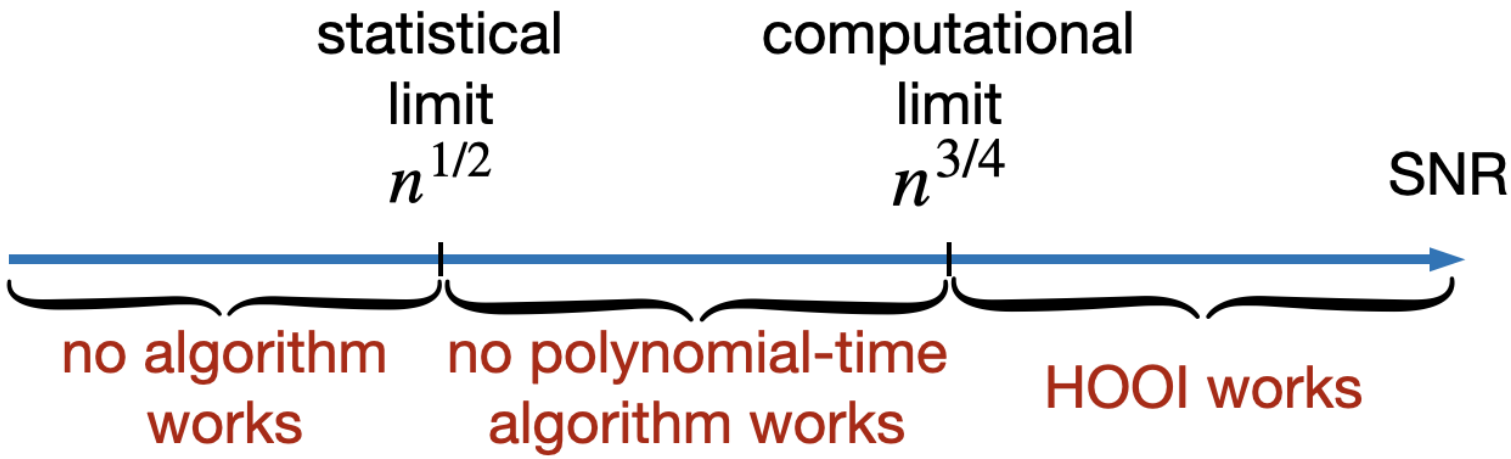
- **Spectral initialization:** $\widehat{\boldsymbol{U}}_i^{(0)}$, $1 \le i \le 3$
- **Sequential updates:** while $t \le t_{\mathsf{max}}$

$$\widehat{\boldsymbol{U}}_1^{(t)} = \mathsf{SVD}_{r_1}\left(\mathcal{M}_1\left(\boldsymbol{\mathcal{A}} \times_2 \widehat{\boldsymbol{U}}_2^{(t-1)\top} \times_3 \widehat{\boldsymbol{U}}_3^{(t-1)\top}\right)\right)$$
$$\widehat{\boldsymbol{U}}_2^{(t)} = \mathsf{SVD}_{r_2}\left(\mathcal{M}_2\left(\boldsymbol{\mathcal{A}} \times_3 \widehat{\boldsymbol{U}}_3^{(t-1)\top} \times_1 \widehat{\boldsymbol{U}}_1^{(t-1)\top}\right)\right)$$
$$\widehat{\boldsymbol{U}}_3^{(t)} = \mathsf{SVD}_{r_3}\left(\mathcal{M}_3\left(\boldsymbol{\mathcal{A}} \times_1 \widehat{\boldsymbol{U}}_1^{(t-1)\top} \times_2 \widehat{\boldsymbol{U}}_2^{(t-1)\top}\right)\right)$$
$$t \leftarrow t+1$$

- **Output:** $\widehat{\boldsymbol{U}}_i := \widehat{\boldsymbol{U}}_i^{(t_{\mathsf{max}})} \longrightarrow$ estimate of $\boldsymbol{U}_i^\star$

$$\widehat{\boldsymbol{\mathcal{G}}} := \boldsymbol{\mathcal{A}} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 \widehat{\boldsymbol{U}}_2^\top \times_3 \widehat{\boldsymbol{U}}_3^\top \longrightarrow \text{estimate of } \boldsymbol{\mathcal{G}}^\star$$

Statistical-computational gap



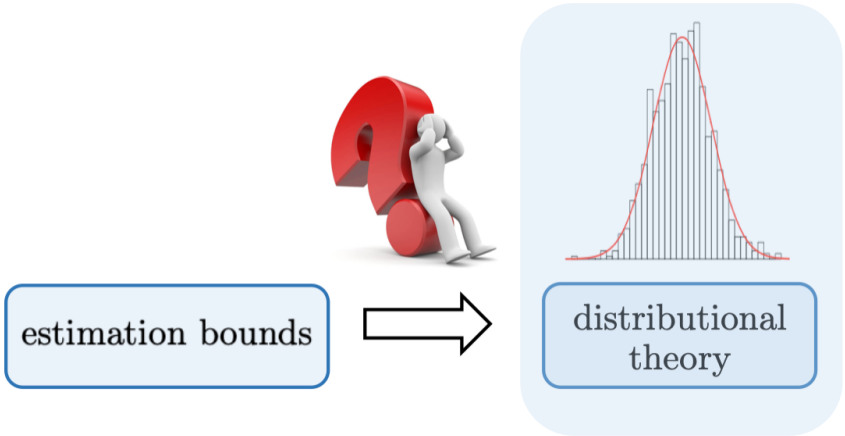
signal strenth $\lambda^\star_{\min} := \lambda_{\min}(\boldsymbol{\mathcal{T}}^\star) = \min_j \sigma_{r_j}\left(\mathcal{M}_j(\boldsymbol{\mathcal{T}}^\star)\right)$

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

SNR $:= \lambda^\star_{\min}/\omega$

Inference for Tucker Low-rank Tensor PCA

- How to make inference?
- Challenge: distribution theory for subspace estimates from non-convex iterations



- The same (or slightly stronger) SNR condition enables us to make inference and no need to debias!

Two-iteration algorithm for inference

- **Input:** Estimators $\widehat{\boldsymbol{U}}_1^{(0)}, \widehat{\boldsymbol{U}}_2^{(0)}, \widehat{\boldsymbol{U}}_3^{(0)}$
- **Sequential updates:** for $t = 0, 1$

$$\widehat{\boldsymbol{U}}_1^{(t+1)} = \mathsf{SVD}_{r_1}\left(\mathcal{M}_1\left(\boldsymbol{\mathcal{A}} \times_2 \widehat{\boldsymbol{U}}_2^{(t)\top} \times_3 \widehat{\boldsymbol{U}}_3^{(t)\top}\right)\right)$$
$$\widehat{\boldsymbol{U}}_2^{(t+1)} = \mathsf{SVD}_{r_2}\left(\mathcal{M}_2\left(\boldsymbol{\mathcal{A}} \times_3 \widehat{\boldsymbol{U}}_3^{(t)\top} \times_1 \widehat{\boldsymbol{U}}_1^{(t)\top}\right)\right)$$
$$\widehat{\boldsymbol{U}}_3^{(t+1)} = \mathsf{SVD}_{r_3}\left(\mathcal{M}_3\left(\boldsymbol{\mathcal{A}} \times_1 \widehat{\boldsymbol{U}}_1^{(t)\top} \times_2 \widehat{\boldsymbol{U}}_2^{(t)\top}\right)\right)$$

- **Output:** Test statistic $\widehat{\boldsymbol{U}}_1 := \widehat{\boldsymbol{U}}_1^{(2)}, \widehat{\boldsymbol{U}}_2 := \widehat{\boldsymbol{U}}_2^{(2)}, \widehat{\boldsymbol{U}}_3 := \widehat{\boldsymbol{U}}_3^{(2)}$, $\widehat{\boldsymbol{\mathcal{G}}} := \boldsymbol{\mathcal{A}} \times_1 \widehat{\boldsymbol{U}}_1^{\top} \times_2 \widehat{\boldsymbol{U}}_2^{\top} \times_3 \widehat{\boldsymbol{U}}_3^{\top}$

Assumptions

- $n_j \asymp n$ for $1 \le j \le 3$.
- **Initialization Error**: $\|\sin\Theta(\widehat{\boldsymbol{U}}_j^{(0)}, \boldsymbol{U}_j^\star)\| \lesssim \sqrt{n}\omega/\lambda_{\min}^\star$ w.h.p.
  - Attainable by HOOI under the essential SNR condition $\lambda_{\min}^\star/\omega \gtrsim n^{3/4}$ (Zhang and Xia, 2018, Theorem 1)!
- **signal-to-noise ratio (SNR)**: $\frac{\lambda_{\min}^\star}{\omega} \gg n^{3/4}$ (ignoring log factors)
- ranks $r_j = O(1)$, condition number $\kappa = O(1)$

Here, $\|\sin\Theta(\widehat{\boldsymbol{U}}_j^{(0)}, \boldsymbol{U}_j^\star)\| = \sqrt{1 - \sigma_{r_j}^2(\boldsymbol{U}_j^{\star\top}\widehat{\boldsymbol{U}}_j^{(0)})}$.

Theoretical guarantees

Theorem 3 (Xia, Zhang, Zhou '22)

$$\frac{\|\sin\Theta(\widehat{\boldsymbol{U}}_j, \boldsymbol{U}_j^\star)\|_{\mathrm{F}}^2 - n_j\omega^2\|\boldsymbol{\Lambda}_j^{\star -1}\|_{\mathrm{F}}^2}{\sqrt{2n_j}\omega^2\|\boldsymbol{\Lambda}_j^{\star -2}\|_{\mathrm{F}}} \sim N(0,1) + \textit{ negligible terms},$$

where $\boldsymbol{\Lambda}_j^\star = \mathrm{diag}(\lambda_1^{\star(j)}, \ldots, \lambda_{r_j}^{\star(j)})$ *is the diagonal matrix containing the singular values of* $\mathcal{M}_j(\boldsymbol{\mathcal{T}}^\star)$.

$$\|\sin\Theta(\widehat{\boldsymbol{U}}_j^{(0)}, \boldsymbol{U}_j^\star)\|_{\mathrm{F}} = \sqrt{r_j - \textstyle\sum_{i=1}^{r_j}\sigma_i^2(\boldsymbol{U}_j^{\star\top}\widehat{\boldsymbol{U}}_j^{(0)})}.$$

Theoretical guarantees

Theorem 3 (Xia, Zhang, Zhou '22)

$$\frac{\|\sin\Theta(\widehat{\boldsymbol{U}}_j, \boldsymbol{U}_j^\star)\|_{\mathrm{F}}^2 - n_j\omega^2\|\boldsymbol{\Lambda}_j^{\star -1}\|_{\mathrm{F}}^2}{\sqrt{2n_j}\omega^2\|\boldsymbol{\Lambda}_j^{\star -2}\|_{\mathrm{F}}} \sim N(0,1) + \textit{ negligible terms},$$

where $\boldsymbol{\Lambda}_j^\star = \mathrm{diag}(\lambda_1^{\star(j)}, \ldots, \lambda_{r_j}^{\star(j)})$ *is the diagonal matrix containing the singular values of* $\mathcal{M}_j(\boldsymbol{\mathcal{T}}^\star)$.

- SNR condition $\lambda_{\min}^\star/\omega \gg n^{3/4}$ is slighter stronger than the one for estimation ($\lambda_{\min}^\star/\omega \gtrsim n^{3/4}$)

Theoretical guarantees

Theorem 3 (Xia, Zhang, Zhou '22)

$$\frac{\|\sin\Theta(\widehat{\boldsymbol{U}}_j, \boldsymbol{U}_j^\star)\|_{\mathrm{F}}^2 - n_j\omega^2\|\boldsymbol{\Lambda}_j^{\star -1}\|_{\mathrm{F}}^2}{\sqrt{2n_j}\omega^2\|\boldsymbol{\Lambda}_j^{\star -2}\|_{\mathrm{F}}} \sim N(0,1) + \textit{ negligible terms},$$

where $\boldsymbol{\Lambda}_j^\star = \mathrm{diag}(\lambda_1^{\star(j)}, \ldots, \lambda_{r_j}^{\star(j)})$ *is the diagonal matrix containing the singular values of* $\mathcal{M}_j(\boldsymbol{\mathcal{T}}^\star)$.

- SNR condition $\lambda_{\min}^\star/\omega \gg n^{3/4}$ is slighter stronger than the one for estimation ($\lambda_{\min}^\star/\omega \gtrsim n^{3/4}$)
- To make inference, we still need to estimate $\boldsymbol{\Lambda}_j^\star$ and ω^2

Theoretical guarantees

Estimates for $\mathbf{\Lambda}_j^\star$ and ω^2:

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

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

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

Theoretical guarantees

Estimates for $\mathbf{\Lambda}_j^\star$ and ω^2:

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

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

$$\widehat{\omega} = \|\boldsymbol{\mathcal{A}} - \underbrace{\boldsymbol{\mathcal{A}} \times_1 \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{U}}_1^\top \times_2 \widehat{\boldsymbol{U}}_2\widehat{\boldsymbol{U}}_2^\top \times_3 \widehat{\boldsymbol{U}}_3\widehat{\boldsymbol{U}}_3^\top}_{\approx \mathcal{T}}\|_{\mathrm{F}}/\sqrt{n_1n_2n_3}.$$

Theorem 3 (Xia, Zhang, Zhou '22)

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

Theoretical guarantees

Estimates for $\mathbf{\Lambda}_j^\star$ and ω^2:

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

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

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

Theoretical guarantees

Estimates for $\mathbf{\Lambda}_j^\star$ and ω^2:

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

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

$$\widehat{\omega} = \|\boldsymbol{\mathcal{A}} - \underbrace{\boldsymbol{\mathcal{A}} \times_1 \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{U}}_1^\top \times_2 \widehat{\boldsymbol{U}}_2\widehat{\boldsymbol{U}}_2^\top \times_3 \widehat{\boldsymbol{U}}_3\widehat{\boldsymbol{U}}_3^\top}_{\approx \mathcal{T}}\|_{\mathrm{F}}/\sqrt{n_1n_2n_3}.$$

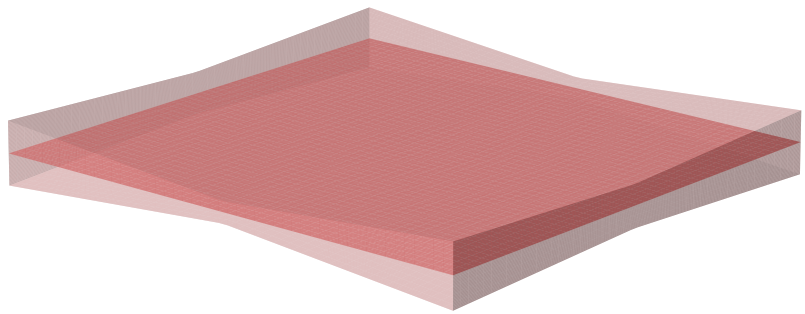
Theorem 4 (Xia, Zhang, Zhou '22)

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

Inference for Tucker Low-rank Tensor PCA

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

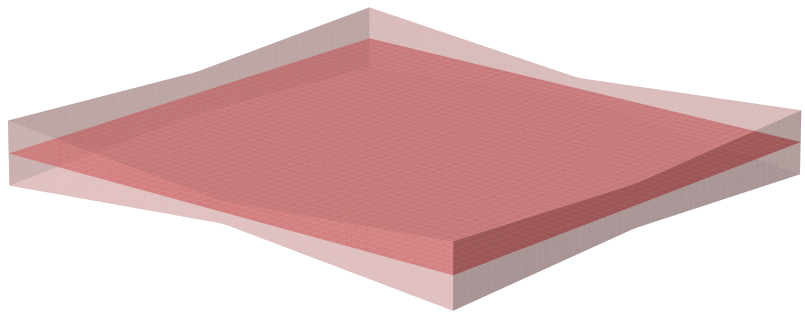
$$\mathrm{CR}_\alpha(\widehat{\boldsymbol{U}}_j) := \left\{ \boldsymbol{V} \in \mathbb{O}_{n_j, r_j} : \|\sin\Theta(\widehat{\boldsymbol{U}}_j, \boldsymbol{V})\|_{\mathrm{F}}^2 \le n_j \widehat{\omega}^2 \|\widehat{\boldsymbol{\Lambda}}_j^{-1}\|_{\mathrm{F}}^2 + z_\alpha \sqrt{2 n_j} \widehat{\omega}^2 \|\widehat{\boldsymbol{\Lambda}}_j^{-2}\|_{\mathrm{F}} \right\}.$$



Inference for Tucker Low-rank Tensor PCA

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

$$\mathrm{CR}_\alpha(\widehat{\boldsymbol{U}}_j) := \left\{ \boldsymbol{V} \in \mathbb{O}_{n_j, r_j} : \|\sin\Theta(\widehat{\boldsymbol{U}}_j, \boldsymbol{V})\|_{\mathrm{F}}^2 \le n_j \widehat{\omega}^2 \|\widehat{\boldsymbol{\Lambda}}_j^{-1}\|_{\mathrm{F}}^2 + z_\alpha \sqrt{2n_j}\widehat{\omega}^2 \|\widehat{\boldsymbol{\Lambda}}_j^{-2}\|_{\mathrm{F}} \right\}.$$



Corollary 1 (Confidence region for tensor PCA)

If $\lambda_{\min}^\star/\omega \gg n^{3/4}$, *then* $\lim_{p\to\infty} \mathbb{P}(\boldsymbol{U}_j^\star \in \mathrm{CR}_\alpha(\widehat{\boldsymbol{U}}_j)) = 1-\alpha$.

Proof Sketch

Assume $\omega = 1$.

$$2\|\sin\Theta(\widehat{\boldsymbol{U}}_1, \boldsymbol{U}_1^\star)\|_{\mathrm{F}}^2 = \|\widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{U}}_1^\top - \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top}\|_{\mathrm{F}}^2 = 2r_1 - 2\big\langle \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{U}}_1^\top, \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top}\big\rangle$$
$$= -2\big\langle \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top}, \widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{U}}_1^\top - \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top}\big\rangle.$$

Step 1: By the spectral representation formula (Xia, 2019, Theorem 1):

$$\widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{U}}_1^\top - \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top} = \boldsymbol{S}_1 + \boldsymbol{S}_2 + \boldsymbol{S}_3 + \sum_{k\geq 4}\boldsymbol{S}_k.$$

Step 2: $\big\langle \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top}, \boldsymbol{S}_1\big\rangle = 0.$

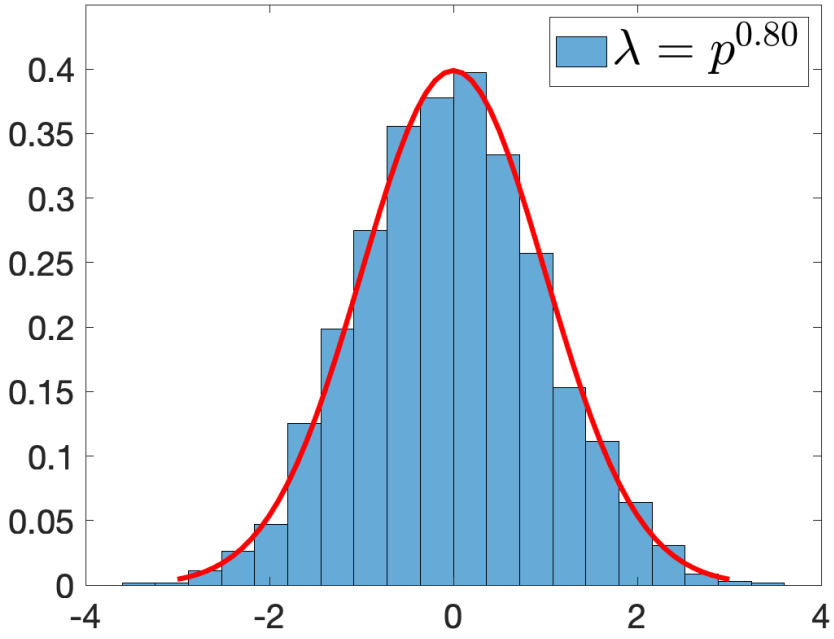
Step 3: $\big\langle \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top}, \boldsymbol{S}_2\big\rangle =$ sum of i.i.d. random variables + high-order terms.

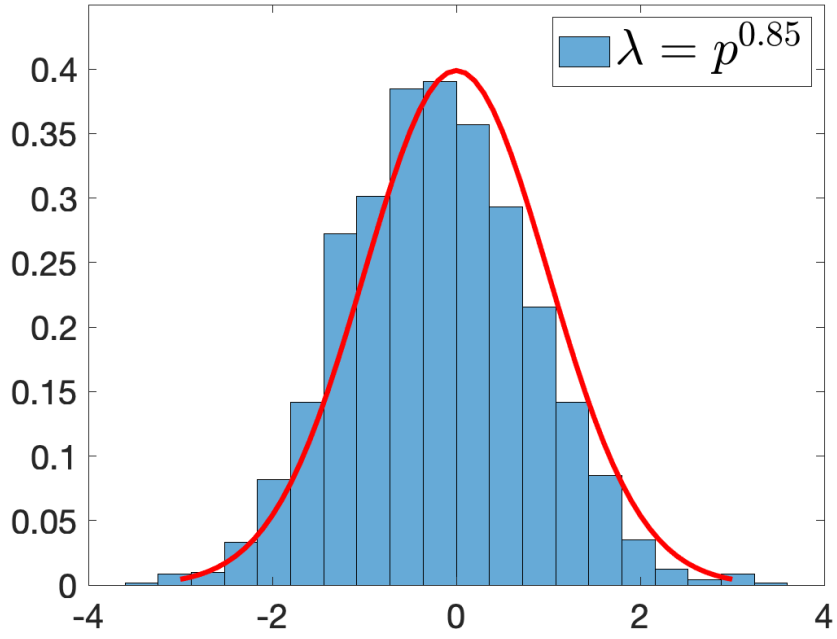
Step 4: $\big|\big\langle \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top}, \boldsymbol{S}_3\big\rangle\big| =$ high-order terms w.h.p.

Step 5: $\|\boldsymbol{S}_k\| \leq (C_1\sqrt{n}/\lambda_{\min})^k$ w.h.p.

$\Rightarrow \big\langle \boldsymbol{U}_1^\star\boldsymbol{U}_1^{\star\top}, \sum_{k\geq 4}\boldsymbol{S}_k\big\rangle =$ high-order terms.

Numerical simulations





Normal approximation of $\dfrac{\|\sin\Theta(\widehat{\boldsymbol{U}}_j,\boldsymbol{U}_j^\star)\|_{\mathrm{F}}^2 - n_j\widehat{\omega}^2\|\widehat{\boldsymbol{\Lambda}}_j^{-1}\|_{\mathrm{F}}^2}{\sqrt{2n_j}\widehat{\omega}^2\|\widehat{\boldsymbol{\Lambda}}_j^{-2}\|_{\mathrm{F}}}$

$n_1 = n_2 = n_3 = n = 200$, $r = 3$, 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!

Summary

- Two-iteration algorithm for inference

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

paper:

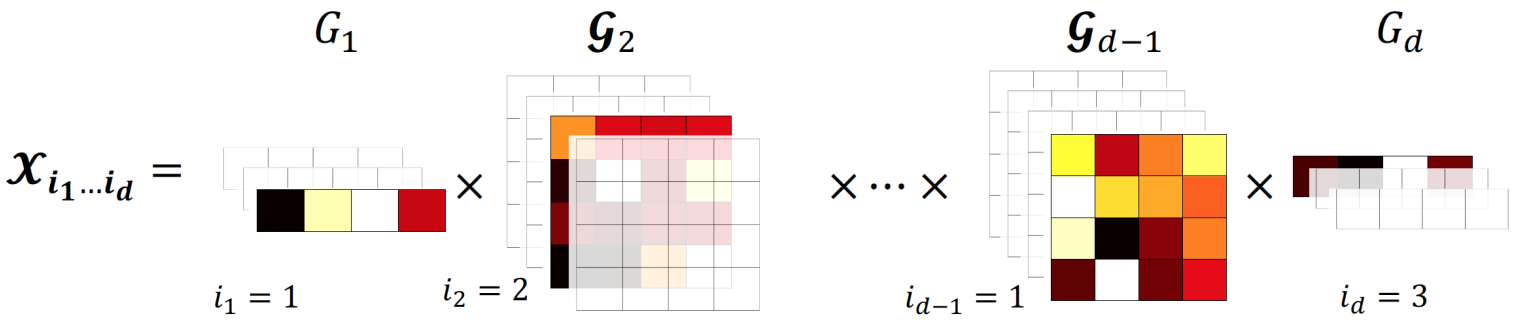
D. Xia, A. Zhang, Y. Zhou, "Inference for Low-rank Tensors – No Need to Debias," *Annals of Statistics*, 2022

Highlight of other works

My research focuses on

- designing computationally fast and statistically accurate algorithms for modern statistical problems
- establishing theoretical foundations for new and existing methods

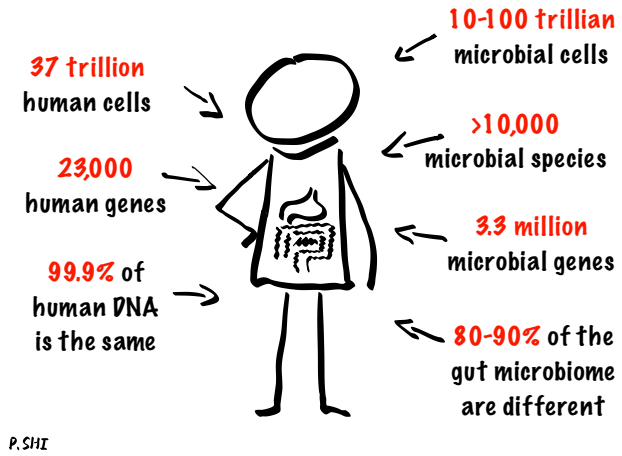
High-order tensor SVD

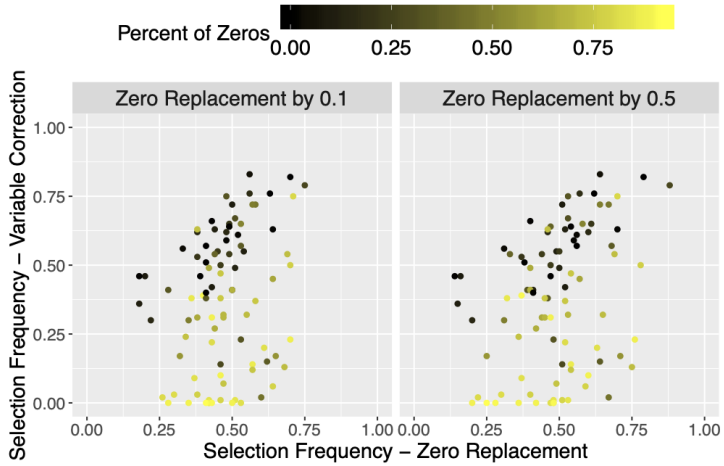


- Novel method: tensor-train orthogonal iteration (TTOI)
- Theory: *minimax optimal* estimation error

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

Microbial compositional data

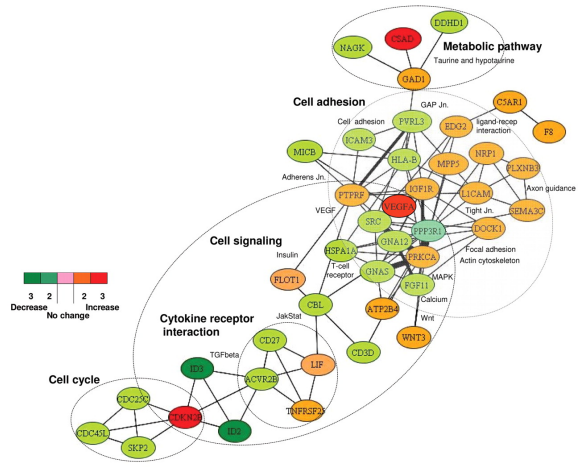




- Goal: Identifying microbial taxa or genes associated with clinicalphenotypes
- Propose log-error-in-variable model & new method
- Theory: optimal guarantee for estimation error

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 Transactions on Information Theory*, 2022

Future directions

- Tensor problems under more realistic assumptions
- Federated statistical learning
- Theoretical foundations of deep learning
- ...

Entry-wise Inference for Rank-1 Tensors

Model: Observe

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

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

Goal: Inference for $T^\star_{i,j,k}$

Entry-wise Inference for Rank-1 Tensors

Theorem 5

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

$$\Big(\frac{\lambda}{\omega}(\widehat{u}_i - u_i^\star),\ \frac{\lambda}{\omega}(\widehat{v}_j - v_j^\star),\ \frac{\lambda}{\omega}(\widehat{w}_k - w_k^\star)\Big)^\top \xrightarrow{\mathrm{d}} N(0, I_3) \quad \text{as} \quad p \to \infty.$$

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

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

Entry-wise Inference for Rank-1 Tensors

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

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

- Asymptotic $(1-\alpha)$ confidence interval for $\mathcal{T}_{ijk}$:
$[\widehat{\mathcal{T}}_{ijk} - z_{\alpha/2}\sigma\sqrt{\widehat{u_i^2}\widehat{v_j^2} + \widehat{v_j^2}\widehat{w_k^2} + \widehat{w_k^2}\widehat{u_i^2}}, \widehat{\mathcal{T}}_{ijk} + z_{\alpha/2}\sigma\sqrt{\widehat{u_i^2}\widehat{v_j^2} + \widehat{v_j^2}\widehat{w_k^2} + \widehat{w_k^2}\widehat{u_i^2}}]$.