Revisit Tensor Decomposition: Statistical Foundations and Computational Guarantees

Anru Zhang
Department of Biostatistics & Bioinformatics
Department of Computer Science
Duke University

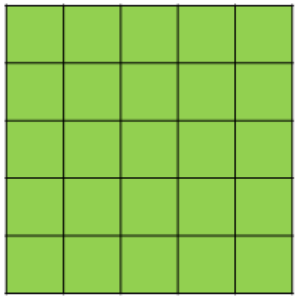
IDS Interdisciplinary Workshop
"Exploring the Foundations: Fundamental AI and Theoretical Machine Learning"
University of Hong Kong

May 26, 2025

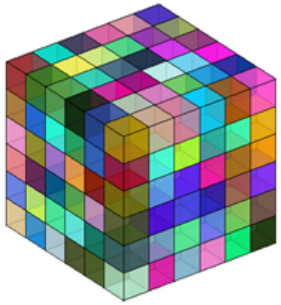
Introduction

- Tensors are arrays with multiple directions.

Order-1 tensor: vector



Order-2 tensor: matrix



Order-3 tensor

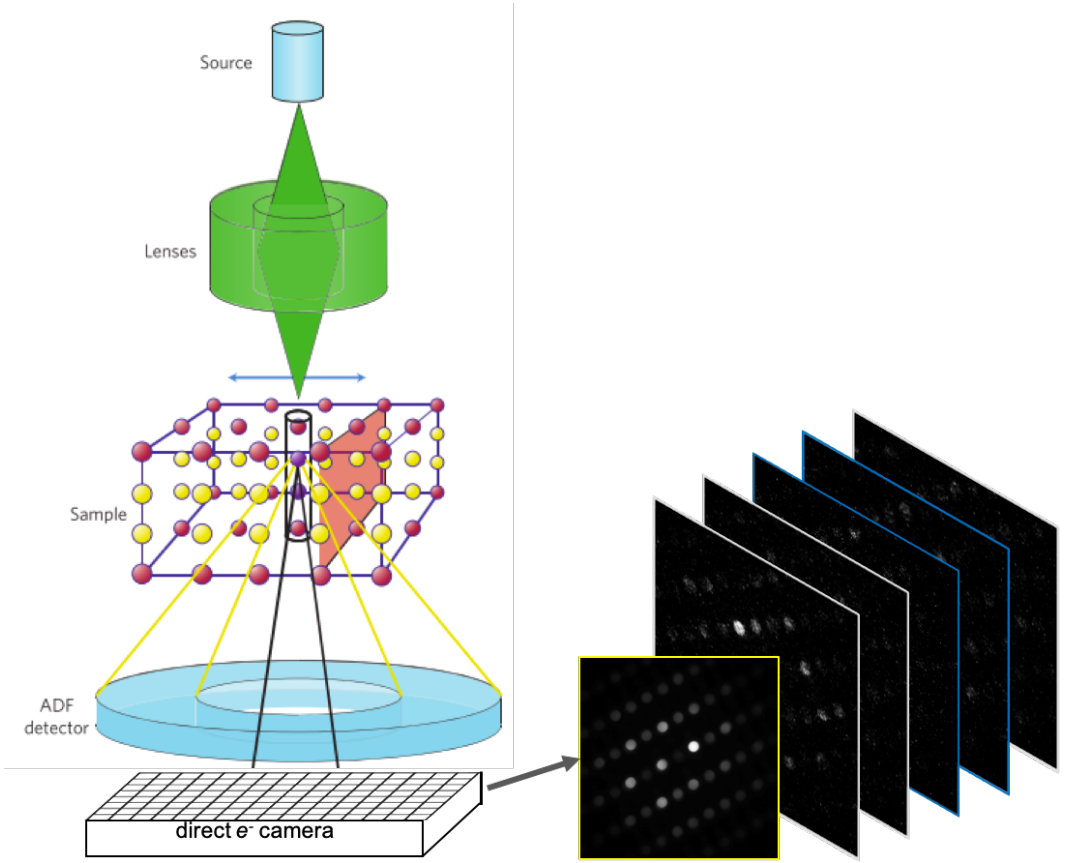
- Tensors of order three or higher are called high-order tensors.

$$\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times \cdots \times p_d}, \qquad \boldsymbol{\mathcal{A}} = (A_{i_1 \cdots i_d}), \qquad 1 \leq i_k \leq p_k, \quad k = 1, \ldots, d.$$

- A survey on tensor algebra (Kolda & Bader, 2009)

Scientific Problems with Tensors

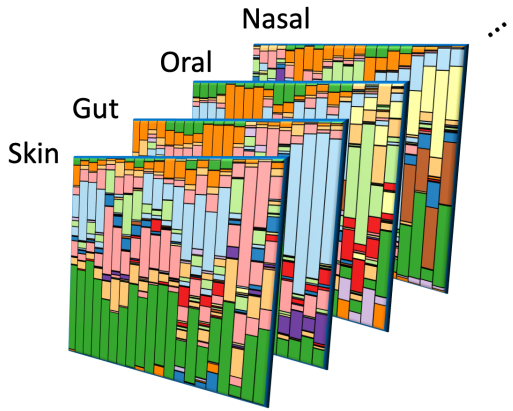
- 4D scanning transmission electron microscopy (4D-STEM)
- Goal: study the nanometer level structure of materials



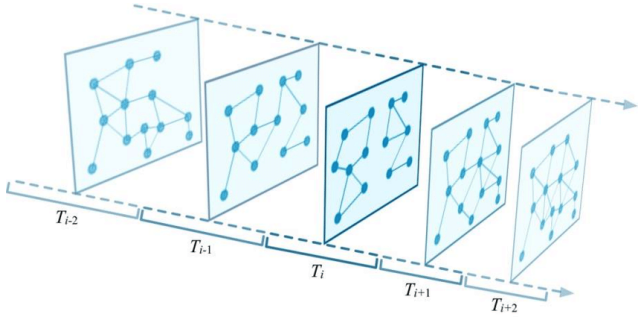
Data source: Voyles Group at UW-Madison

More High-Order Data Are Emerging

- Sequencing data analysis



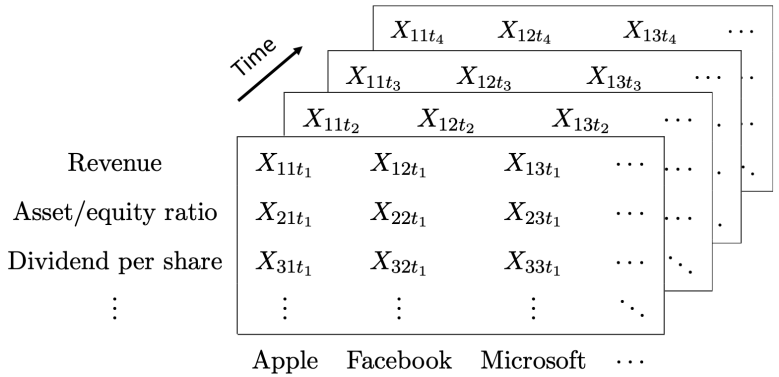
- Dynamic social network



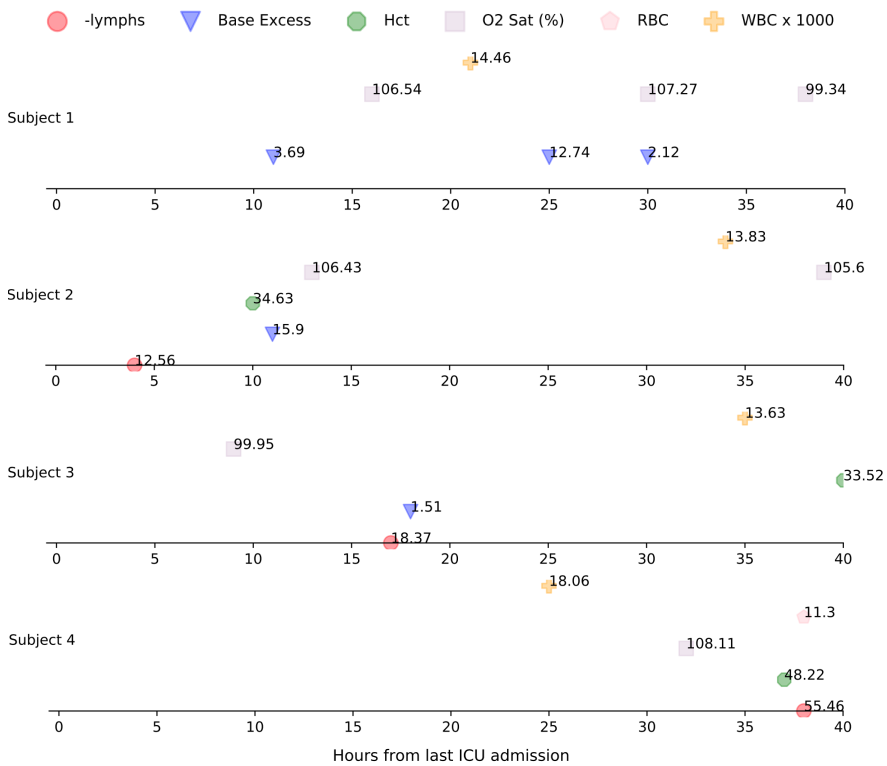
Picture source: Dakiche, N., Tayeb, F. B. S., Slimani, Y. ,& Benatchba, K. 2018 FUZZ-IEEE.

More High-Order Data Are Emerging

- Matrix-valued time series



- Multivariate/high-dimensional longitudinal data (Electronic health records, wearable measurements, etc)



Interaction Pursuit

- Model (Hao, **Z.**, Cheng, *IEEE TIT*, 2018)

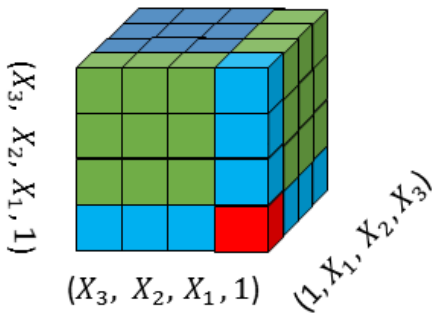
$$y = \beta_0 + \underbrace{\sum_i X_i\beta_i}_{\text{Main effect}} + \underbrace{\sum_{i,j} \gamma_{ij} X_i X_j}_{\text{Pairwise interaction}} + \underbrace{\sum_{i,j,k} \eta_{ijk} X_i X_j X_k}_{\text{Triple-wise}} + \varepsilon.$$

- Rewrite as

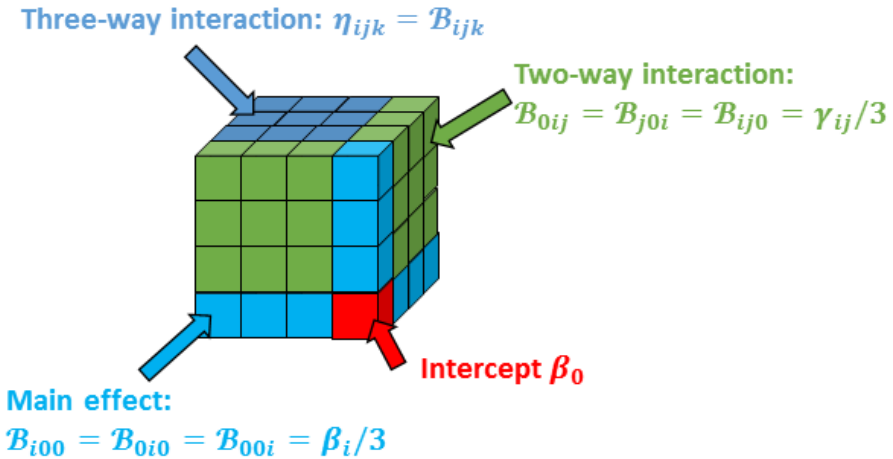
$$y = \langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{B}} \rangle + \varepsilon.$$

Tensorized Covariates

$\boldsymbol{\mathcal{X}} = (1, \boldsymbol{X}) \circ (1, \boldsymbol{X}) \circ (1, \boldsymbol{X})$



High-order Interaction Effects $\boldsymbol{\mathcal{B}}$



High Order Enables Solutions for Harder Problems

Estimation of Mixture Models

- A mixture model incorporates subpopulations in an overall population.
- Examples:
  - Gaussian mixture model (Lindsay & Basak, 1993; Hsu & Kakade, 2013; Wu & Yang, 2019)
  - Topic modeling (Arora et al, 2013)
  - Hidden Markov process (Anandkumar, Hsu, & Kakade, 2012)
  - Independent component analysis (Miettinen, et al., 2015)
  - Additive index model (Balasubramanian, Fan & Yang, 2018)
  - Mixture regression model (De Veaux, 1989; Jordan & Jacobs, 1994)
  - Generative model (Chen, Li, Li, **Z., 2022**)
  - ...
- Method of Moment (MoM):
  - First moment $\rightarrow$ vector;
  - Second moment $\rightarrow$ matrix;
  - **High-order moment $\rightarrow$ high-order tensors.**

Fast Matrix Multiplication

nature › articles › article

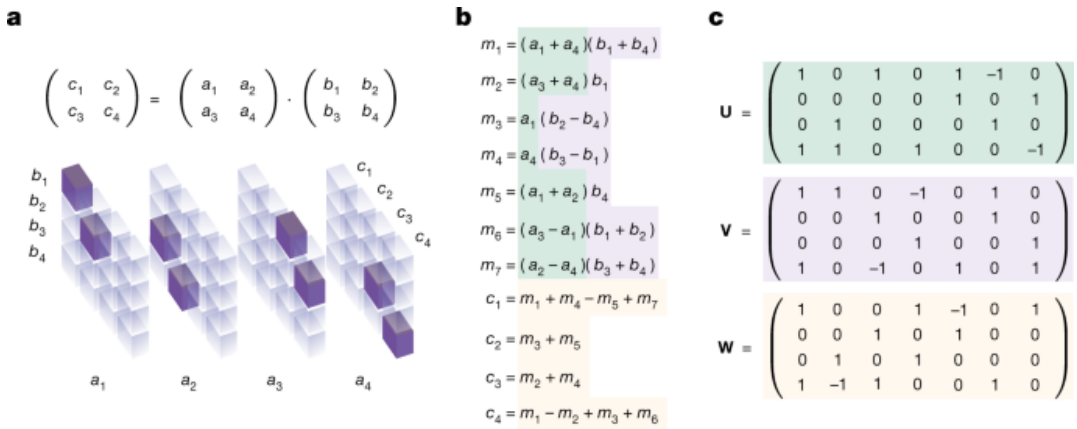
Article | Open Access | Published: 05 October 2022

Discovering faster matrix multiplication algorithms with reinforcement learning

Alhussein Fawzi ✉, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J. R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, David Silver, Demis Hassabis & Pushmeet Kohli

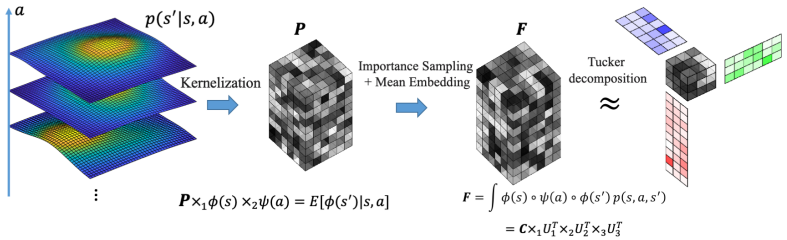
Nature **610**, 47–53 (2022) | Cite this article

1971 Altmetric | Metrics



Tensors in Machine Learning

- Markov decision process (MDP) in reinforcement learning

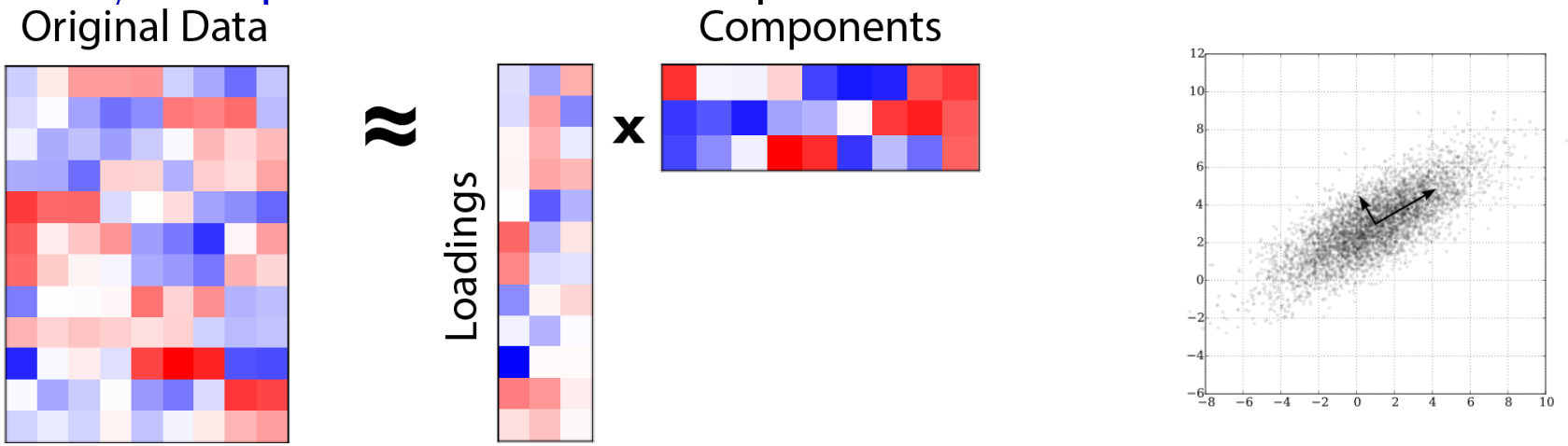


Ni, **Zhang**, Duan, Wang (2020). Learning Good State and Action Representations via Tensor Decomposition.

- **Tensor provides a testing ground to study complicated phenomena/methods in modern machine learning.**
  - Overparemetrization (Wang, Wu, Lee, Ma, Ge, 2020)
  - implicit bias/implicit regularization (Noam Razin, Asaf Maman, Nadav Cohen, 2021)
  - neuron collapse (Qu, Zhai, Li, Zhang, and Zhu, 2021)
  - generative models (Chen, Li, Li, **Z.**, 2022)
  - ...

Dimension reduction, SVD, PCA

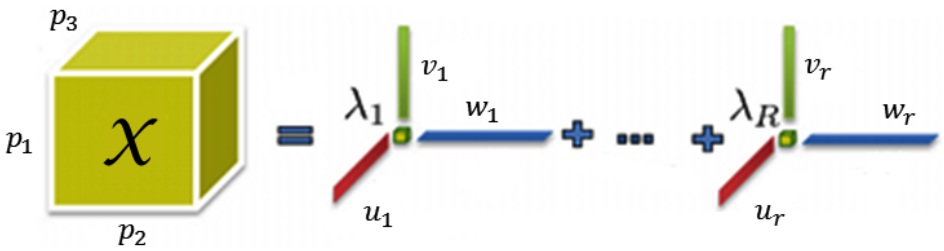
- Singular value decomposition (SVD): one of the most important tools for dimension reduction.
- Goal: Find the most representative low-dimensional structure of the data matrix.
- Closely related to Principal component analysis (PCA): Find the one/multiple directions that explain most of the variance.



- Tensor PCA model (Montanari & Richard, 2014): $r = 1$
- Statistical framework for tensor SVD in general rank is not well defined or solved.

Tensor Decomposition

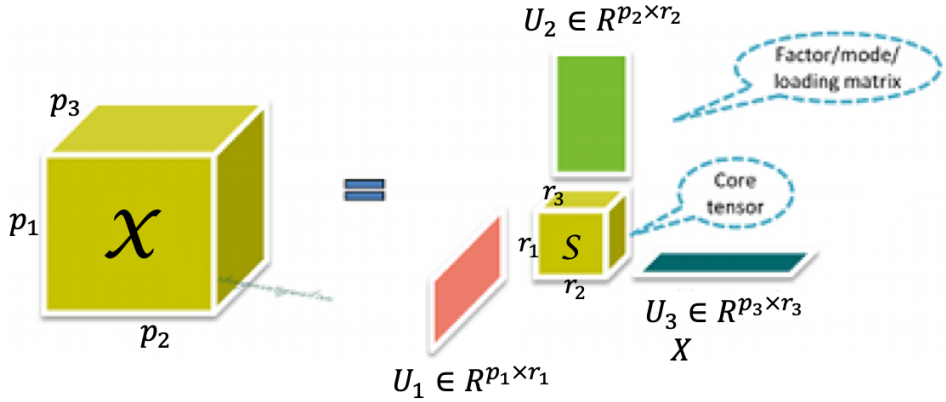
Canonical polyadic (CP) rank:
Smallest r s.t.



$$\boldsymbol{\mathcal{X}} = \sum_{i=1}^{r} \lambda_i u_i \circ v_i \circ w_i$$

Extension of matrix SVD:
$\boldsymbol{X} = \sum_{i=1}^{r} \lambda_i u_i v_i^\top$

Tucker Rank: Smallest (r_1, r_2, r_3) s.t.



$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{S}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3$$

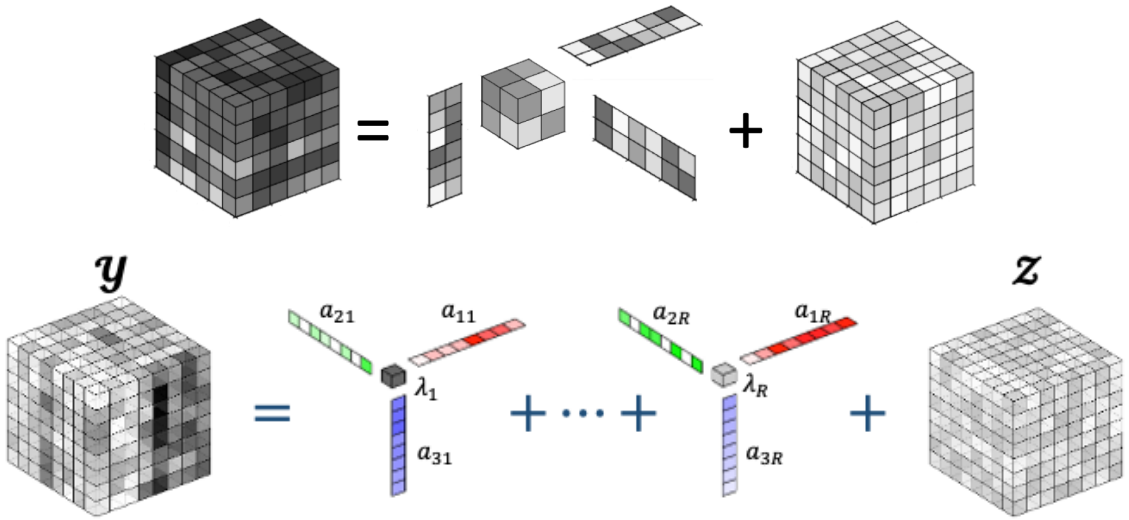
Extension of matrix SVD:
$\boldsymbol{X} = \boldsymbol{U}_1 \boldsymbol{S} \boldsymbol{U}_2^\top$

Foundamental Models for Tensor Decomposition

This talk focuses on:

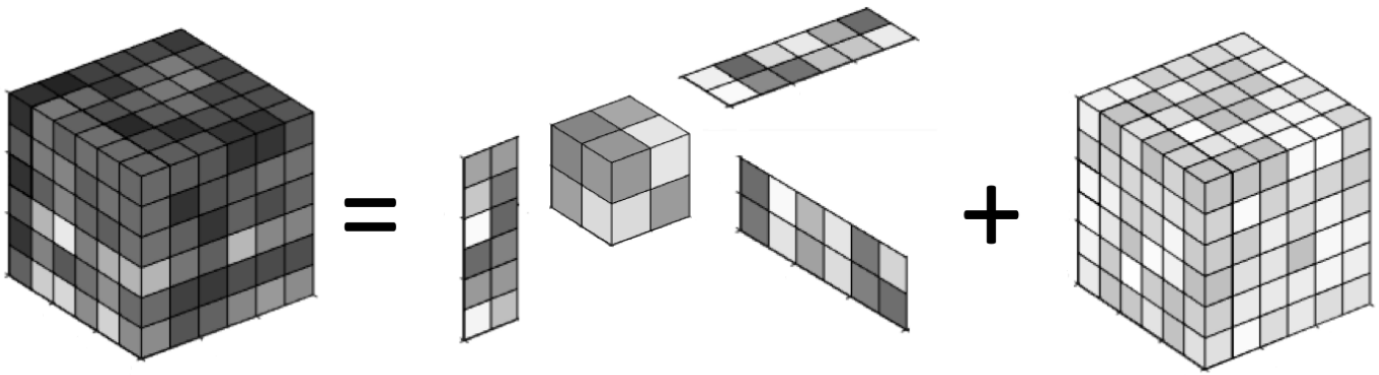
$$\mathcal{Y} = \mathcal{X} + \mathcal{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$$

- $\mathcal{Y}$: observation tensor
- $\mathcal{X}$: signal tensor of low Tucker rank or CP rank
- $\mathcal{Z}$: noise tensor; $\mathcal{Z} \overset{iid}{\sim} N(0, \sigma^2)$



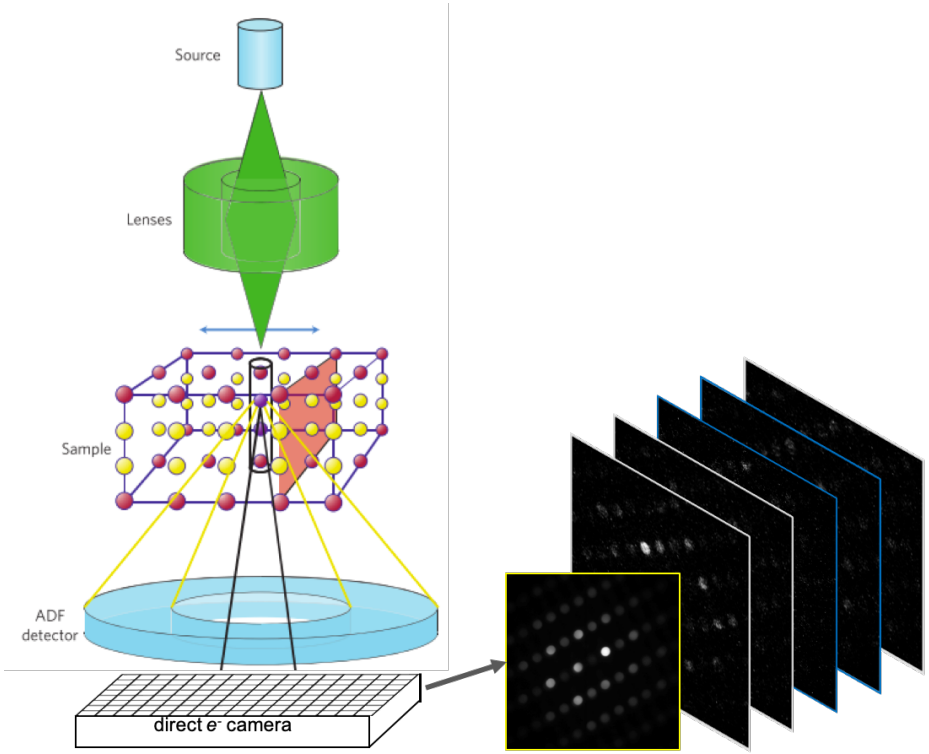
Goal: statistical inference on low-rank components of $\mathcal{X}$.

Part I: Tucker Tensor Decomposition: A Statistical and Computational Perspective



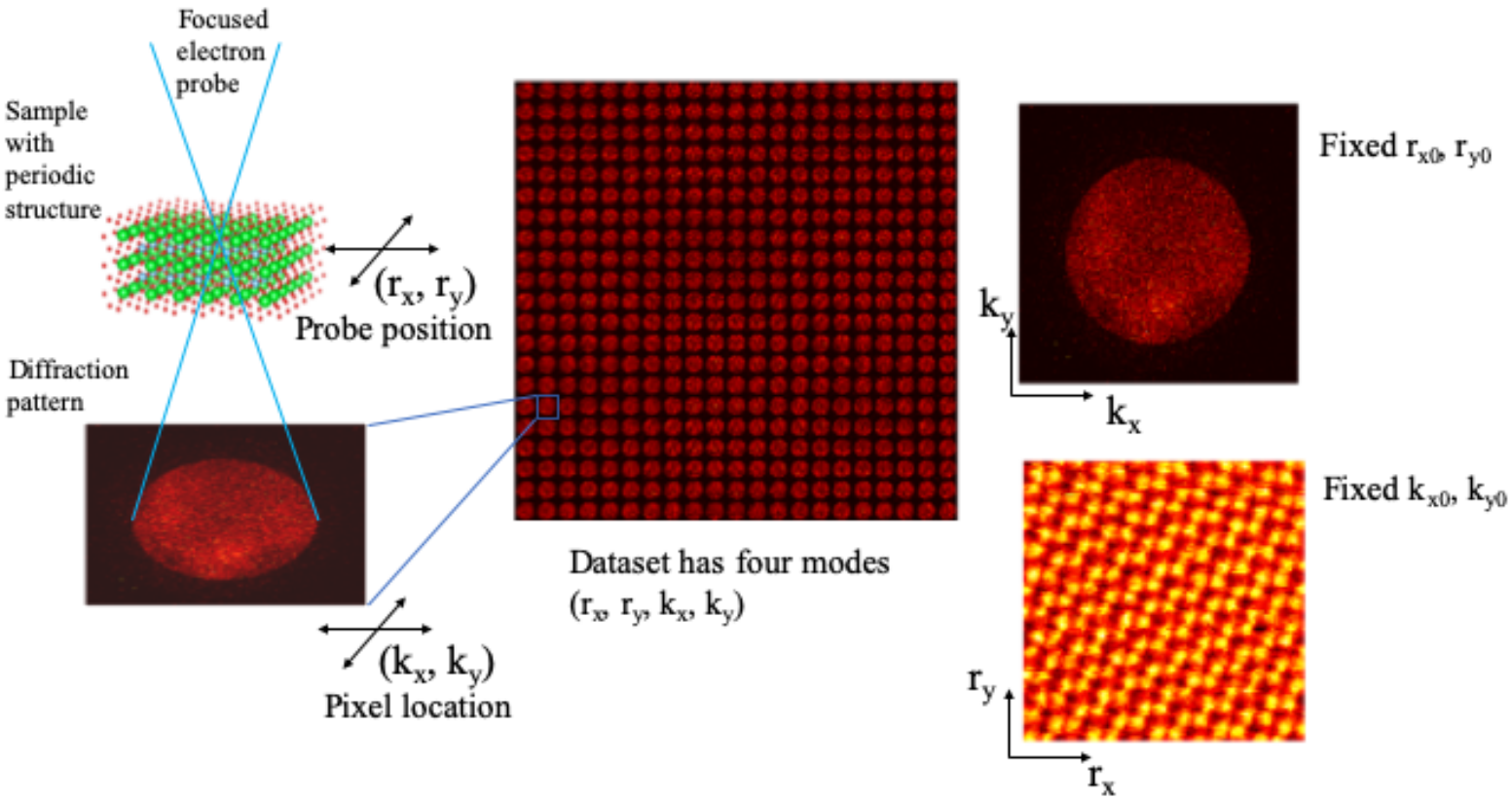
Motivation: Computational Image Denoising

4D Scanning Transmission Electron Microscopy (4D-STEM)



- Final data: 4-D tensor of pixels.
- Goal: Recovery of atomic or molecular structures
- Usage: Identify abnormalities, improve materials,...

4D-STEM Image Denoising



- 4D-STEM images are photon-limited and highly noisy
  $\rightarrow$ Adequate denoising is crucial!
- Tensorial structure
  $\rightarrow$ Denoise using tensor SVD method.

SVD for Tensor with Tucker Low Rank

- A general statistical framework for SVD for tensor with Tucker low rank:

$$\mathcal{Y} = \mathcal{X} + \mathcal{Z},$$

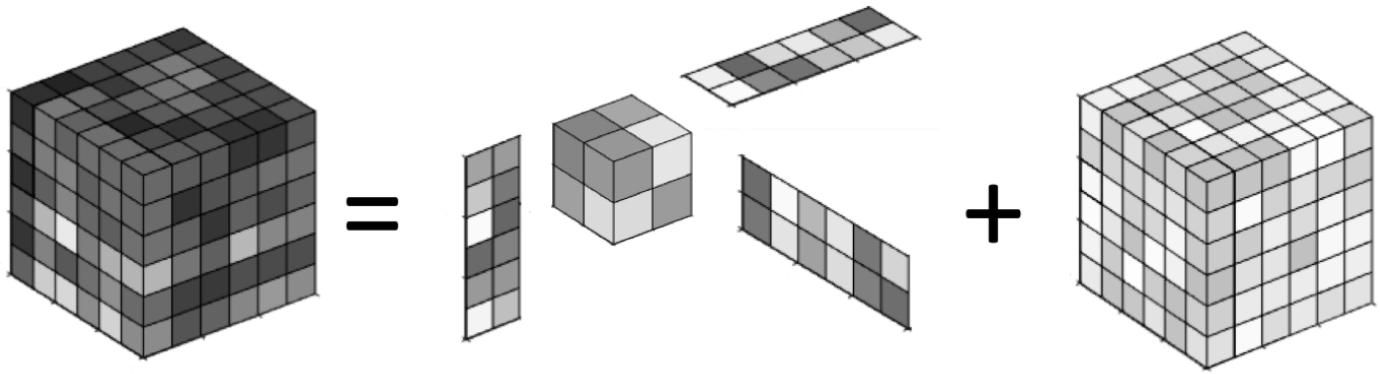
  where
  - $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is the observation (e.g., noisy images);
  - $\mathcal{X}$ is a low-rank tensor (e.g., ground truth images).
  - $\mathcal{Z}$ is the noise;
- Goal: **recovery** of the high-dimensional low-rank structure $\mathcal{X}$.

Model

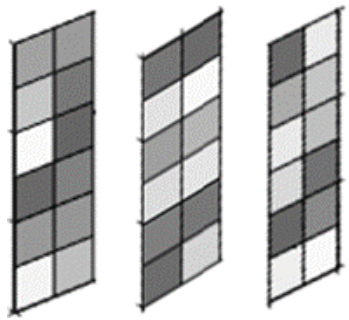
- Observations: $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$,

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{S}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3 + \boldsymbol{\mathcal{Z}},$$

$$\boldsymbol{\mathcal{Z}} \overset{iid}{\sim} N(0, \sigma^2), \quad \boldsymbol{U}_k \in \mathbb{O}_{p_k, r_k}, \quad \boldsymbol{\mathcal{S}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}.$$



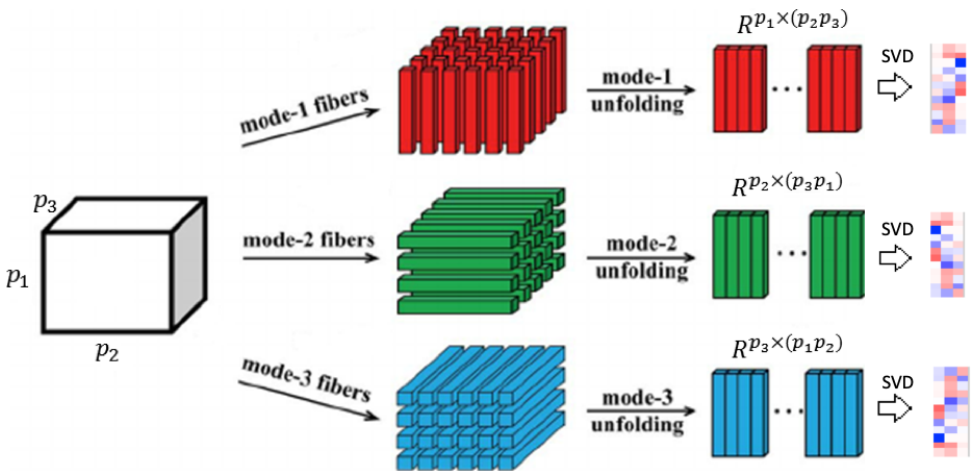
- Goal: estimate the original tensor $\boldsymbol{\mathcal{X}}$ and loadings $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$.



Methodology - Step 1: Initialization

- We can apply higher-order orthogonal iteration (HOOI).
- Initialize with HOSVD (De Lathauwer, Moor, and Vandewalle, SIAM. J. Matrix Anal. & Appl. 2000a)

$$\hat{\boldsymbol{U}}_k^{(0)} = \text{SVD}_{r_k}\left(\mathcal{M}_k(\boldsymbol{\mathcal{Y}})\right), \quad k = 1, 2, 3.$$

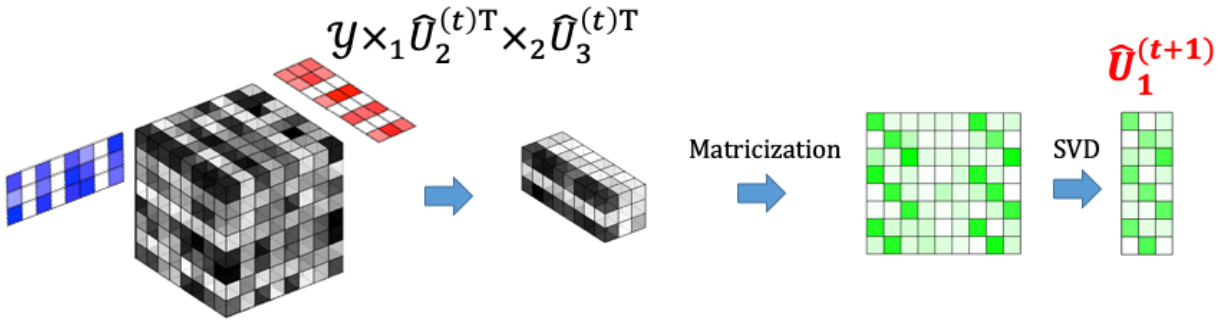


- Initialization $\hat{\boldsymbol{U}}_k^{(0)}$ provides a starting point.
- Faster procedure: sequential truncated HOSVD (Vannieuwenhoven, Vandebril, & Meerbergen, 2012)

Methodology - Step 2: Power Iteration

- **Repeat** Let $t = t + 1$. Calculate

$$\hat{\boldsymbol{U}}_1^{(t)} = \text{SVD}_{r_1}\Big(\mathcal{M}_1(\boldsymbol{\mathcal{Y}} \times_2 (\hat{\boldsymbol{U}}_2^{(t-1)})^\top \times_3 (\hat{\boldsymbol{U}}_3^{(t-1)})^\top)\Big),$$



$$\hat{\boldsymbol{U}}_2^{(t)} = \text{SVD}_{r_2}\Big(\mathcal{M}_2(\boldsymbol{\mathcal{Y}} \times_1 (\hat{\boldsymbol{U}}_1^{(t)})^\top \times_3 (\hat{\boldsymbol{U}}_3^{(t-1)})^\top)\Big),$$

$$\hat{\boldsymbol{U}}_3^{(t)} = \text{SVD}_{r_3}\Big(\mathcal{M}_3(\boldsymbol{\mathcal{Y}} \times_1 (\hat{\boldsymbol{U}}_1^{(t)})^\top \times_2 (\hat{\boldsymbol{U}}_2^{(t)})^\top)\Big).$$

  Until $t = t_{\max}$ or convergence.

- Power iteration refines the initializations.

Theoretical Analysis

Define signal-to-noise ratio (SNR): λ/σ,

$$\text{Signal strength: } \lambda = \min_{k=1,2,3} \sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$$

$$= \text{least non-zero singular value of } \mathcal{M}_k(\boldsymbol{\mathcal{X}}), k = 1, 2, 3,$$

$$\text{Noise level: } \sigma = \text{SD}(\boldsymbol{\mathcal{Z}}).$$

Theorem (Z. and Xia, *IEEE Trans. Inform. Theory*, 2018)

Under strong SNR, $\lambda/\sigma > Cp^{3/4}$,

(Recovery of loadings) $$\mathbb{E} \min_{O \in \mathbb{O}_r} \left\| \hat{\boldsymbol{U}}_k - \boldsymbol{U}_k O \right\|_F \leq \frac{C\sqrt{p_k r_k}}{\lambda/\sigma}, \quad k = 1, 2, 3;$$

(Recovery of $\boldsymbol{\mathcal{X}}$) $$\mathbb{E} \left\| \hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}} \right\|_F^2 \leq C \left(p_1 r_1 + p_2 r_2 + p_3 r_3 \right) \sigma^2.$$

- These upper bounds are rate-optimal!

Brief of Theoretical Results

Order-d Tensor SVD exhibits three phases (Z. and Xia, 2018):

- (Strong SNR) $\lambda/\sigma \geq Cp^{d/4}$,
  $\rightarrow$ an efficient algorithm for optimal estimation of loadings and $\boldsymbol{\mathcal{X}}$.

- (Weak SNR) $\lambda/\sigma < cp^{1/2}$,
  $\rightarrow$ no algorithm can consistently recover loadings or $\boldsymbol{\mathcal{X}}$.

- (Moderate SNR) $p^{1/2} \ll \lambda/\sigma \ll p^{d/4}$,
  $\rightarrow$ MLE optimally estimates loadings and $\boldsymbol{\mathcal{X}}$ with $\gg$ polynomial time.
  $\rightarrow$ No polynomial-time algorithm performs consistently.

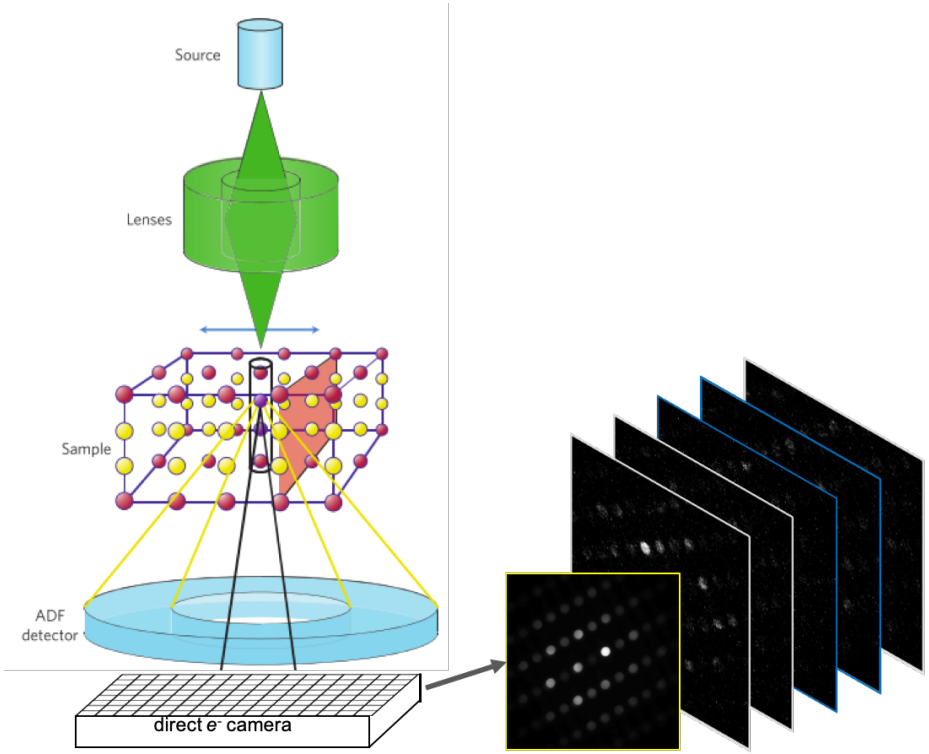
Remark

- $d = 2$, i.e. matrix SVD: computation and statistical gap closes.
- $d \geq 3$: tensor SVD is with not only statistical, but also computational challenges.

Computational Image Denoising

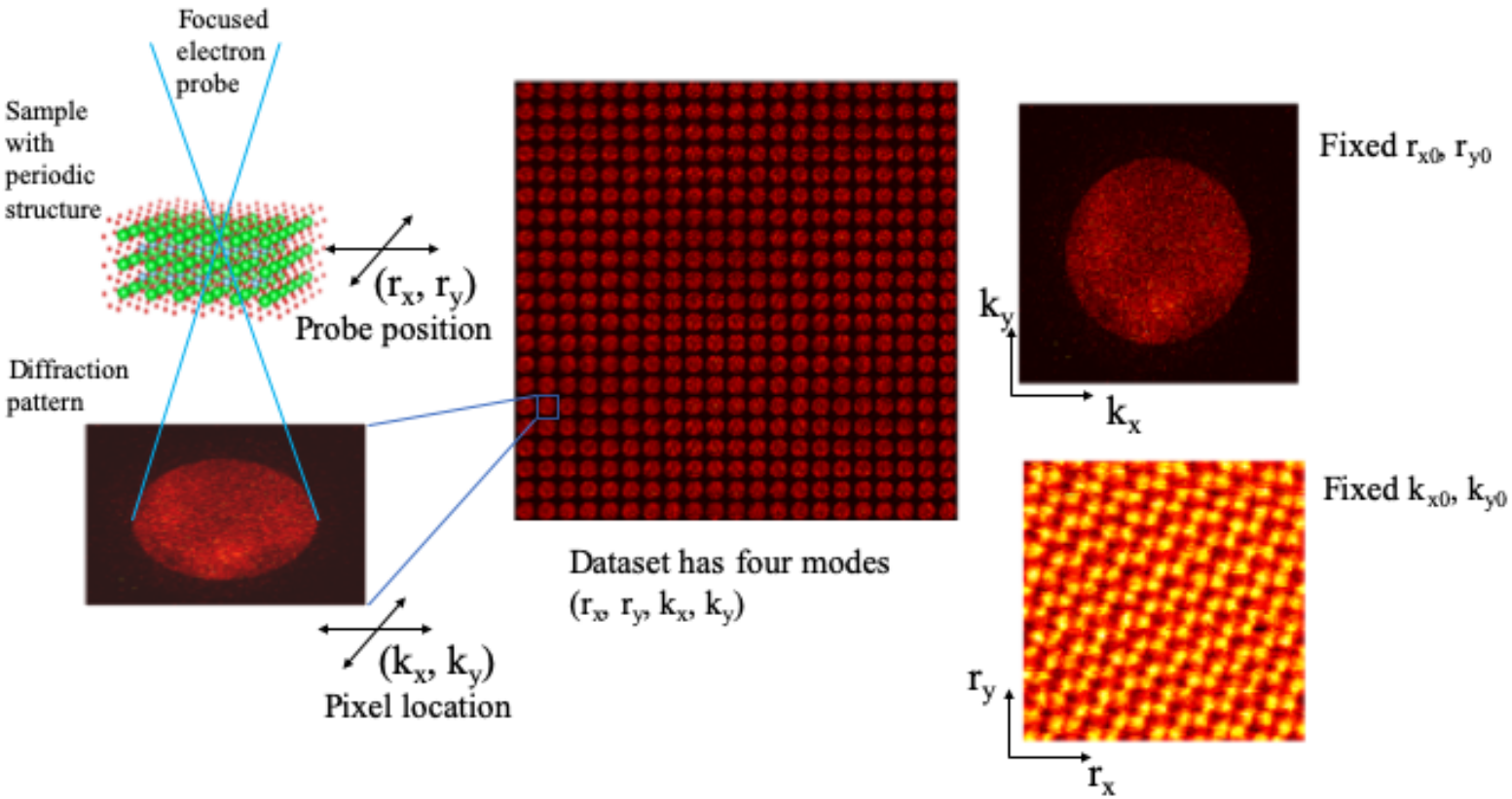
Computational Image Denoising

4D Scanning Transmission Electron Microscopy (4D-STEM)



- Final data: 4-D tensor of pixels.
- Goal: Recovery of atomic or molecular structures
- Usage: Identify abnormalities, improve materials,...

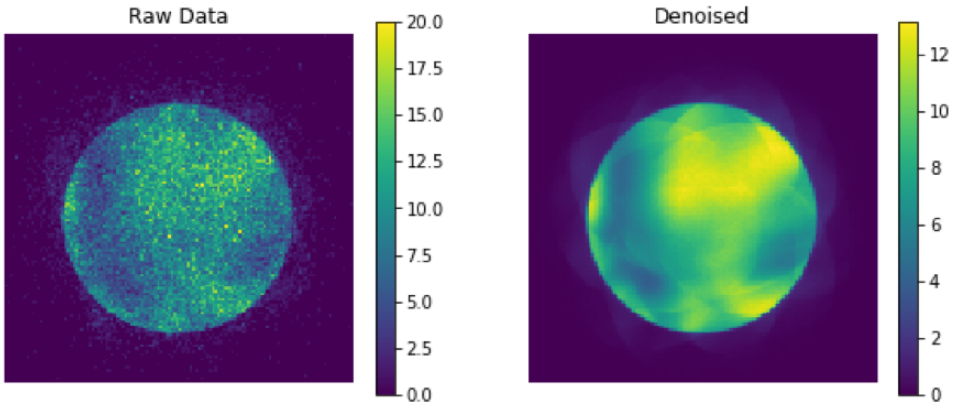
4D-STEM Image Denoising



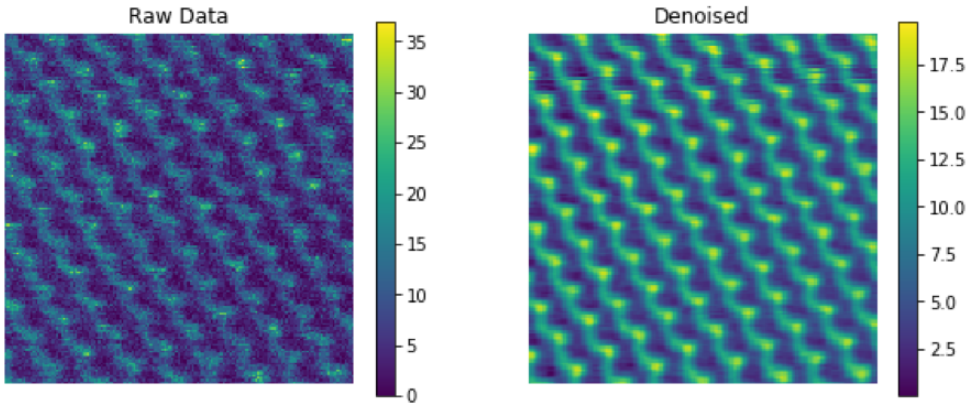
- 4D-STEM images are photon-limited and highly noisy
  $\rightarrow$ Adequate denoising is crucial!
- Tensorial structure
  $\rightarrow$ Denoise using tensor SVD method.

Results on Whole 4D Real Data

Denoised results on the whole 4D dataset

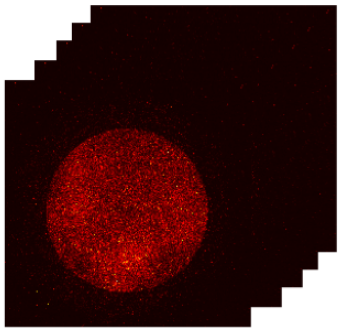


Noisy single CBED vs denoised CBED

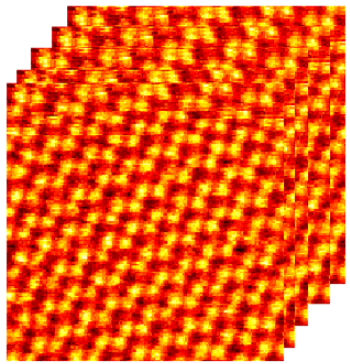


3D image stack from 4D datacube:

- (r_x, r_y, k_x, k_y) -> $(r_x, r_y, [k_x, k_y])$, concatenate the two dimensions of k_x and k_y into one single dimension.



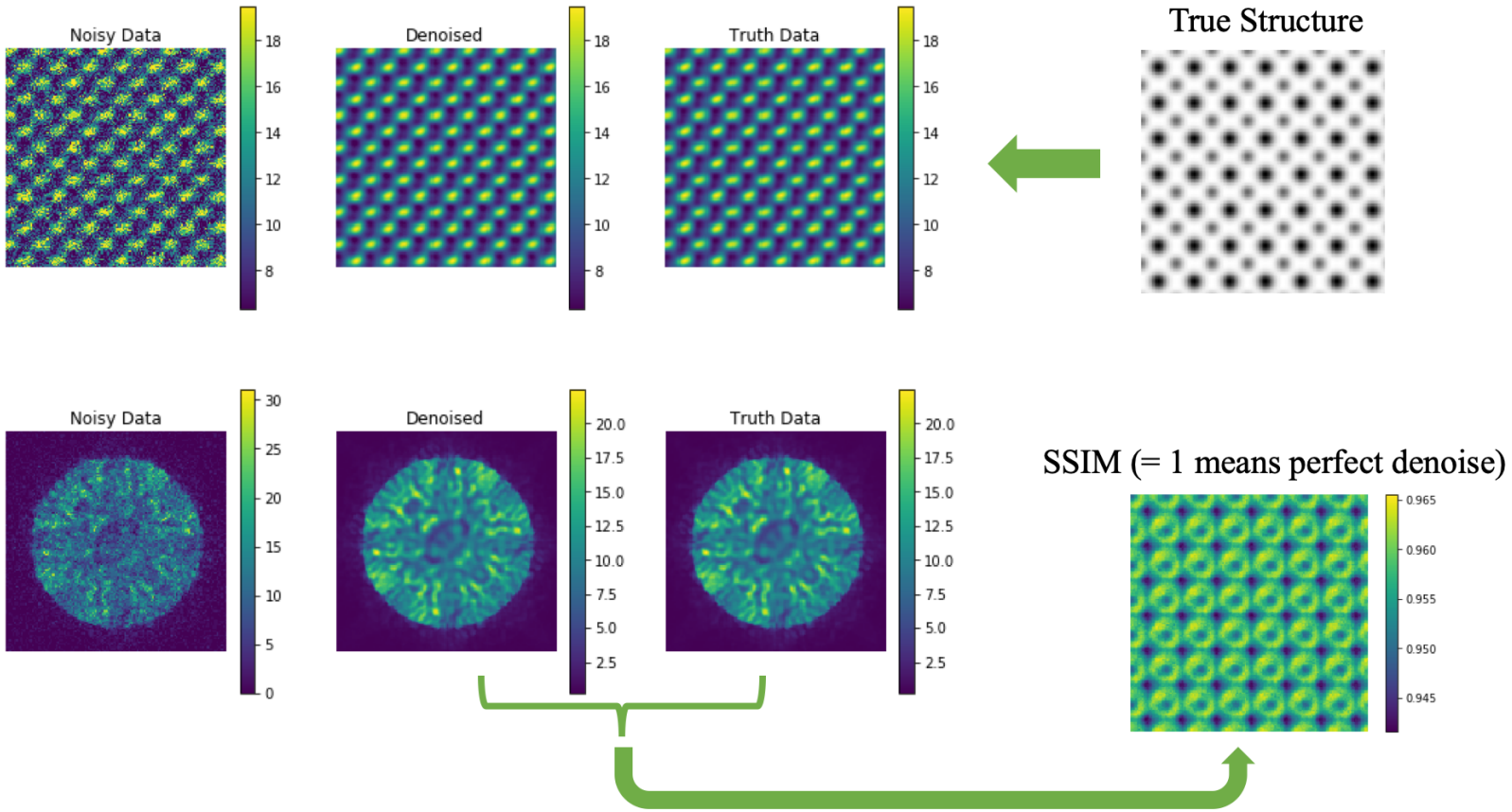
Previous scheme, low redundancy inside each frame.



Current scheme, rearranged dimension, high redundancy inside each frame, same redundancy across different frames.

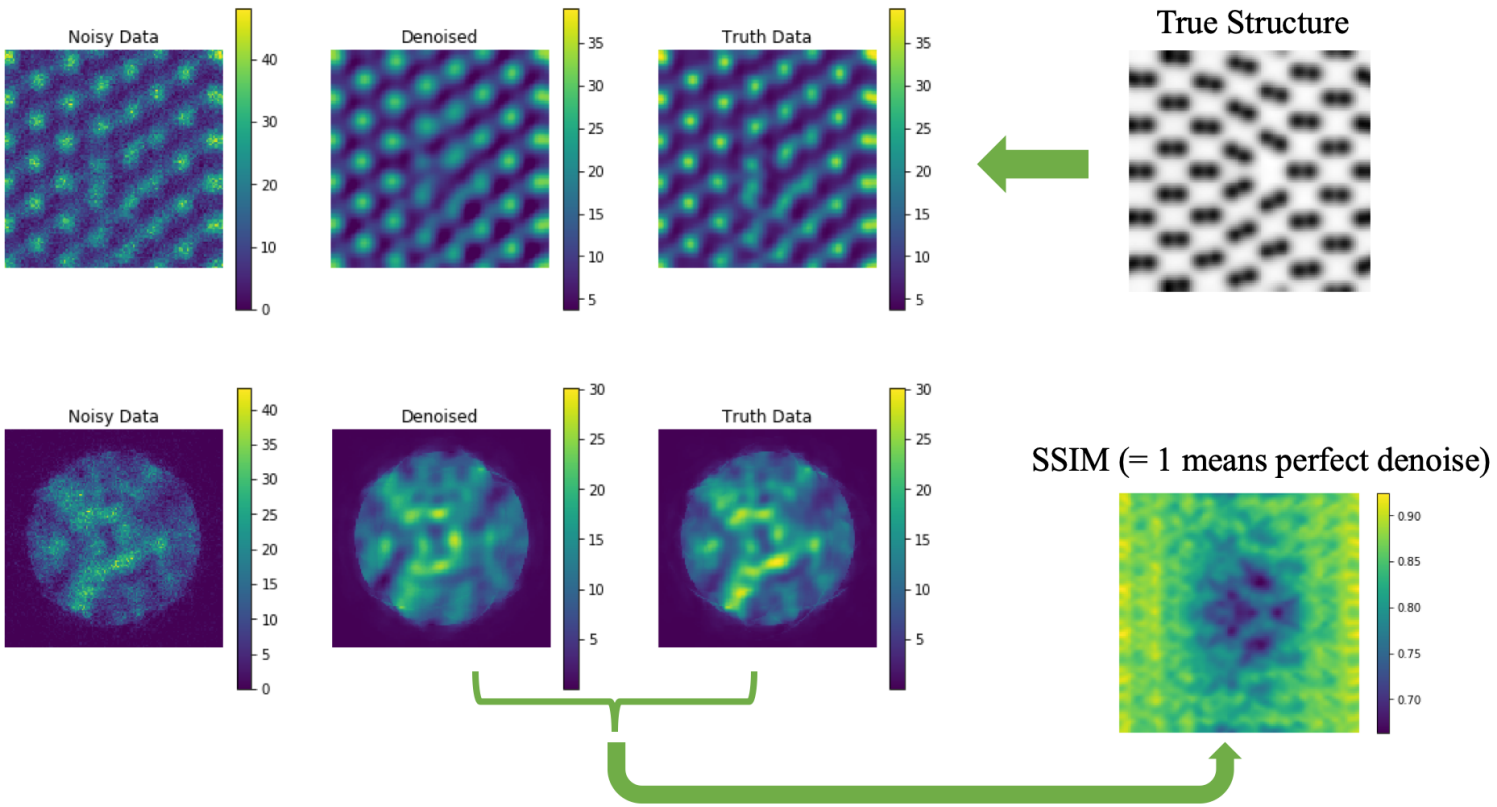
Tensor Denoising on Pure Repeating Structure

- Simulated data on strontium titanate



Tensor Denoising on Aperiodic Structure

- Simulated data on monocrystalline silicon with anomalies



Statistical Inference

- In addition to point estimation, uncertainty quantification is important.
- Principle:

$$\hat{\theta} = \theta + b + v;$$

Estimator = True parameter + Bias + Value with tractable distribution (standard error).

- Hopefully, Bias $\ll$ Standard error.
- Otherwise, debiased estimator is needed.

Debiasing in Literature

- Statistical inference in high-dimensional settings often relies on debiased estimators.
- Debiased estimators in literature:
  Low-rank matrix models
  - Matrix completion (Chen et al., 2019)

$$M^{\mathrm{d}} = \mathcal{P}_{\mathsf{rank}-r}\left[Z - \frac{1}{p}\mathcal{P}_{\Omega}(Z - M)\right].$$

  - Matrix regression (Xia, 2019a)

$$M^{\mathrm{d}} = \hat{M}^{\mathsf{nuc}} + \frac{1}{n}\sum_{i=n+1}^{2n}\left(y_i - tr(X_i^{\top}\hat{M}^{\mathsf{nuc}})\right)X_i.$$

- Sparse linear regression
  - Debiased Lasso (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014)

$$\theta^{\mathrm{d}} = \hat{\theta} + \frac{1}{n}MX^{\top}(Y - X\hat{\theta}).$$

Inference for Tucker Decomposition

The inference procedure for tensor SVD does not involve debiasing.

Assumption

- **Initialization Error**: *The estimators* $(\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)})$ *satisfy* $\max_{j=1,2,3} \|\sin\Theta(\hat{U}_j^{(0)}, U_j)\| \le C_2\sqrt{p}\sigma/\lambda_{\min}$[a] *for some absolute constant* $C_2 > 0$ *w.h.p.*

[a] $\|\sin\Theta(\hat{U}_j^{(0)}, U_j)\| = \sqrt{1 - \sigma_{r_j}^2(U_j^\top \hat{U}_j^{(0)})}$.

Attainable by HOOI under the essential SNR condition $\lambda_{\min}/\sigma \ge Cp^{3/4}$.

For better presentation, assume $r_j = O(1), \kappa = O(1)$, where

$$\lambda_{\max} := \max_j \sigma_1\big(\mathcal{M}_j(\boldsymbol{\mathcal{X}})\big), \quad \text{condition number} \quad \kappa(\boldsymbol{\mathcal{X}}) := \lambda_{\max}/\lambda_{\min}.$$

Inference for Tucker Decomposition

Theorem (Asymptotic normality of principal components)

If $\lambda_{\min}/\sigma \gg p^{3/4}$, *then*

$$\frac{\|\sin\Theta(\hat{U}_k, U_k)\|_{\mathrm{F}}^2 - p_k\sigma^2\|\Lambda_k^{-1}\|_{\mathrm{F}}^2}{\sqrt{2p_k}\sigma^2\|\Lambda_k^{-2}\|_{\mathrm{F}}} \xrightarrow{\mathrm{d.}} N(0,1) \quad \text{as} \quad p_k \to \infty.$$

Here, $\Lambda_k = \mathrm{diag}(\lambda_1^{(k)}, \ldots, \lambda_{r_k}^{(k)})$ *is the diagonal matrix containing the singular values of* $\mathcal{M}_k(\boldsymbol{\mathcal{X}})$; $\|\sin\Theta(\hat{U}_k, U_k)\|_{\mathrm{F}} = \sqrt{r_k - \sum_{i=1}^{r_k}\sigma_i^2(U_k^\top \hat{U}_k)}$.

Inference for Tucker Low-rank Tensor PCA

Theorem (Asymptotic normality of principal components in tensor PCA)

If $\lambda_{\min}/\sigma \gg p^{3/4}$, *then*

$$\frac{\|\sin\Theta(\hat{U}_k, U_k)\|_{\mathrm{F}}^2 - p_k\sigma^2\|\Lambda_k^{-1}\|_{\mathrm{F}}^2}{\sqrt{2p_k}\sigma^2\|\Lambda_k^{-2}\|_{\mathrm{F}}} \xrightarrow{\mathrm{d.}} N(0,1) \quad \text{as} \quad p_k \to \infty,$$

where $\Lambda_k = \mathrm{diag}(\lambda_1^{(k)}, \ldots, \lambda_{r_k}^{(k)})$ *is the diagonal matrix containing the singular values of* $\mathcal{M}_k(\boldsymbol{\mathcal{T}})$.

- SNR condition $\lambda_{\min}/\sigma \gg p^{3/4}$ is slighter stronger than the one for estimation ($\lambda_{\min}/\sigma \geq Cp^{3/4}$)
- To make inference, we still need to estimate Λ_k and σ^2

Inference for Tucker Low-rank Tensor

Estimates for Λ_k and σ^2:

$$
\begin{aligned}
&\hat{\Lambda}_k : \text{diagonal matrix containing the top } r_k \text{ singular values of} \\
&\qquad \mathcal{M}_k\big(\boldsymbol{\mathcal{Y}} \times_{k+1} \hat{U}_{k+1}^\top \times_{k+2} \hat{U}_{k+2}^\top\big), \\
&\hat{\sigma} = \big\|\boldsymbol{\mathcal{Y}} - \underbrace{\boldsymbol{\mathcal{Y}} \times_1 \hat{U}_1\hat{U}_1^\top \times_2 \hat{U}_2\hat{U}_2^\top \times_3 \hat{U}_3\hat{U}_3^\top}_{\approx \boldsymbol{\mathcal{X}}}\big\|_{\mathrm{F}} / \sqrt{p_1p_2p_3}.
\end{aligned} \tag{1}
$$

Theorem (Inference for Tucker Low-rank Tensor SVD)

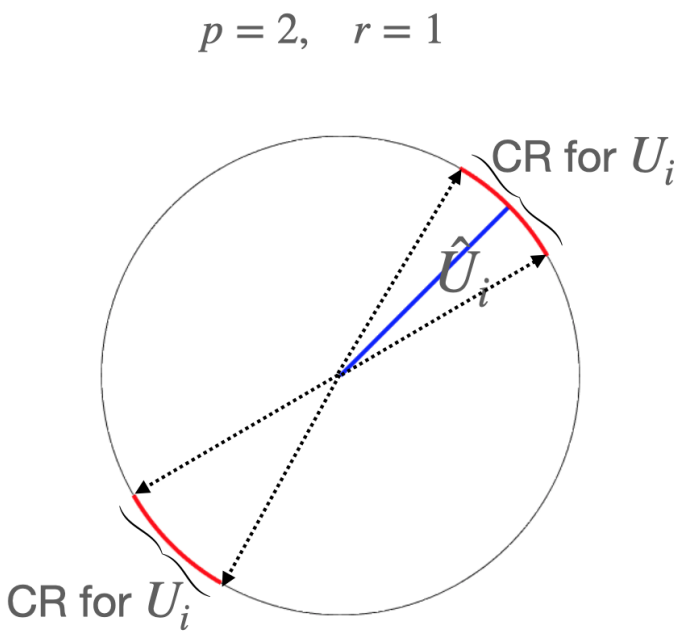
If $\lambda_{\min}/\sigma \gg p^{3/4}$, *then*

$$
\frac{\|\sin\Theta(\hat{U}_k, U_k)\|_{\mathrm{F}}^2 - p_k\hat{\sigma}^2\|\hat{\Lambda}_k^{-1}\|_{\mathrm{F}}^2}{\sqrt{2p_k}\hat{\sigma}^2\|\hat{\Lambda}_k^{-2}\|_{\mathrm{F}}} \xrightarrow{\text{d.}} N(0,1) \quad \text{as} \quad p_k \to \infty.
$$

Inference for Tucker Low-rank Tensor

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

$$\mathrm{CR}_\alpha(\hat{U}_k) := \left\{ V \in \mathbb{O}_{p_1,r_1} : \|\sin\Theta(\hat{U}_k, V)\|_{\mathrm{F}}^2 \le p_k\hat{\sigma}^2\|\hat{\Lambda}_k^{-1}\|_{\mathrm{F}}^2 + z_\alpha\sqrt{2p_k}\hat{\sigma}^2\|\hat{\Lambda}_k^{-2}\|_{\mathrm{F}} \right\}.$$



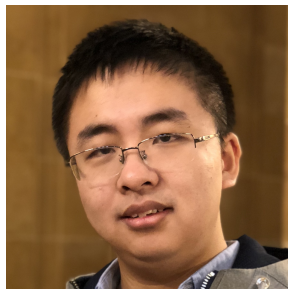
Theorem (Confidence region for tensor SVD)

If $\lambda_{\min}/\sigma \gg p^{3/4}$, *then* $\lim_{p\to\infty} \mathbb{P}\big(U_k \in \mathrm{CR}_\alpha(\hat{U}_k)\big) = 1-\alpha.$

"Blessing of Computational Barrier"

- When $\lambda_{\min}/\sigma \ll p^{3/4}$, even if good estimate is obtained, debiasing would be required.
- "Blessing of Computational Barrier"
  "Due to the statistical-computational-gap in low-rank tensor estimation, one usually requires stronger conditions than the information-theoretic limit to ensure the computationally feasible estimation is achievable. Surprisingly, such conditions 'incidentally' render a feasible low-rank tensor inference without debiasing."

Xia, D., Zhang, A. R., & Zhou, Y. (2022). Inference for low-rank tensors–no need to debias. The Annals of Statistics, 50(2), 1220-1245.

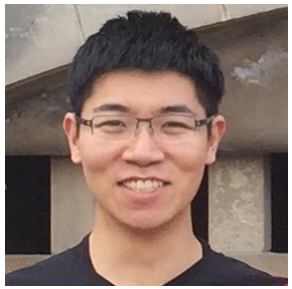



Computational Barriers Meets Over-parameterization

★ "blessing of computational barriers" in tensor regression:

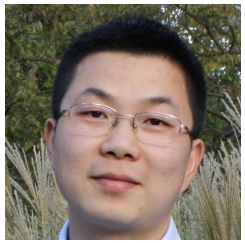
- A "prepaid lunch" exists in over-parameterized tensor regression when $d \geq 3$, but not in over-parameterized matrix regression.

Luo, Y., & Zhang, A. R. (2024). Tensor-on-tensor regression: Riemannian optimization, over-parameterization, statistical-computational gap and their interplay. *The Annals of Statistics*, 52(6), 2583-2612.



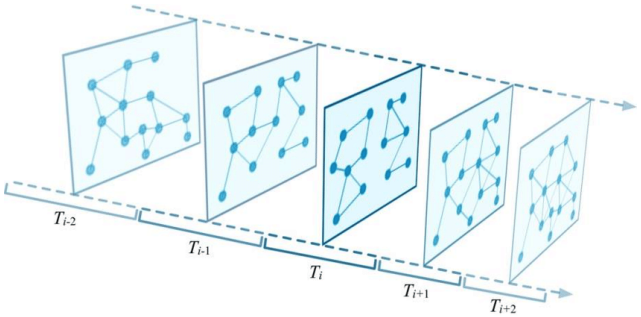
Extension: Statistical-Computational Limits in Multilayer Networks

Joint work with Jing Lei and Zihan Zhu




Statistical-Computational Limits in Multilayer Networks

A canonical multilayer stochastic block model



- n nodes, T snapshots, 2 communities
- Connectivity probability matrices:

$$B^{(1)} = \begin{bmatrix} (3/2)\rho & (1/2)\rho \\ (1/2)\rho & (3/2)\rho \end{bmatrix} \quad B^{(2)} = \begin{bmatrix} (1/2)\rho & (3/2)\rho \\ (3/2)\rho & (1/2)\rho \end{bmatrix}$$

- $\delta \sim \text{Uniform}(\{1,2\})$, $\tau \sim \text{Uniform}(\{1,2\})$
- Under H_1, for $1 \le i < j \le n, 1 \le t \le T$,

$$A_t(i,j) \overset{\text{indep.}}{\sim} Bernoulli(B^{(\tau_t)}(\delta_i, \delta_j))$$

- Under H_0,

$$A_t(i,j) \overset{\text{indep.}}{\sim} Bernoulli(\rho)$$

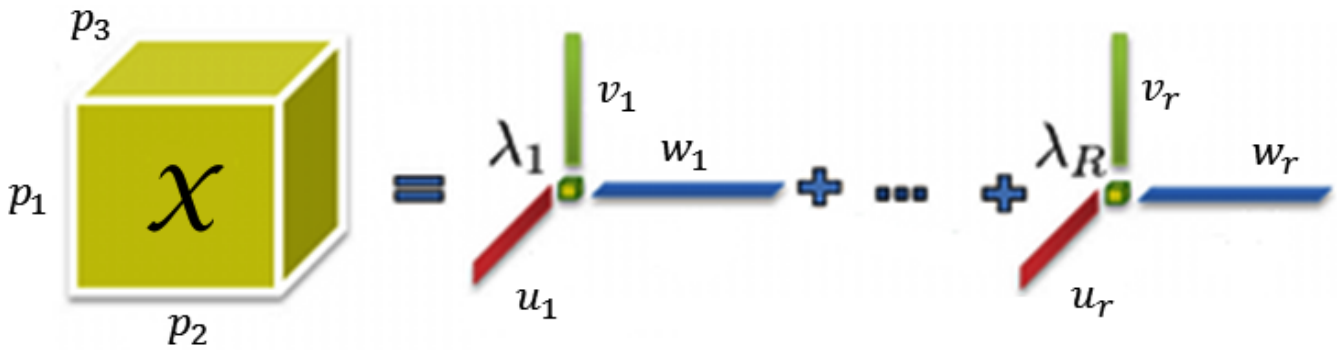
Statistical-Computational Limits in Multilayer Networks

Computational-statistical limits: $T_n = \Theta(n^a)$, $\rho_n = \Theta(n^{-b})$, $a > 0$, $b \in (0, 2)$

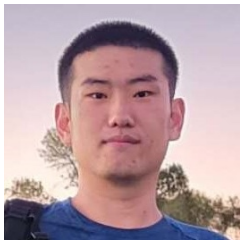
- If $b < 1 + a/2$, detection/recovery possible in polynomial time
- If $b > 1 + a$, detection/recovery impossible even without computational constraints
- If $1 + a/2 < b < 1 + a$, detection/recovery possible without computational constraints
  Assuming low-degree polynomial conjecture, detection/recovery impossible in polynomial time

Lei, J., Zhang, A. R., & Zhu, Z. (2024). Computational and statistical thresholds in multi-layer stochastic block models. *The Annals of Statistics*, 52(5), 2431-2455.

Part II: Tensor CP Decomposition: Statistical Optimality and Fast Convergence



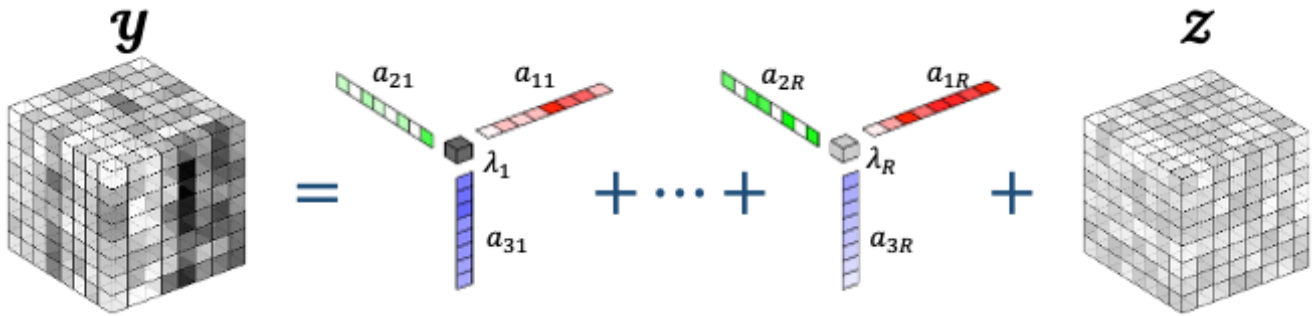
Joint work with Runshi Tang, Julien Chhor, and Olga Klopp



Statistical Model for CP Decomposition

$$\mathcal{Y} = \mathcal{X} + \mathcal{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$$

- $\mathcal{X} = \sum_{r=1}^{R} a_{1,r} \circ \cdots \circ a_{d,r}$: low-rank signal with CP rank R
- Outer product $\circ$ across loading vectors $a_{k,r} \in \mathbb{R}^{p_k}$
- $\mathcal{Z} \overset{iid}{\sim} N(0,1)$: random noise



Goal: Estimate the set of loading vectors $\{a_{k,r}\}$

Why CP Decomposition?

- Most classic tensor decomposition method, dates back to 1927
- Under mild conditions, CP decomposition is unique (unlike matrix factorizations or Tucker decomposition)
- Reveals underlying structure in data (e.g., hidden patterns or clusters)
- Broad applications:
  - psychometry
  - signal processing
  - neuroscience
  - chemometrics
  - sequecing data analysis
  - analysis for moment tensor
  - density estimation & mixture models
  - ...

Alternating Least Squares (ALS)

- ALS is standard method for CP decomposition.
- Iterate for $t = 0, 1, \ldots, k = 1, \ldots, d$,

$$\begin{aligned}\hat{B}_k^{(t+1)} &= \underset{b_{k,r}, r\in[R]}{\operatorname{argmin}} \left\| \boldsymbol{\mathcal{Y}} - \sum_{r=1}^{R} \hat{a}_{1,r}^{(t)} \circ \cdots \circ b_{k,r} \circ \cdots \circ \hat{a}_{d,r}^{(t)} \right\|_F^2 \\ &= \underset{B\in\mathbb{R}^{p_k\times R}}{\operatorname{argmin}} \left\| \mathcal{M}_k(\boldsymbol{\mathcal{Y}}) - B[(\odot_{i\neq k}\hat{A}_i^{(t)})^\top] \right\|_F^2 \\ &= \mathcal{M}_k(\boldsymbol{\mathcal{Y}})[(\odot_{i\neq k}\hat{A}_i^{(t)})^\top]^\dagger.\end{aligned}$$

  $\odot$: Khatri-Rao product. After that, normalize each column of $\hat{B}_k^{(t+1)}$ to unit norm to obtain $\hat{A}_k^{(t+1)}$.
- Objective decreases monotonically, but:
  - Non-convex loss surface
  - Convergence is not guaranteed in general

Open Question 1: Statistical Guarantees

Does ALS have statistical guarantees?

- Previous theory focuses on:
  - $R = 1$: Rank-One ALS/Tensor power iteration (Richard and Montanari (2014); Huang et al. (2022); Wu and Zhou (2024))
  - Orthogonal tensors with Sequential Rank-One ALS (Anandkumar et al. (2014); Huang et al. (2022))

- **General CP decomposition lacks theory.** Some reasons:
  - Non-orthogonality of $[a_{k,1} \cdots a_{k,R}]$
  - Lack of best rank-k approximation property (no Eckart-Young-Mirsky Theorem)

Open Question 2: Initialization of ALS

How can we effectively initialize ALS to ensure it converges to the desired estimator?

- Initialization is critical due to exponentially many local minima (Ben Arous, Mei, Montanari, Nica, 2019).
- Classical methods (e.g., Simultaneous Diagonalization) work in the noiseless case, but highly sensitive to noise in practice.
- **No robust, statistically consistent initializer known for general** R.

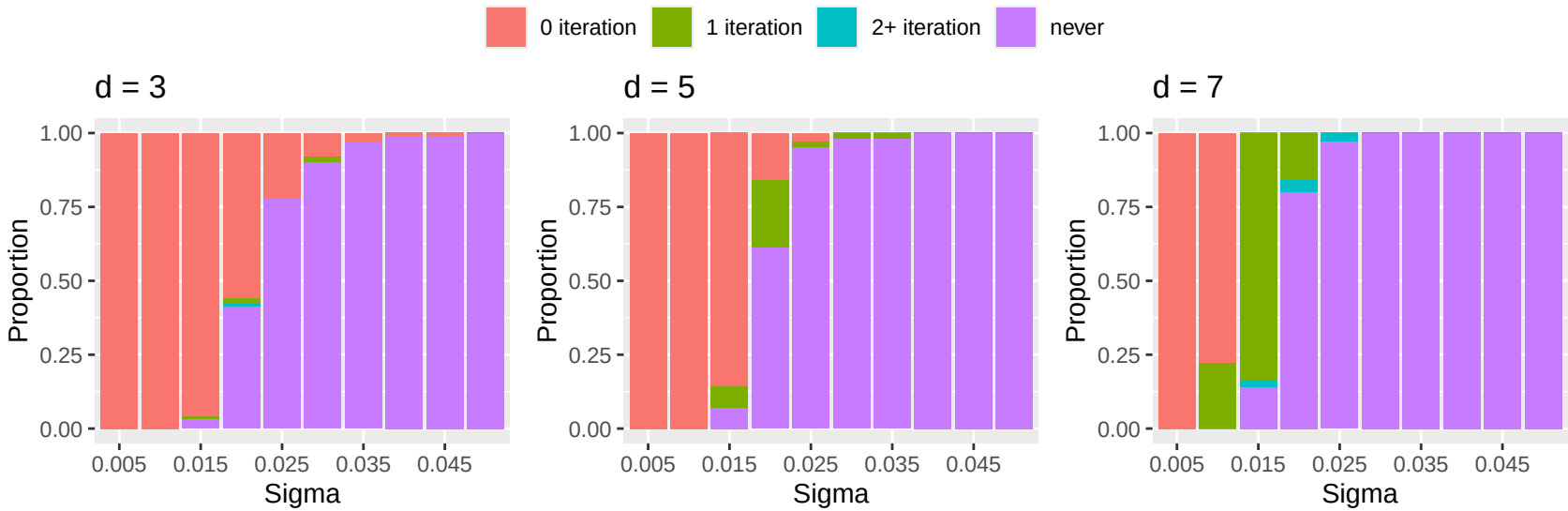
Open Question 3: Convergence Speed of ALS

How does ALS converge for CP decomposition?

- Very fast convergence of ALS was empirically observed.
- Some theoretical results for $R = 1$:
  - To optimal error in $\log p$ iterations (Huang et al. (2022))
  - To any pre-specified error in $\log \log p$ iterations (Wu and Zhou, 2024)
- General R: remains largely unexplored.

Simulation: $R = 1$

- It was observed empirically that a few iterations lead to convergence.
- Consider simulations:
  - $p_k = 5$, $d \in \{3, 5, 7\}$, $R = 1$, $\sigma \in [0.005, 0.05]$
  - Count number of iterations required for $\varepsilon_t < 0.05$ over 100 simulations

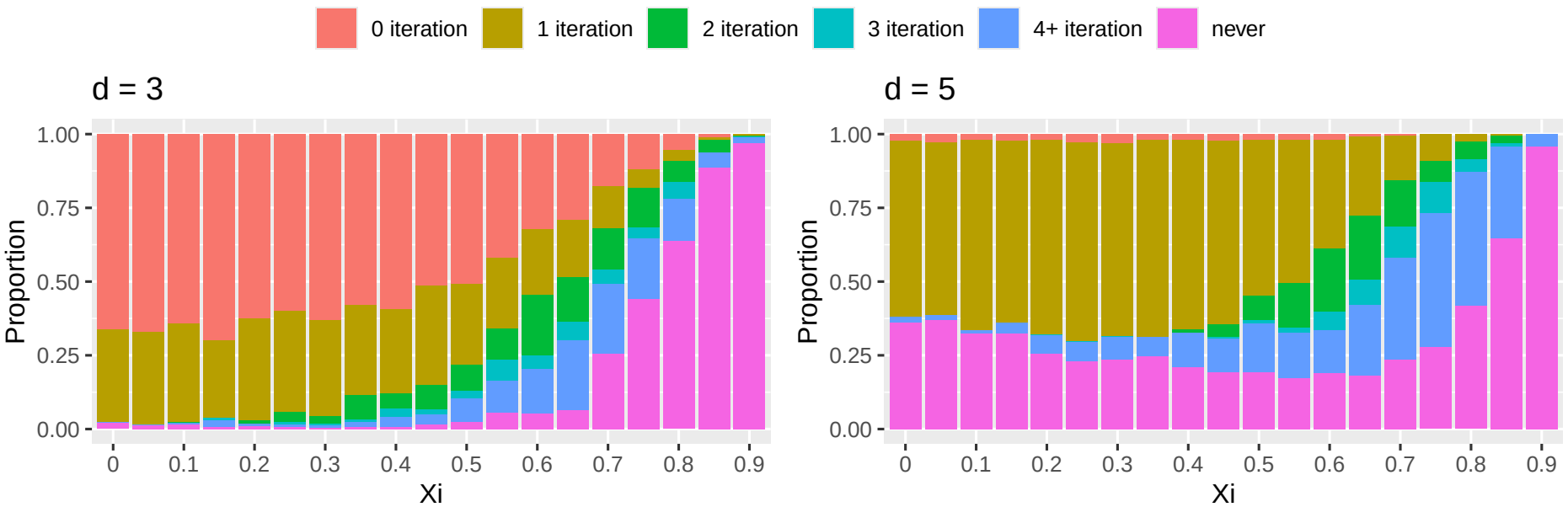


- Either 1 or 2 iterations (small error), or never converge (big error). ("All or nothing" phenomenon)

Simulation: $R > 1$

Consider simulation:

- Dimension $p_k = 5$, order $d \in \{3, 5\}$, rank $R = 3$;
- Incoherence $\xi \in [0, 0.9]$
- Count number of iterations required for $\varepsilon_t < 0.05$ over 300 simulations



- Higher coherence coefficient ξ requires more iterations.
- Never converge for large ξ.

tl;dr (too long; didn't read)

1. **Does ALS have statistical guarantees?**
   A: Yes, we will prove this.

2. **How can we effectively initialize ALS to ensure it converges to the desired estimator?**
   A: We introduce a method named TASD (Tucker-based Approximation with Simultaneous Diagonalization) that achieves good empirical performance and has provable guarantees.

3. **How does ALS converge for CP decomposition?**
   A: Depending on the SNR, incoherence, and other parameters,
   - When $R = 1$: Either 1, 2 iterations (small error) converges to statistically-optimal estimator, or never converge (big error).
   - When $R > 1$: the convergence varies (small coherence: quadratic; large coherence: linear rate).

A Simplified Case: Rank-One Tensor Decomposition

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{Z}}, \quad \boldsymbol{\mathcal{X}} = \lambda a_1 \circ \cdots \circ a_d$$

- ALS update reduces to power iteration:

$$\hat{a}_k^{(t+1)} = \frac{\boldsymbol{\mathcal{Y}} \times_{i \neq k} \hat{a}_i^{(t)\top}}{\|\boldsymbol{\mathcal{Y}} \times_{i \neq k} \hat{a}_i^{(t)\top}\|_2}$$

- Initialization via unfolding:

$$\hat{a}_k^{(0)} = \mathrm{SVD}_1(\mathcal{M}_k(\boldsymbol{\mathcal{Y}}))$$

- Loss function that accounts for sign ambiguity:

$$\varepsilon_t = \max_{i \in [d]} \min\{\|\hat{a}_i^{(t)} - a_i\|, \|\hat{a}_i^{(t)} + a_i\|\}$$

Rank-One Theory: Initialization and ALS Accuracy

Theorem (Informal: Rank-one tensor decomposition)

When signal-to-noise ratio $\lambda \gtrsim \sigma p_{\max}^{d/4}$, *with high probability,*

- *Initialization error:*

$$\varepsilon_0 \lesssim \sigma \left(\frac{1}{\lambda^2} \vee \frac{p_{\max}^d}{\lambda^4} \right).$$

- *After 2 iterations of ALS/power iteration,*

$$\varepsilon_t \leq C\sigma\sqrt{p_{\max}}/|\lambda|.$$

- SNR requirement $\lambda \gtrsim \sigma p_{\max}^{d/4}$ is essential; if it fails, the problem is computationally infeasible (Zhang and Xia, 2018).
- Error rate $\sigma\sqrt{p_{\max}}/|\lambda|$ is information-theoretic optimal, supported by minimax lower bound.
- Key technical tool: new perturbation bound that works for $R = 1$ only.

Summary of ALS Guarantees ($R = 1$)

	Initialization	Required λ	Iterations	Error
Richard and Montanari (2014)	Unfolding	$p^{\lfloor d/2 \rfloor/2}$	NA	$p^{1/2}/\lvert\lambda\rvert$
	Random	$p^{d/2}$	NA	
Huang et al. (2022)		$p^{d/2+\varepsilon}$	$\log_d p$	
Wu and Zhou (2024)		$p^{d/2}(\log p)^{-C}$	$\log_d \log_d p$	$\forall \delta > 0$
Our result	Unfolding	$p^{d/4}$	2	$p^{1/2}/\lvert\lambda\rvert$

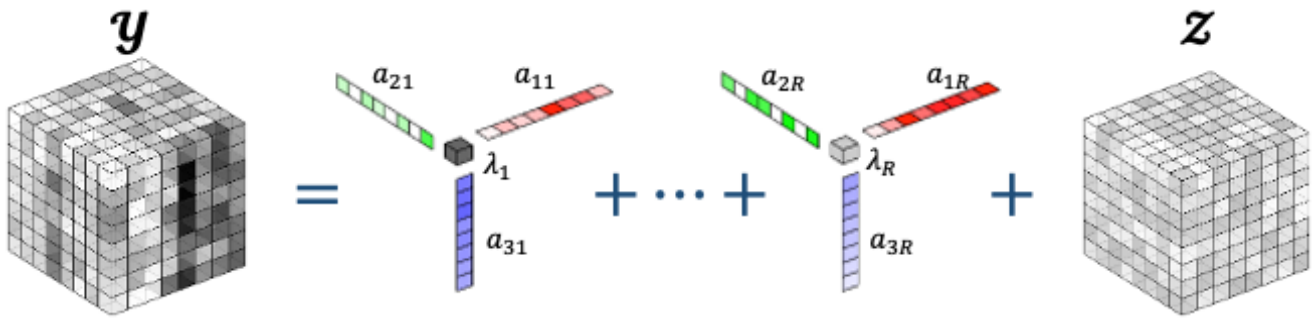
Table: Results for CP decomposition in $R = 1$

CP Tensor Decomposition Model: General Rank

$$\mathcal{Y} = \mathcal{X} + \mathcal{Z}, \quad \mathcal{X} = \sum_{r=1}^{R} \lambda_r \, a_{1,r} \circ \cdots \circ a_{d,r}$$

- $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$: observed tensor
- $\mathcal{X}$: low-rank tensor (CP rank R)
- $\mathcal{Z}$: noise, i.i.d. sub-Gaussian with variance σ^2

Goal: estimate loading matrices $A_k = [a_{k,1}, \ldots, a_{k,R}]$



ALS for General Rank $R > 1$

- ALS: iterate for $t = 0, 1, \ldots, k = 1, \ldots, d$,

$$\begin{aligned}\hat{B}_k^{(t+1)} &= \underset{B \in \mathbb{R}^{p_k \times R}}{\operatorname{argmin}} \left\| \mathcal{M}_k(\boldsymbol{\mathcal{Y}}) - B[(\odot_{i \neq k} \hat{A}_i^{(t)})^\top] \right\|_F \\ &= \mathcal{M}_k(\boldsymbol{\mathcal{Y}})[(\odot_{i \neq k} \hat{A}_i^{(t)})^\top]^\dagger.\end{aligned}$$

  After that, normalize each column of $\hat{B}_k^{(t+1)}$ to unit norm to obtain $\hat{A}_k^{(t+1)}$

- Error metric:

$$\varepsilon_t = \max_{k \in [d]} \left\{ \max_{r \in [R]} \left\{ \min\{ \| \hat{a}_{k,r}^{(t)} \pm a_{k,r} \| \} \right\} \right\}$$

ALS Convergence Guarantee ($R > 1$)

Denote

- $\xi = \max_{k\in[d]} \max_{i\neq j\in R} |a_{i,k}^\top a_{j,k}|$
- $\lambda_{\max} = \max_{r\in[R]}\{|\lambda_r|\}$, and $\lambda_{\min} = \min_{r\in[R]}\{|\lambda_r|\}$

Assumptions:

- Initialization: $\varepsilon_0 \frac{\lambda_{\max}}{\lambda_{\min}} dR \lesssim C_1^{-1}$
- Incoherence: if $d > 3$, $\xi \leq \left(C_1 d R^{\frac{1}{d-3}}\right)^{-1}$; if $d = 3$, $\xi \leq (C_1 R^{1/2})^{-1}$
- Signal strength:
  $\lambda_{\min} \geq C_1\sigma\left(d^2 \log d\sqrt{p_{\max}} \vee \frac{\lambda_{\max}}{\lambda_{\min}}\sqrt{dp_{\max}\log d}(d+R)\right)$

Coherence and Two-Phase Convergence

Theorem (Informal: Convergence of ALS under General Rank)

- *ε_t converges to the minimax error bound $\sigma\sqrt{p_{\max}}/\lambda_{\min}$*
- *Convergence speed relies on coherence ξ:*
  - *When $(R-1)\xi^{d-1} \leq \sigma\sqrt{p_{\max}}/\lambda_{\min}$, ALS converges quadratically.*
  - *When $(R-1)\xi^{d-1} > \sigma\sqrt{p_{\max}}/\lambda_{\min}$ and $\xi^{d-2} \leq c\frac{\lambda_{\min}}{Rd\lambda_{\max}}$, ALS converges linearly.*

- Coherence among loadings hinder convergence
- Intuition: High coherence $\rightarrow$ singular design matrix in least squares $\rightarrow$ slower convergence
- ALS requires warm initialization

Question: How to initialize?

Simultaneous Diagonalization (SimDiag)

SimDiag can exactly recover loadings in noiseless settings:

- $\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \lambda_r a_{1,r} \circ \cdots \circ a_{d,r}$. Goal: recover $a_{k,r}$.
- Draw two sets of vectors

$$w_{1k}, w_{2k} \in \mathbb{R}^{p_k}, k = 3, \ldots, d.$$

- Compute two matrices

$$M_1 := \boldsymbol{\mathcal{X}} \times_{k=3}^{d} w_{1k}^\top = \sum_{r=1}^{R} \left(\lambda_r \prod_{k=3}^{d} \langle a_{k,r}, w_{1k} \rangle \right) \cdot a_{1,r} a_{2,r}^\top := A_1 D_1 A_2^\top$$

$$M_2 := \boldsymbol{\mathcal{X}} \times_{k=3}^{d} w_{2k}^\top = \sum_{r=1}^{R} \left(\lambda_r \prod_{k=3}^{d} \langle a_{k,r}, w_{2k} \rangle \right) \cdot a_{1,r} a_{2,r}^\top := A_1 D_2 A_2^\top$$

- $M_1 M_2^\dagger = A_1 D_1 A_2^\top (A_1 D_2 A_2^\top)^\dagger = A_1 D_1 D_2^\dagger A_1^\dagger$
  $\rightarrow$ **We can recover columns of A_1 from eigenvectors of $M_1 M_2^\dagger$.**

Limitations of SimDiag

SimDiag

- is most effective in noiseless cases
- is sensitive to noise and input dimension p_k
  - Under noisy settings,

$$M_1 = \sum_{r=1}^{R} \left(\lambda_r \prod_{k=3}^{d} \langle a_{k,r}, w_{1k} \rangle \right) \cdot a_{1,r} a_{2,r}^\top + Z_1;$$

$$M_2 = \sum_{r=1}^{R} \left(\lambda_r \prod_{k=3}^{d} \langle a_{k,r}, w_{2k} \rangle \right) \cdot a_{1,r} a_{2,r}^\top + Z_2;$$

  - Coefficients in the signal parts, $\left(\lambda_r \prod_{k=3}^{d} \langle a_{k,r}, w_{1k} \rangle \right)$, shrink when p_k and d grow.
- lacks robustness under realistic settings

TASD: Tucker-based Approximation + SimDiag

Instead, we propose TASD:

- Perform Tucker decomposition: $\mathcal{Y} \approx \mathcal{S} \times_{k=1}^{d} U_k$
- SimDiag applied to $\mathcal{S}$ (smaller tensor of size $\mathbb{R}^{R \times \cdots \times R}$) instead of orginal tensor $\mathcal{X}$ (bigger tensor of size $\mathbb{R}^{p_1 \times \cdots \times p_d}$)
- Estimate A_k via U_k and the outputs of SimDiag

Algorithm Summary of TASD Part I.

Input: Order-d Tensor $\mathcal{Y}$ and Target CP Rank R

1. Obtain $\hat{U}_k$ and $\hat{\mathcal{S}}$ from Tucker decomposition of $\mathcal{Y}$ with Tucker rank $(R, \ldots, R)$;
2. Generate w_{1k} and w_{2k} in $\mathbb{R}^{p_k}$ for $k \in [d]$;
3. Calculate $\hat{M}_i = \mathcal{M}_h(\hat{\mathcal{S}} \times_{k \notin \{h, h+1\}} w_{ik})$ for $i \in \{1, 2\}$;
4. Obtain eigenvectors $\hat{V}_h$ from the eigen-decomposition of $\hat{M}_1 (\hat{M}_2)^{\dagger}$;
5. Calculate the estimation of hth loading matrices $\hat{A}_h = \hat{U}_h \operatorname{Re}(\hat{V}_h)$

TASD Algorithm: Part II

- After obtain $\hat{A}_h$, we solve

$$\hat{D} = \underset{D \in \mathbb{R}^{R \times \prod_{k \neq h} p_k}}{\operatorname{argmin}} \left\| \mathcal{M}_h(\boldsymbol{\mathcal{Y}}) - \hat{A}_h D \right\|_F,$$

- Motivation: $\mathcal{M}_h(\boldsymbol{\mathcal{X}}) = A_h \operatorname{diag}(\lambda)(\odot_{k \neq h} A_k)^\top$
  $\rightarrow$ each row of $\hat{D}$ approximates a row of $(\odot_{k \neq h} A_k)^\top$
- Then perform Rank-One-ALS on each row of $\hat{D}$

Algorithm Summary of TASD Part II.

1. Calculate $\hat{D}$ by $\hat{D} = (\hat{A}_h)^\dagger \mathcal{M}_h(\boldsymbol{\mathcal{Y}})$;
2. Let $\boldsymbol{\mathcal{Y}}_r \in \mathbb{R}^{\prod_{k \neq h} p_k}$ be the rth row of $\hat{D}$ for $r \in [R]$;
3. For $r \in [R]$ and $k \neq h \in [d]$, estimate $\hat{a}_{k,r}$ by R1-ALS with input $\boldsymbol{\mathcal{Y}}_r$;
4. Let $\hat{A}_k = [\hat{a}_{k,1}, \cdots, \hat{a}_{k,R}]$.

Advantages of TASD

- Exact recovery when noise $\mathcal{Z} = 0$ and A_k's have full column rank.
- TASD applied on R^d tensor instead of $p_1 \times \cdots \times p_d$
  Since, $R \ll p_k$ in most practical settings, this improves numerical stability and robustness.

Theoretical Guarantee for TASD

- $\lambda_{\text{Tucker}} = \min_{k\in[d]} \sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$, $\bar{\lambda}_{\text{Tucker}} = \max_{k\in[d]} \sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$ and $\kappa_T = \frac{\bar{\lambda}_{\text{Tucker}}}{\lambda_{\text{Tucker}}}$
- $\hat{A}_h = [\hat{a}_{1,h}, \ldots, \hat{a}_{R,h}]$: the output of TASD
- w_{ik} generated with i.i.d. standard normal entries
- Tucker decomposition performed by HOOI initialized by HOSVD.

Assumptions

- Loading matrices A_k's have full column rank
- $p_{\min} \geq C \log d$,
- $\lambda_{\text{Tucker}} \geq C\sigma p_{\max}^{d/4} \vee \sigma C^d R^{\frac{d-1}{2}}$,
- $\lambda_{\min} \gtrsim \sigma\kappa_T \frac{\lambda_{\max}}{\lambda_{\min}} C^d d^{3d-5} R^{\frac{5d}{2}-4} (\log R)^{\frac{d-2}{2}} (R \vee \log d)^{d-2} \sqrt{p_{\max} \vee R^{d-1}}$, and
- $C\xi d^3 R^{9/2} < 1$;

Theoretical Guarantee for TASD

Theorem (Informal; Theory for TASD)

For any r with probability more than 0.99:

$$\min_{\tilde{r}\in[R]}\{\|\hat{a}_{r,h}\pm a_{\tilde{r},h}\|\}$$
$$\lesssim\sigma\frac{\sqrt{p_{\max}\vee R^{d-1}}}{\lambda_{\text{Tucker}}}+$$
$$\sigma\kappa_T\frac{\lambda_{\max}}{\lambda_{\min}}\frac{C^d d^{3d-5}R^{\frac{5d}{2}-4}(\log R)^{\frac{d-2}{2}}(R\vee\log d)^{d-2}\sqrt{p_{\max}\vee R^{d-1}}}{\lambda_{\min}}.$$

Combining theories for ALS and TASD, we have:

Corollary (Informal: theory for TASD-ALS)

Under mild conditions, TASD-ALS converges globally with:

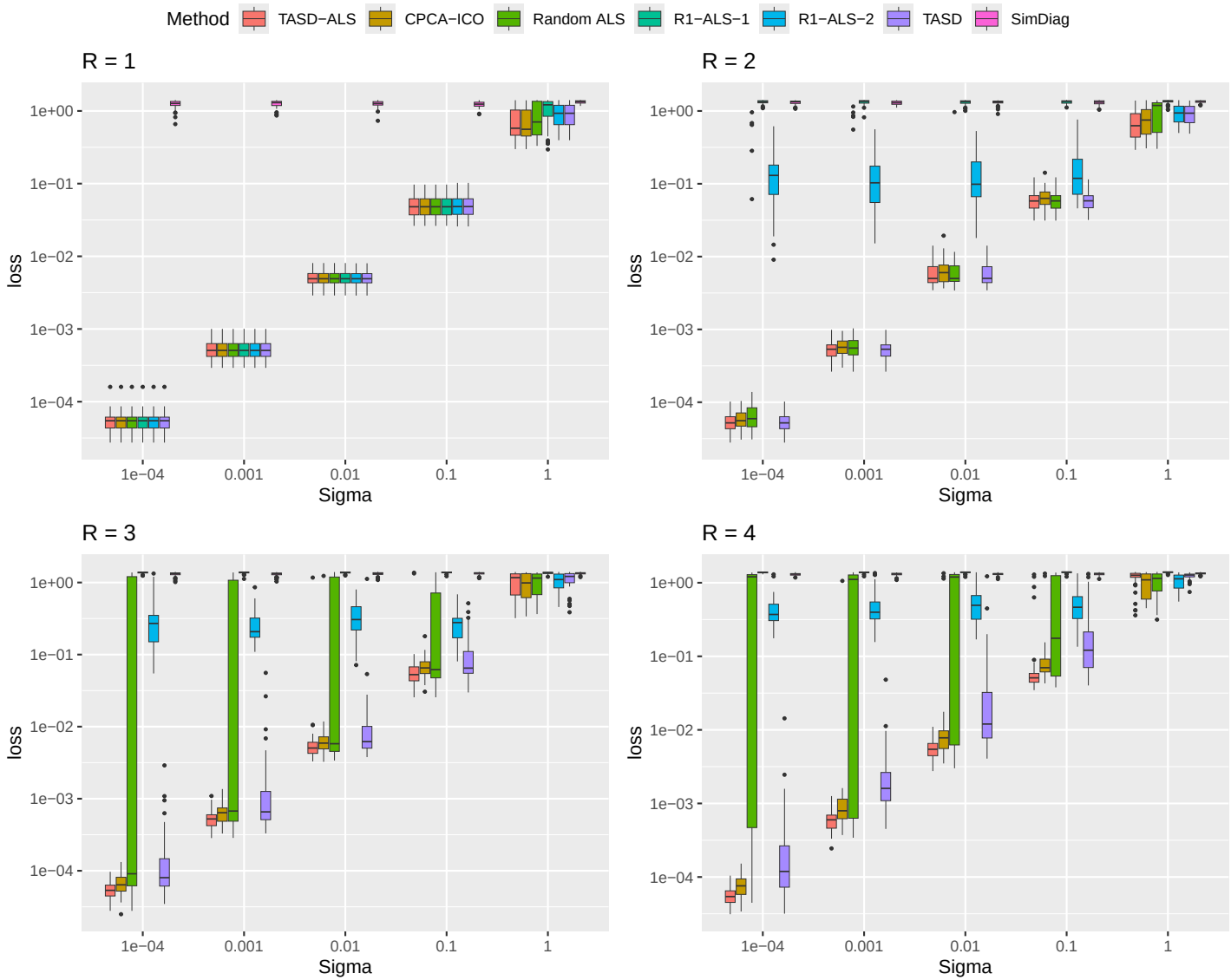
$$\varepsilon_t\lesssim\sigma\sqrt{p_{\max}}/\lambda_{\min}.$$

Summary

- Tensor methods are useful.
- Tensor Tucker decomposition:
  - Polynomial time algorithm that is statistically optimal for strong SNR
  - Analysis of computational difficulty under different SNR scenarios
  - Excellent denoising performance on 4D-STEM images
- Tensor CP decomposition:
  - TASD+ALS achieves good performance with provable guarantees.
  - For main message, see tl; dr.
- Other topics...

Thank you! Questions?

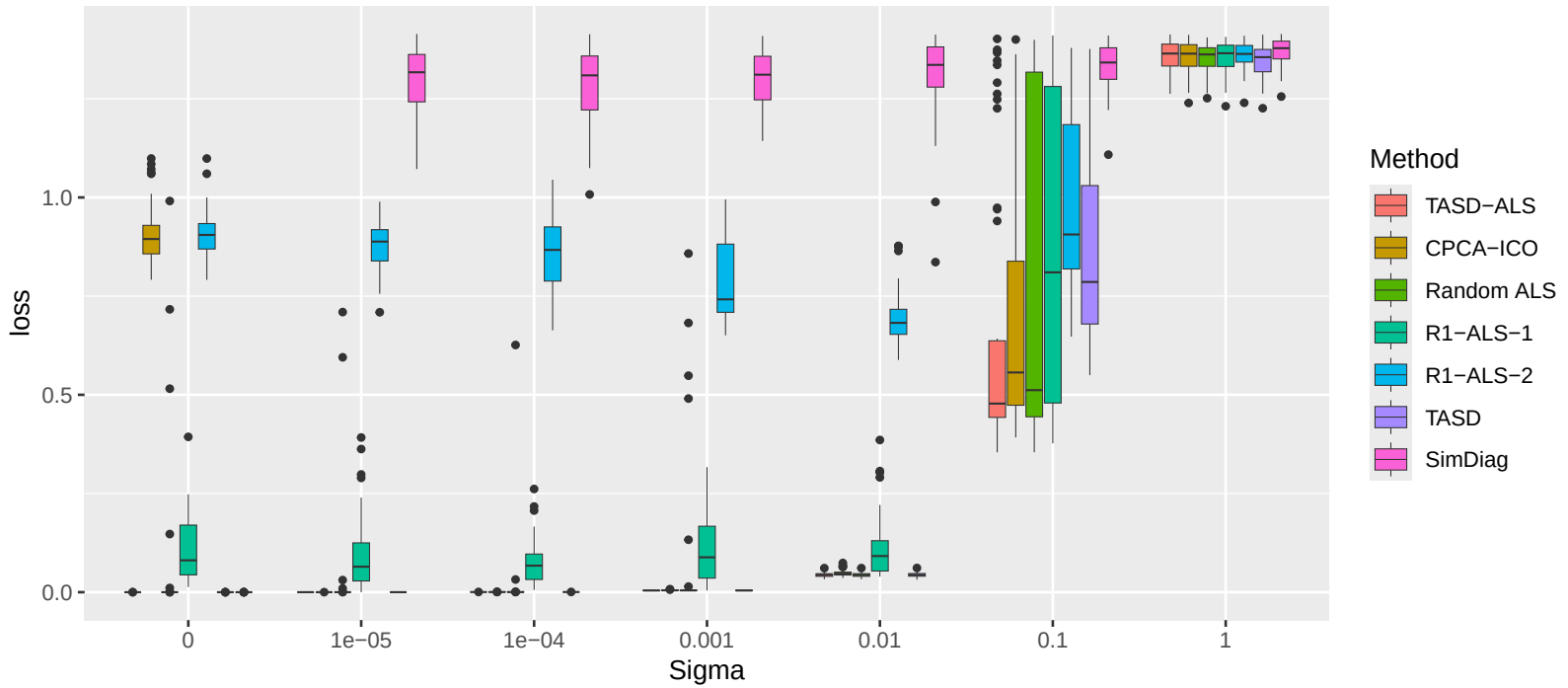
Simulation I



Recovery Performance Across Methods

- **SimDiag** fails consistently under noise.
- **R1-ALS-1/2** fails for $R > 1$ due to poor best rank-1 approximations.
- **TASD-ALS** outperforms **CPCA-ICO** and shows more stability than **Random ALS**.
- For $R = 3, 4$, **TASD-ALS** improves on **TASD**.

Simulation I: Additional Study with $\lambda_r = 1$ and $R = 2$

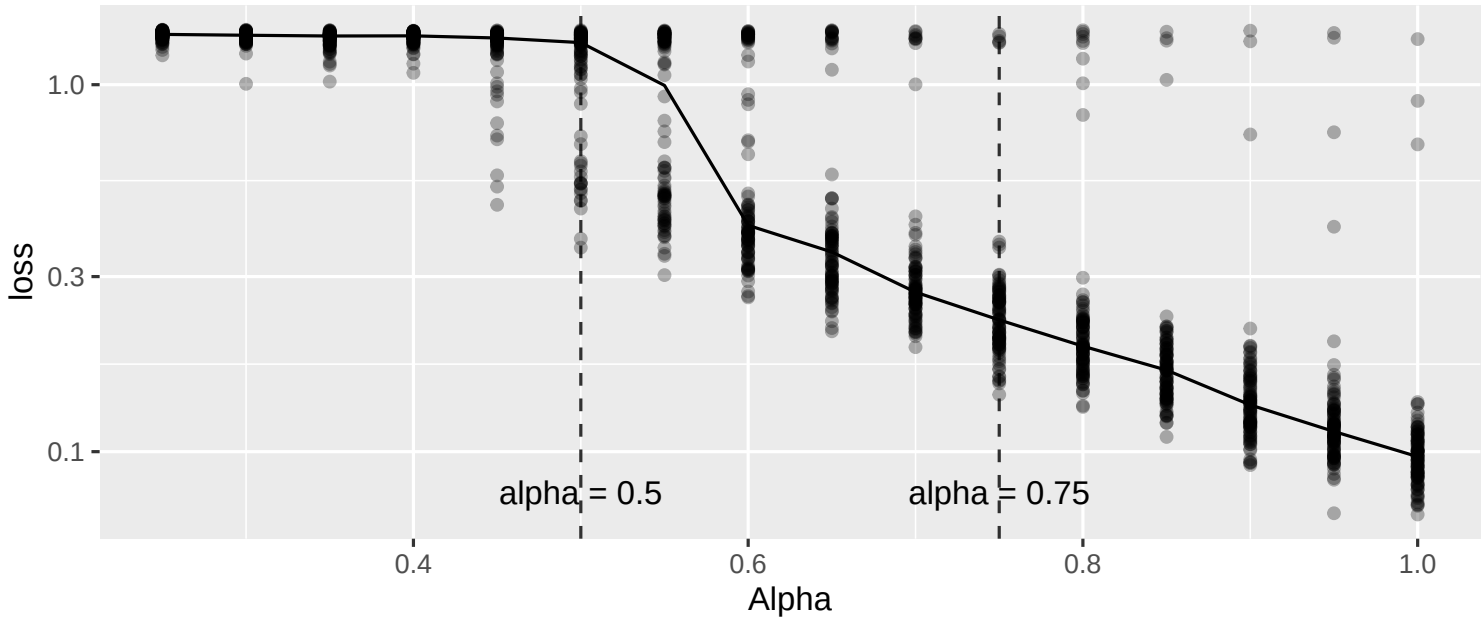


Recovery Performance Across Methods

- **SimDiag** only works for noiseless case.
- The best rank-one approximation $\approx\in$ CPD in this well-conditioned setting.
- **CPCA-ICO** doesn't work when eigengap $= 0$

Simulation II: TASD-ALS in varying SNR

- Setup: $p_k = p = 30$, $R = 3$, $\sigma = 15/p^{\alpha}$.
- Theory:
  - ALS convergence with warm initialization: $\alpha > 0.5$
  - TASD success: $\alpha > 0.75$
- Median loss decreases near-linearly for $\alpha > 0.5$ (log scale).
- Flat loss for $\alpha < 0.5$ implies failure of ALS convergence.

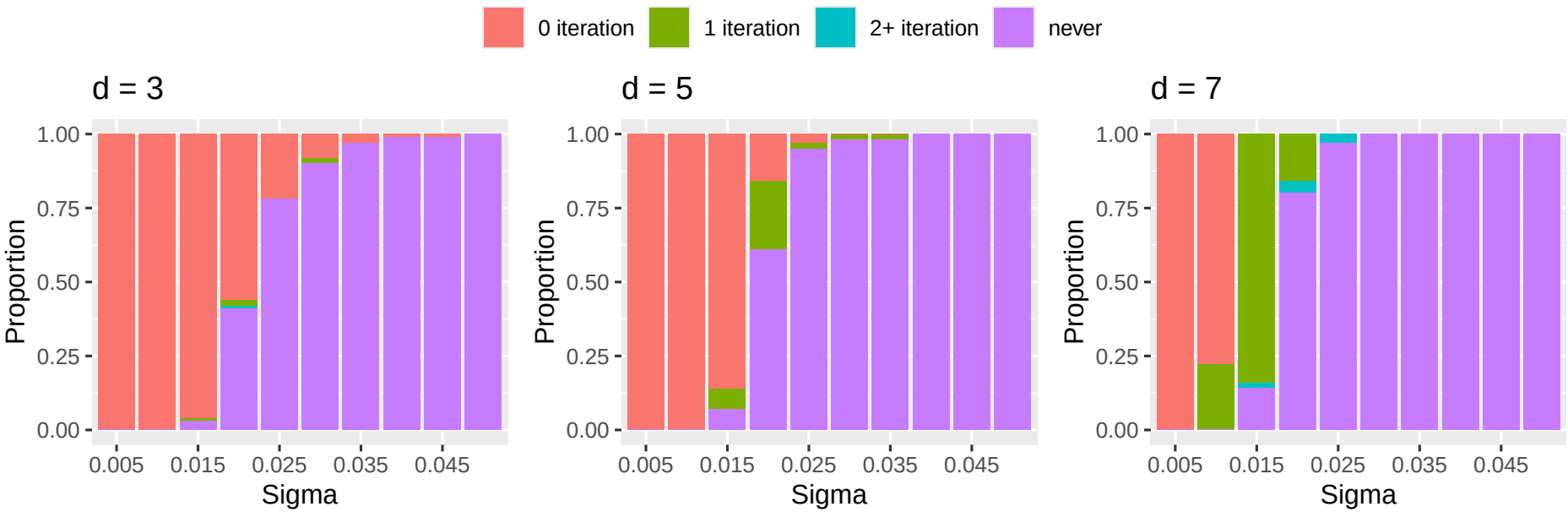


Simulation III: Tracking Convergence in High Order Setting

- $p_k = 5$; $d \in \{3, 5, 7\}$
- $R = 1$; $\sigma \in [0.005, 0.05]$
- Count number of iterations required for $\varepsilon_t < 0.05$ over 100 simulations

Results: Convergence Speed

- Higher order tensor requires stronger signal strength and more iterations
- Fail to converge once the noise level is greater than some threshold

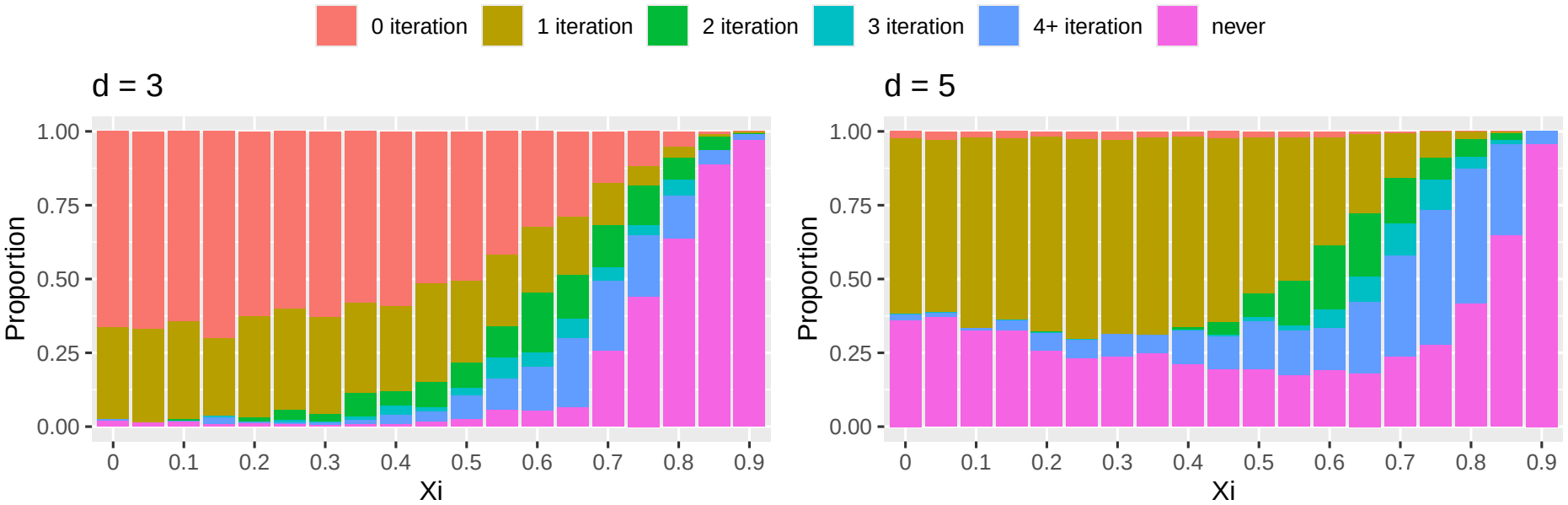


Simulation III: Tracking Convergence in High Order Setting

- $p_k = 5$; $d \in \{3, 5\}$
- $R = 3$; $\xi \in [0, 0.9]$
- Count number of iterations required for $\varepsilon_t < 0.05$ over 300 simulations

Results: Convergence Speed

- Higher coherence coefficient ξ requires more iterations
- Fail to converge for large ξ



Bibliography

Anandkumar, A., Ge, R., Hsu, D. J., Kakade, S. M., Telgarsky, M., et al. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832.

Huang, J., Huang, D. Z., Yang, Q., and Cheng, G. (2022). Power iteration for tensor pca. *Journal of Machine Learning Research*, 23(128):1–47.

Richard, E. and Montanari, A. (2014). A statistical model for tensor pca. *Advances in neural information processing systems*, 27.

Wu, Y. and Zhou, K. (2024). Sharp analysis of power iteration for tensor pca. *arXiv preprint arXiv:2401.01047*.