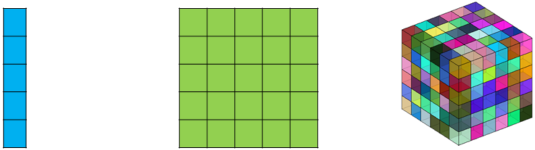
Singular Value Decomposition for High-dimensional Tensor Data

Anru Zhang
Department of Statistics
University of Wisconsin-Madison

Introduction

- Tensors are arrays with multiple directions.



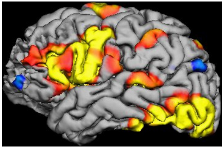
Order-1 tensor: vector   Order-2 tensor: matrix   Order-3 tensor

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

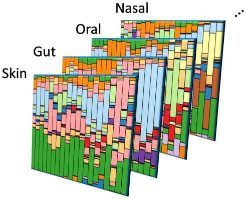
$$\mathcal{A} \in \mathbb{R}^{p_1 \times \cdots \times p_d}, \qquad \mathcal{A} = (A_{i_1 \cdots i_d}), \qquad 1 \leq i_k \leq p_k, \quad k = 1, \ldots, d.$$

More High-Order Data Are Emerging

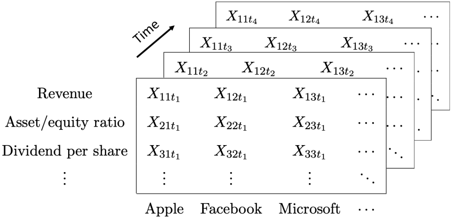
- Brain imaging



- Microbiome studies



- Matrix-valued time series



High Order Enables Solutions for Harder Problems

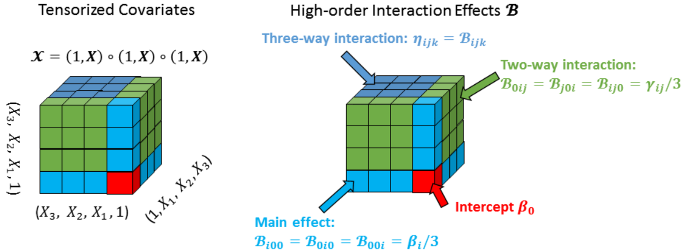
High-order Interaction Pursuits

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

$$y_i = \beta_0 + \underbrace{\sum_i X_i\beta_i}_{\text{Main effect}} + \underbrace{\sum_{i,j} \gamma_{ij} X_i X_j}_{\text{Pairwise interaction}} + \underbrace{\sum_{i,j,k} \eta_{ijk} X_i X_j X_k}_{\text{Triple-wise}} + \varepsilon_i, \quad i = 1, \ldots, n.$$

- Rewrite as

$$y_i = \langle \mathcal{B}, \mathcal{X}_i \rangle + \varepsilon_i.$$



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)
  - 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)
  - ...
- Method of Moment (MoM):
  - First moment → vector;
  - Second moment → matrix;
  - **High-order moment → high-order tensors.**

High Order is ...

- **High order is more charming!**

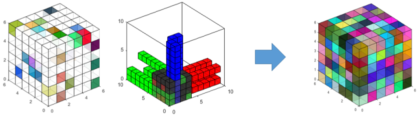
- **High order is harder!**

  Tensor problems are **far more than** extension of matrices.

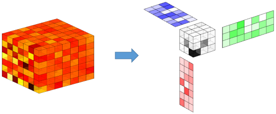
  - More structures
  - High-dimensionality
  - Computational difficulty
  - Many concepts not well defined or NP-hard

High Order Casts New Problems and Challenges

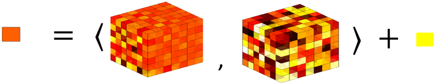
- Tensor Completion



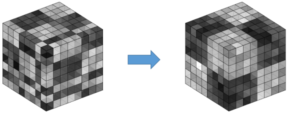
- Tensor SVD



- Tensor Regression

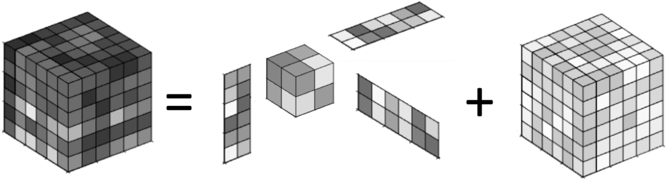


- Biclustering/Triclustering



- ...

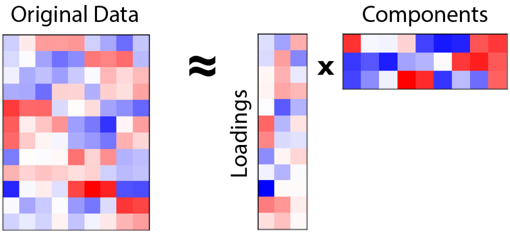
In this talk, we focus on **tensor SVD**.

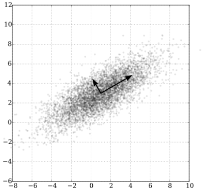


Part I: Tensor SVD: Statistical and Computational Limits

SVD and PCA

- Singular value decomposition (SVD) is one of the most important tools in multivariate analysis.
- Goal: Find the **underlying low-rank structure** from the data matrix.
- Closely related to Principal component analysis (PCA): Find the **one/multiple directions** that explain most of the **variance**.





Tensor SVD

- We propose a general framework for tensor SVD.
- $$\mathcal{Y} = \mathcal{X} + \mathcal{Z},$$

  where
  - $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is the observation;
  - $\mathcal{Z}$ is the noise of small amplitude;
  - $\mathcal{X}$ is a low-rank tensor.
- We wish to **recover** the high-dimensional **low-rank** structure $\mathcal{X}$.

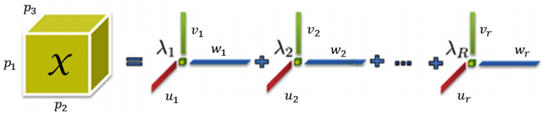
  $\rightarrow$ Unfortunately, there is no uniform definition for tensor rank.

Tensor Rank Has No Uniform Definition

- Canonical polyadic (CP) rank:

$$r_{cp} = \min r \quad \text{s.t.}$$

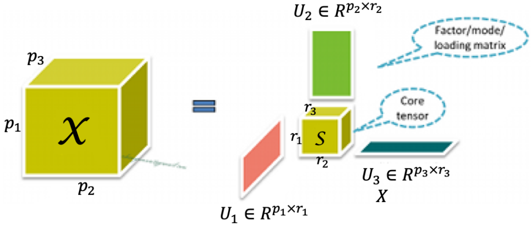
$$\mathcal{X} = \sum_{i=1}^{r} \lambda_i \cdot u_i \circ v_i \circ w_i$$



- Tucker rank:

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

$$\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \boldsymbol{U}_k \in \mathbb{R}^{p_k \times r_k}$$



Smallest possible (r_1, r_2, r_3) are Tucker rank of $\mathcal{X}$.

- See Kolda and Balder (2009) for a comprehensive survey.

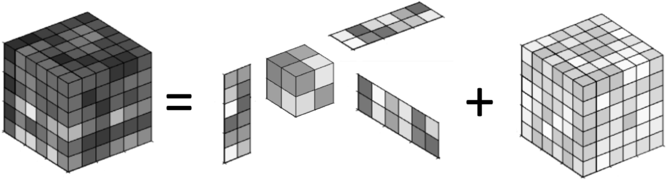
Picture Source: Guoxu Zhou's website. http://www.bsp.brain.riken.jp/ zhougx/tensor.html

Model

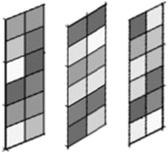
- Observations: $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$,

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

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



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

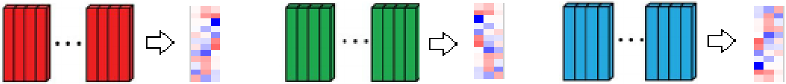


Straightforward Idea 1: Higher order SVD (HOSVD)

- Since $\boldsymbol{U}_k$ is the subspace for $\mathcal{M}_k(\boldsymbol{\mathcal{X}})$, let

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

  i.e. the leading r_k singular vectors of all mode-k fibers.



Note: $\text{SVD}_r(\cdot)$ represents the first r left singular vectors of any given matrix.

Straightforward Idea 1: Higher order SVD (HOSVD)

(De Lathauwer, De Moor, and Vandewalle, SIAM J. Matrix Anal. & Appl. 2000a)

A multilinear singular value decomposition
L De Lathauwer, B De Moor, J Vandewalle - SIAM journal on Matrix Analysis …, 2000 - SIAM
We discuss a multilinear generalization of the singular value decomposition. There is a strong analogy between several properties of the matrix and the higher-order tensor decomposition; uniqueness, link with the matrix eigenvalue decomposition, first-order
Cited by 2826 Related articles All 18 versions

- **Advantage**: easy to implement and analyze.
- **Disadvantage:** perform sub-optimally.
  Reason: simply unfolding the tensor fails to utilize the tensor structure!

Straightforward Idea 2: Maximum Likelihood Estimator

- Maximum-likelihood estimator

$$\hat{\boldsymbol{U}}_1^{mle}, \hat{\boldsymbol{U}}_2^{mle}, \hat{\boldsymbol{U}}_3^{mle}, \hat{\boldsymbol{\mathcal{S}}}^{mle} = \operatorname*{argmax}_{\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3, \boldsymbol{\mathcal{S}}} \|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{S}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3\|_F^2$$

- Equivalently, $\hat{\boldsymbol{U}}_1^{mle}, \hat{\boldsymbol{U}}_2^{mle}, \hat{\boldsymbol{U}}_3^{mle}$ can be calculated via

$$\begin{aligned} \max \quad & \left\|\boldsymbol{\mathcal{Y}} \times_1 \boldsymbol{V}_1^\top \times_2 \boldsymbol{V}_2^\top \times_3 \boldsymbol{V}_3^\top\right\|_F^2 \\ \text{subject to} \quad & \boldsymbol{V}_1 \in \mathbb{O}_{p_1,r_1}, \boldsymbol{V}_2 \in \mathbb{O}_{p_2,r_2}, \boldsymbol{V}_3 \in \mathbb{O}_{p_3,r_3}. \end{aligned}$$

- **Advantage**: achieves statistical optimality. (will be shown later)
- **Disadvantage**:
  - Non-convex, computational intractable.
  - NP-hard to approximate even $r = 1$ (Hillar and Lim, 2013).

Phase Transition in Tensor SVD

- The difficulty is driven by signal-to-noise ratio (SNR).

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

$$\sigma = \text{SD}(Z) = \text{noise level.}$$

- Suppose $p_1 \asymp p_2 \asymp p_3 \asymp p$. Three phases:

$$\lambda/\sigma \geq Cp^{3/4} \quad \text{(Strong SNR case)},$$
$$\lambda/\sigma < cp^{1/2} \quad \text{(Weak SNR case)},$$
$$p^{1/2} \ll \lambda/\sigma \ll p^{3/4} \quad \text{(Moderate SNR case)}.$$

Strong SNR Case: Methodology

- When $\lambda/\sigma \geq Cp^{3/4}$, apply higher-order orthogonal iteration (HOOI).
  (De Lathauwer, Moor, and Vandewalle, SIAM. J. Matrix Anal. & Appl. 2000b)
- (Step 1. Spectral initialization)

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

- (Step 2. Power iterations)
  Repeat Let $t = t + 1$. Calculate

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

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

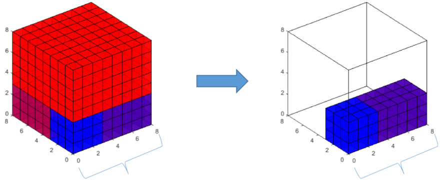
$$\hat{\boldsymbol{U}}_3^{(t)} = \text{SVD}_{r_3}\left(\mathcal{M}_3(\boldsymbol{\mathcal{Y}} \times_1 (\hat{\boldsymbol{U}}_1^{(t)})^\top \times_2 (\hat{\boldsymbol{U}}_2^{(t)})^\top)\right).$$

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

Interpretation

1. Spectral initialization provides a "warm start."
2. Power iteration refines the initializations.
   Given $\hat{\boldsymbol{U}}_1^{(t-1)}, \hat{\boldsymbol{U}}_2^{(t-1)}, \hat{\boldsymbol{U}}_3^{(t-1)}$, denoise $\boldsymbol{\mathcal{Y}}$ via:

$$\boldsymbol{\mathcal{Y}} \times_2 \hat{\boldsymbol{U}}_2^{(t-1)} \times_3 \hat{\boldsymbol{U}}_3^{(t-1)}.$$



- Mode-1 singular subspace is reserved;
- Noise can be highly reduced.

Thus, we update

$$\hat{\boldsymbol{U}}_1^{(t)} = \mathrm{SVD}_{r_1}\left(\mathcal{M}_{r_1}\left(\boldsymbol{\mathcal{Y}} \times_2 \hat{\boldsymbol{U}}_2^{(t-1)} \times_3 \hat{\boldsymbol{U}}_3^{(t-1)}\right)\right).$$

Higher-order orthogonal iteration (HOOI)

(De Lathauwer, Moor, and Vandewalle, SIAM. J. Matrix Anal. & Appl. 2000b)

On the Best Rank-1 and Rank-(R_1 ,R_2 ,. . .,R_N) Approximation of Higher-Order Tensors

L De Lathauwer, B De Moor, J Vandewalle - SIAM journal on Matrix Analysis …, 2000 - SIAM

In this paper we discuss a multilinear generalization of the best rank-R approximation problem for matrices, namely, the approximation of a given higher-order tensor, in an optimal least-squares sense, by a tensor that has prespecified column rank value, row rank

Cited by 1196 Related articles

Strong SNR Case: Theoretical Analysis

Theorem (Upper Bound)

Suppose $\lambda/\sigma > Cp^{3/4}$ and other regularity conditions hold, after at most $O\left(\log(p/\lambda) \vee 1\right)$ iterations,

- *(Recovery of $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$)*

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

- *(Recovery of $\boldsymbol{\mathcal{X}}$)*

$$\sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}_{p,r}(\lambda)} \max_{k=1,2,3} \mathbb{E} \left\| \hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}} \right\|_F^2 \leq C\left(p_1 r_1 + p_2 r_2 + p_3 r_3\right)\sigma^2,$$

$$\sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}_{p,r}(\lambda)} \max_{k=1,2,3} \mathbb{E} \frac{\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\|_F^2}{\|\boldsymbol{\mathcal{X}}\|_F^2} \leq \frac{C\left(p_1 + p_2 + p_3\right)\sigma^2}{\lambda^2}.$$

Strong SNR Case: Lower Bound

Define the following class of low-rank tensors with signal strength λ.

$$\mathcal{F}_{p,r}(\lambda) = \{\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3} : \text{rank}(\mathcal{X}) = (r_1, r_2, r_3), \sigma_{r_k}(\mathcal{M}_k(\mathcal{X})) \geq \lambda\}$$

Theorem (Lower Bound)

(Recovery of $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$)

$$\inf_{\tilde{\boldsymbol{U}}_k} \sup_{\mathcal{X} \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \min_{O \in \mathbb{O}_r} \left\| \tilde{\boldsymbol{U}}_k - \boldsymbol{U}_k O \right\|_F \geq c \frac{\sqrt{p_k r_k}}{\lambda/\sigma}, \quad k = 1, 2, 3.$$

(Recovery of $\mathcal{X}$)

$$\inf_{\hat{\mathcal{X}}} \sup_{\mathcal{X} \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \left\| \hat{\mathcal{X}} - \mathcal{X} \right\|_F^2 \geq c(p_1 r_1 + p_2 r_2 + p_3 r_3)\sigma^2,$$

$$\inf_{\hat{\mathcal{X}}} \sup_{\mathcal{X} \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \frac{\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2}{\|\mathcal{X}\|_F^2} \geq \frac{c(p_1 + p_2 + p_3)\sigma^2}{\lambda^2}.$$

HOSVD vs. HOOI

$$\mathbb{E}\min_{O\in\mathbb{O}_r}\|\hat{\boldsymbol{U}}_k^{HOSVD}-\boldsymbol{U}_kO\|_F \asymp \frac{\sqrt{p_kr_k}}{\lambda/\sigma}+\frac{\sqrt{p_1p_2p_3r_k}}{(\lambda/\sigma)^2};$$

$$\mathbb{E}\min_{O\in\mathbb{O}_r}\|\hat{\boldsymbol{U}}_k^{HOOI}-\boldsymbol{U}_kO\|_F \asymp \frac{\sqrt{p_kr_k}}{\lambda/\sigma}.$$

- When $\lambda/\sigma \leq cp$, HOOI significantly improves upon HOSVD.
- The analysis for rank-r tensor SVD is more difficult than both rank-1 tensor SVD or rank-r matrix SVD.
  - Many concepts (e.g. singular values) are not well defined for tensors.

Weak SNR Case

Under the weak SNR case $\lambda/\sigma < cp^{1/2}$, $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$, or $\boldsymbol{\mathcal{X}}$ cannot be stably estimated in general.

Theorem

(Recovery of $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$)

$$\inf_{\hat{\boldsymbol{U}}_k} \sup_{X \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \min_{O \in \mathbb{O}_r} r_k^{-1/2} \|\hat{\boldsymbol{U}}_k - \boldsymbol{U}_k O\|_F \geq c, \quad k = 1, 2, 3.$$

(Recovery of $\boldsymbol{\mathcal{X}}$)

$$\inf_{\hat{\boldsymbol{\mathcal{X}}}} \sup_{\mathcal{X} \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \frac{\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\|_F^2}{\|\boldsymbol{\mathcal{X}}\|_F^2} \geq c.$$

Moderate SNR Case

- Recall the SNR λ/σ measures the problem difficulty.

$$\lambda = \min_{k=1,2,3} \sigma_r(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$$
$$\sigma = \text{SD}(Z).$$

- For moderate signal case: $Cp^{1/2} \leq \lambda/\sigma \leq cp^{3/4}$, there exists a gap between computational and statistical optimality.

Moderate SNR Case: Statistical Optimality

- First, MLE achieves statistical optimality.

Theorem (Performance of MLE Estimator)

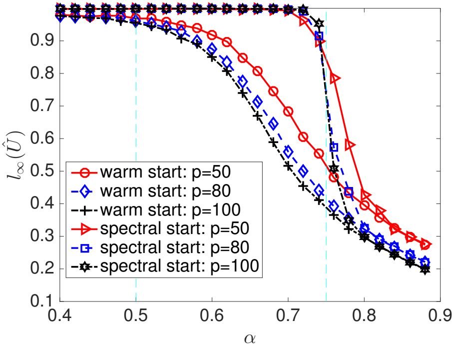
When $\lambda/\sigma \geq Cp^{1/2}$,

- *(Recovery of* $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$*)*
$$\sup_{X \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \min_{O \in \mathbb{O}_r} \left\| \hat{\boldsymbol{U}}_k^{mle} - \boldsymbol{U}_k O \right\|_F \leq C \frac{\sqrt{p_k r_k}}{\lambda/\sigma}, \quad k = 1, 2, 3;$$
- *(Recovery of* $\boldsymbol{\mathcal{X}}$*)*
$$\sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \left\| \hat{\boldsymbol{\mathcal{X}}}^{mle} - \boldsymbol{\mathcal{X}} \right\|_F^2 \leq C \left(p_1 r_1 + p_2 r_2 + p_3 r_3 \right) \sigma^2,$$
$$\sup_{X \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \frac{\|\hat{\boldsymbol{\mathcal{X}}}^{mle} - \boldsymbol{\mathcal{X}}\|_F^2}{\|\boldsymbol{\mathcal{X}}\|_F^2} \leq \frac{C \left(p_1 + p_2 + p_3 \right) \sigma^2}{\lambda^2}.$$

- However MLE is computationally intractable.

Simulation Analysis

- Consider random settings: $\lambda = p^{\alpha}$, $\alpha \in [.4, .9]$, $\sigma = 1$.



- Two phase transitions:
  - The computational inefficient method performs well starting at $\lambda/\sigma \approx p^{1/2}$;
  - The computational efficient HOOI performs well starting at $\lambda/\sigma \approx p^{3/4}$.

Moderate SNR Case: Computational Optimality

Moreover, the following theorem shows the computational hardness for polynomial-time algorithms under moderate SNR.

Theorem

Assume the conjecture of hypergraphic planted clique holds, and $\lambda/\sigma = O(p^{3(1-\tau)/4})$ for any $\tau > 0$, then for any polynomial-time algorithm $\hat{\boldsymbol{U}}_1, \hat{\boldsymbol{U}}_2, \hat{\boldsymbol{U}}_3, \hat{\boldsymbol{\mathcal{X}}}$,

(Recovery of $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$)

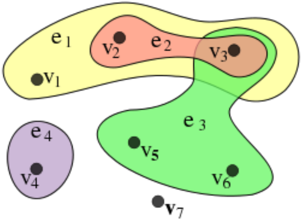
$$\liminf_{p\to\infty} \sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}_{p,r}(\lambda)} \mathbb{E}\left\|\sin\Theta(\hat{\boldsymbol{U}}_k^{(p)}, \boldsymbol{U}_k)\right\|^2 \geq c_1, \quad k = 1, 2, 3,$$

(Recovery of $\boldsymbol{\mathcal{X}}$)

$$\liminf_{p\to\infty} \sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}_{p,r}(\lambda)} \frac{\mathbb{E}\|\hat{\boldsymbol{\mathcal{X}}}^{(p)} - \boldsymbol{\mathcal{X}}\|_F^2}{\|\boldsymbol{\mathcal{X}}\|_F^2} \geq c_1.$$

Remarks

- The analysis relies on the hypergrahic planted clique detection assumption.



- Result shows the hardness of tensor SVD in moderate SNR case.
- More recently, Ben Arous, Mei, Montanari, Nica (2017) analyzed the landscape of rank-1 spiked tensor model.
  - MLE is with exponentially growing many critical points.

Summary

Tensor SVD exhibits three phases,

- (Strong SNR) $\lambda/\sigma \geq Cp^{3/4}$,
  $\rightarrow$ there is efficient algorithm to estimate $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$, and $\boldsymbol{\mathcal{X}}$.

- (Weak SNR) $\lambda/\sigma < cp^{1/2}$,
  $\rightarrow$ no algorithm can stably recover $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$, or $\boldsymbol{\mathcal{X}}$.

- (Moderate SNR) $p^{1/2} \ll \lambda/\sigma \ll p^{3/4}$,
  - non-convex MLE stably recovers $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$, and $\boldsymbol{\mathcal{X}}$;
  - Maybe no polynomial time algorithm performs stably.

Further Generalization to Order-d Tensors

- The results can be generalized to order-d tensors.
- Three phases
  - (Strong SNR) $\lambda/\sigma \geq Cp^{d/4}$,
    $\rightarrow$ Efficient algorithm exists.
  - (Weak SNR) $\lambda/\sigma < cp^{1/2}$,
    $\rightarrow$ No algorithm exists.
  - (Moderate SNR) $p^{1/2} \ll \lambda/\sigma \ll p^{d/4}$,
    - Inefficient algorithm exists;
    - Maybe no polynomial time algorithm performs stably.
- 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.

Part II: Sparse Tensor SVD

Limitation of tensor SVD model

- Higher-order orthogonal iteration (HOOI) is both efficient and minimax-optimal.

$$\inf_{\tilde{U}_k} \sup_{\mathcal{X}} \mathbb{E} \max_{O \in \mathbb{O}_{r_k}} \left\| \tilde{U}_k - U_k O \right\|_F \asymp \frac{\sqrt{p_k r_k}}{\lambda/\sigma}.$$

- **The problem is not completely solved by HOOI!**

- Pitfalls:
  1. SNR requirement: $\lambda/\sigma \geq p^{d/4}$.
     → It is necessary without further conditions.
     → may be too stringent for high-dimensional data.

  2. HOOI is suboptimal when tensor data satisfy structural assumption.
     → Sparsity commonly appear in high-dimensional applications.

Sparsity may occur only in part of modes (directions).

- Motivating example: electroencephalogram (EEG) dataset:

Brain electrical Activity vs. Subject $\times$ Electrodes $\times$ Time.

1. Data are likely to be dense on Mode Subject;
2. Data along Mode Electrodes may be sparse.
3. Data along Mode Time after transformation is possibly sparse.

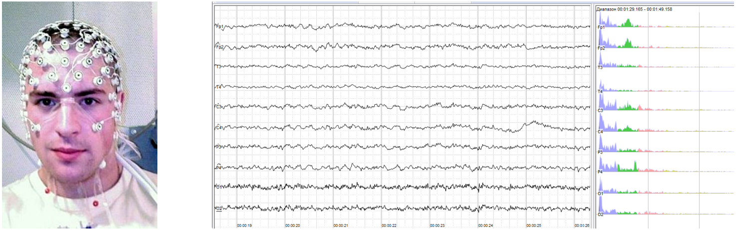


Figure: Illustration of electroencephalogram (Source: Wikipedia)

Sparse Tensor SVD Model

$$\mathcal{Y} = \mathcal{X} + \mathcal{Z} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times \cdots \times_d \boldsymbol{U}_d + \mathcal{Z},$$

- $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is the observation;
- $\mathcal{Z}$ is the noise of small amplitude;
- $\mathcal{X}$ is the sparse low-rank tensor;
- Loadings: $\boldsymbol{U}_k \in \mathbb{R}^{p_k \times r_k}$.
  A subset of modes $J_s \subseteq [d]$ satisfy row-wise sparsity,

$$\|\boldsymbol{U}_k\|_0 = \sum_{i=1}^{p_k} 1_{\{\boldsymbol{U}_{k,[i,:]} \neq 0\}} \leq s_k,;$$

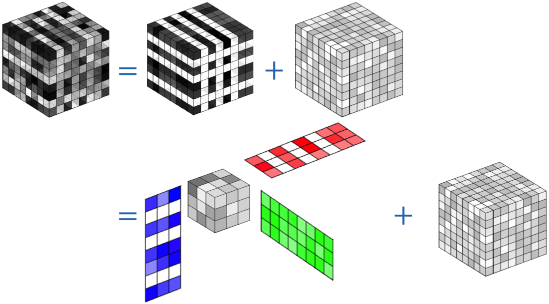
$$s_k \ll p_k, \quad k \in J_s; \quad s_k = p_k, \quad k \notin J_s.$$

A specific setting of sparse tensor SVD model

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

$$\mathcal{Z} \overset{iid}{\sim} N(0, \sigma^2), \quad \mathcal{S} \in \mathbb{R}^{r \times r \times r}, \quad J_s = \{1, 3\}.$$

$$\boldsymbol{U}_k \in \mathbb{O}_{p,r}, \quad \|\boldsymbol{U}_1\|_0 \leq s, \quad \|\boldsymbol{U}_3\|_0 \leq s, \quad \|\boldsymbol{U}_2\|_0 \leq p.$$



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

Straightforward Ideas

- Penalized MLE:

$$\min_{\boldsymbol{U}_1,\boldsymbol{U}_2,\boldsymbol{U}_3,\boldsymbol{\mathcal{S}}} \|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{S}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3\|_F^2 + \lambda\|\boldsymbol{U}_1\|_1 + \lambda\|\boldsymbol{U}_3\|_1.$$

  → computationally difficult

- High-order orthogonal iteration (HOOI) and high-order SVD (HOSVD):
  → ignore sparse patterns.

- S-HOOI and S-HOSVD:
  → In each update of HOOI or HOSVD, apply matrix sparse SVD.
  References: Lee, Shen, Huang, Marron, 2010; Yang, Ma, Buja, 2014, 2016.
  → ignore tensor structures.

Methodology

Step 1. Initialization

- (Support initialization) Select the index set

$$\hat{I}_k^{(0)} = \left\{ i_k : \|\boldsymbol{\mathcal{Y}}_{[\cdots i_k \cdots]}\|_2^2 \geq \lambda_1 \text{ or } \left\|\boldsymbol{\mathcal{Y}}_{[\cdots i_k \cdots]}\right\|_\infty \geq \lambda_2 \right\}, \quad k = 1, 3.$$

  Here, $\lambda_1 = \sigma^2 \left(p^2 + 2\sqrt{p^2 \log p} + 2 \log p \right)$; $\lambda_2 = 2\sigma \sqrt{\log(p^2)}$.

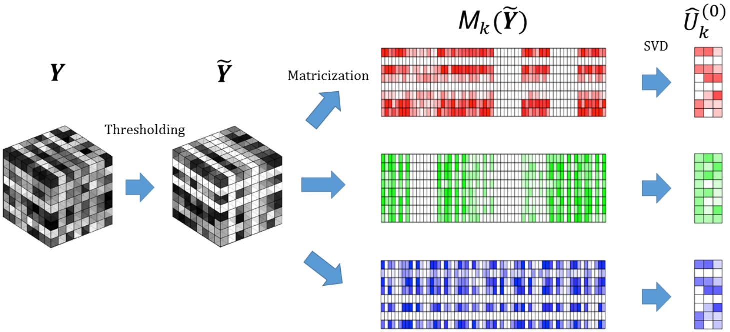
- (Singular subspace initialization) Construct

$$\tilde{\boldsymbol{\mathcal{Y}}}_{[i_1, i_2, i_3]} = \begin{cases} \boldsymbol{\mathcal{Y}}_{[i_1, i_2, i_d]}, & i_1 \in \hat{I}_1^{(0)}, i_3 \in \hat{I}_3^{(0)}, \\ 0, & \text{otherwise.} \end{cases}$$

  and initialize

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

Methodology: initialization



- $\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)}$ provide convenient initial estimates for U_1, U_2, U_3.

Methodology: Iterative Updates

Step 2. Alternating Updates

- For $t = 0, 1, \ldots$, perform alternating updates

$$\hat{\boldsymbol{U}}_1^{(t)} \to \hat{\boldsymbol{U}}_1^{(t+1)} \quad \text{with} \quad \boldsymbol{\mathcal{Y}}, \hat{\boldsymbol{U}}_2^{(t)}, \hat{\boldsymbol{U}}_3^{(t)};$$
$$\hat{\boldsymbol{U}}_2^{(t)} \to \hat{\boldsymbol{U}}_2^{(t+1)} \quad \text{with} \quad \boldsymbol{\mathcal{Y}}, \hat{\boldsymbol{U}}_1^{(t+1)}, \hat{\boldsymbol{U}}_3^{(t)};$$
$$\hat{\boldsymbol{U}}_3^{(t)} \to \hat{\boldsymbol{U}}_3^{(t+1)} \quad \text{with} \quad \boldsymbol{\mathcal{Y}}, \hat{\boldsymbol{U}}_1^{(t+1)}, \hat{\boldsymbol{U}}_2^{(t+1)}.$$

- Two scenarios:

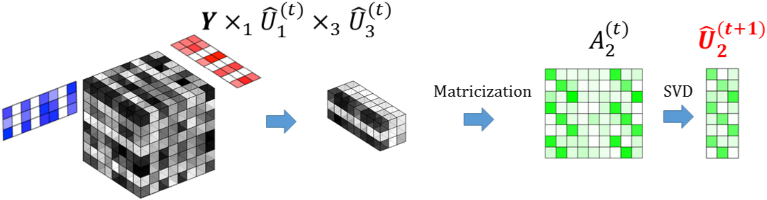
non-sparse mode $k \notin J_s$ and sparse mode $k \in J_s$.

Step 2(a): Update for non-sparse mode

- When $k \notin J_s$, such as $k = 2$, calculate

$$A_2^{(t)} = \mathcal{M}_k\left(\boldsymbol{\mathcal{Y}} \times_1 \hat{\boldsymbol{U}}_1^{(t+1)} \times_3 \hat{\boldsymbol{U}}_3^{(t)}\right) \in \mathbb{R}^{p\times r^2}.$$

$$\hat{\boldsymbol{U}}_2^{(t)} = \mathrm{SVD}_r\left(A_2^{(t)}\right) \in \mathbb{O}_{p,r}.$$



- The update is similar to HOOI.

Step 2(b): Update for sparse mode: double projection & thresholding

- When $k \in J_s$, for example $k = 1$,
  - (i) (First Projection)

$$A_1^{(t)} = \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} \times_2 (\hat{\boldsymbol{U}}_2^{(t)})^\top \times_3 (\boldsymbol{U}_3^{(t)})^\top\right).$$

  - (ii) (First Thresholding)

$$B_{1,[i,:]}^{(t)} = A_{1,[i,:]}^{(t)} 1_{\{\|A_{1,[i,:]}^{(t)}\|_2^2 \geq \eta\}}.$$

  - (iii) (Second Projection)

$$\bar{B}_1^{(t)} = B_1^{(t)} \hat{V}_1^{(t)}, \quad \hat{V}_1^{(t)} = \text{leading } r \text{ right singular vectors of } B_1^{(t)}.$$

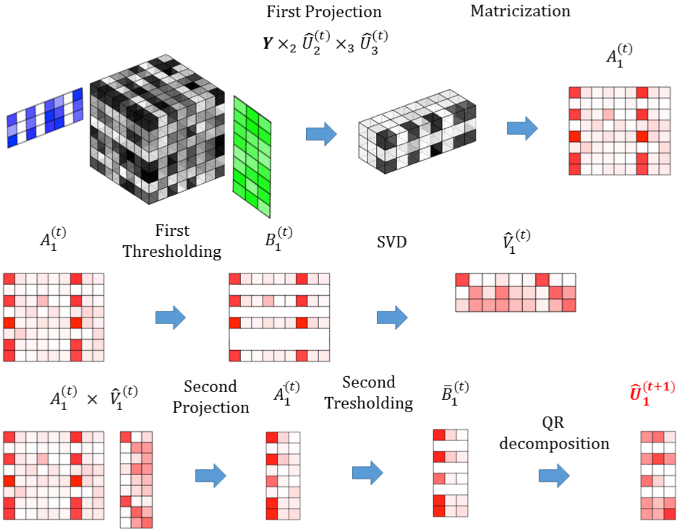
  - (iv) (Second Thresholding)

$$\bar{B}_{1,[i,:]}^{(t)} = \bar{A}_{1,[i,:]}^{(t)} 1_{\{\|\bar{A}_{1,[i,:]}^{(t)}\|_2^2 \geq \bar{\eta}\}}.$$

  - (v) (Orthogonalization)

    Apply QR decomposition to $\bar{B}_1^{(t)}$, assign Q part to $\hat{\boldsymbol{U}}_1^{(t+1)}$.

Methodology: Iterative Updates



Methodology: Final Estimation

Step 3: Final Estimation

- Break from the iterative loop after
    1. maximum of number iteration is reached; or
    2. convergence.
- Obtain
$$\hat{U}_1, \hat{U}_2, \hat{U}_3$$
- Estimate $\boldsymbol{\mathcal{X}}$ by
$$\hat{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{Y}} \times_1 P_{\hat{U}_1} \times_2 P_{\hat{U}_2} \times_3 P_{\hat{U}_3}$$

Remarks

Sparse Tensor Alternating Thresholding SVD **(STAT-SVD)**

- Why so complicated, especially in Step 2(b)?
  - In each step, we need to truncate after an appropriate projection.
  - Double projection & thresholding ensure better statistical accuracy.
  - Analogy: tumor surgery.

Theoretical Analysis

Assume

(1)
$$\begin{aligned}\lambda_k =& \sigma_{\min}(\mathcal{M}(\boldsymbol{\mathcal{X}}_k))\\ \geq& C\sigma\left(\sqrt{(\Pi_k s_k)\cdot \log p} \vee \max_k s_k r_k \vee \frac{r_1\cdots r_d}{\min_k r_k}\right).\end{aligned}$$

Theorem (Upper Bound)

Under (1), after at most a logarithm factor of iterations, STAT-SVD yields,

$$\left\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\right\|_F^2 \leq C\sigma^2\left(r_1\cdots r_d + \sum s_k r_k + \sum_{k\in J_s} s_k \log p_k\right),$$

$$\max_{O\in\mathbb{O}_{r_k}} \left\|\hat{\boldsymbol{U}}_k - \boldsymbol{U}_k O\right\|_F \leq \begin{cases} C(\sqrt{s_k r_k} + \sqrt{s_k \log p_k})/\lambda_k, & k \in J_s,\\ C\sqrt{s_k r_k}/\lambda_k, & k \notin J_s,\end{cases}$$

with high probability.

Remark

Error Bound:

$$\left\|\hat{\mathcal{X}} - \mathcal{X}\right\|_F^2 \leq C\sigma^2\left(r_1\cdots r_d + \sum s_k r_k + \sum_{k\in J_s} s_k \log p_k\right),$$

- $\sigma^2 r_1 \cdots r_d$: complexity in estimating the core tensor;
- $\sigma^2 s_k r_k$: complexity in estimating the values of loadings;
- $\sigma^2 s_k \log p_k$: complexity in estimating the support of loadings $\rightarrow$ only exists in sparse modes $k \in J_s$.

SNR Assumption:

$$\lambda/\sigma \geq C\left(\sqrt{(\Pi_k s_k)\cdot \log p} \vee \max_k s_k r_k \vee \frac{r_1\cdots r_d}{\min_k r_k}\right).$$

- p only appear in logarithms.

Theoretical Analysis

We define the following class of sparse and low-rank tensors,

$$\mathcal{F}_{p,r}(s,\lambda) = \left\{ \boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1\times\cdots\times p_d} : \begin{array}{l} \text{rank}(X) \leq (r_1,\ldots,r_d); \\ \sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{X}})) \geq \lambda_k; \|\boldsymbol{U}_k\|_0 \leq s_k \end{array} \right\}.$$

Theorem (Lower Bound)

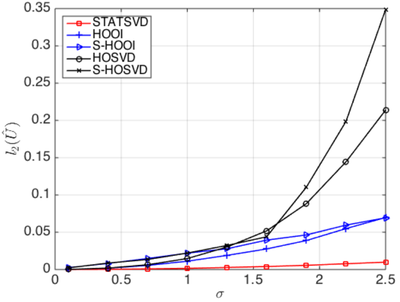
Suppose $p_k \geq s_k \geq r_k$, $r_{-k} \geq 4r_k$,

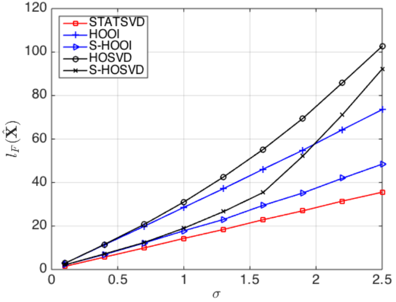
$$\inf_{\hat{\boldsymbol{\mathcal{X}}}} \sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}_{p,s,r}} \mathbb{E}\left\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\right\|_F^2 \geq c\sigma^2\left(r_1\cdots r_d + \sum s_k r_k + \sum_{k\in J_s} s_k \log p_k\right).$$

$$\inf_{\hat{\boldsymbol{U}}_k} \sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}_{p,r}(s,\lambda)} \mathbb{E} \max_{O\in\mathbb{O}_{p_k,r_k}} \left\|\hat{\boldsymbol{U}}_k - \boldsymbol{U}_k O\right\|_F \geq \begin{cases} \dfrac{c\left(\sqrt{s_k r_k} + \sqrt{s_k \log(p_k/s_k)}\right)}{\lambda_k}, & k\in J_s; \\ \dfrac{c\sqrt{s_k r_k}}{\lambda_k}, & k\notin J_s. \end{cases}$$

Simulation Study

- $p = 50$, $s = 10$, $r = 5$.

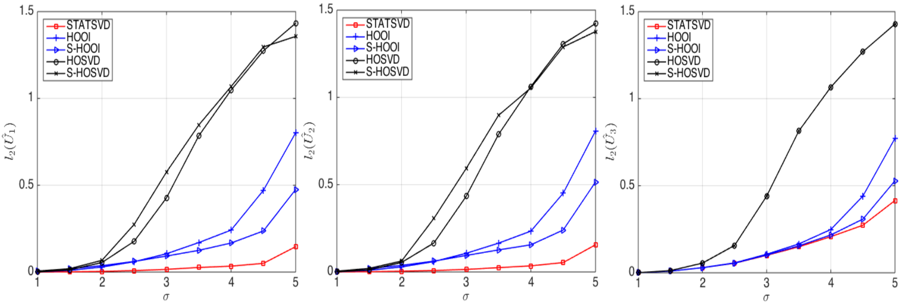




- STAT-SVD outperforms HOOI, HOSVD, S-HOOI, S-HOSVD.

Simulation Study 2

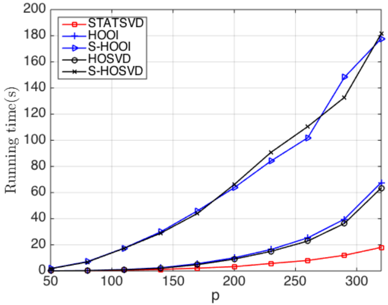
- $p = 50$, $r = 5$, $J_s = 1, 2$, $s_1 = s_2 = 10$. $s_3 = 50$.



- Mode-3 is non-sparse, but STAT-SVD still outperforms other methods. → **Three modes of a tensor are a union.**

Simulation Study 3

- $r = 5$, $s = 10$, p grows.
- We record the running time for each method



- **STAT-SVD is fast.**

Summary

- We propose a general framework for sparse tensor SVD, and an efficient algorithm: STAT-SVD.

- STAT-SVD achieves
  - optimal rate of convergence;
  - good numercial performance.

- Applications: Longitudinal data, EEG data, molecule tomography, ...

- Further questions:
  - Results are all based on strong SNR assumption.
    $\rightarrow$ What if SNR is not strong?
    $\rightarrow$ Any phase transition effect in sparse tensor SVD model?

References

- Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and Computational Limits. *IEEE Transactions on Information Theory*, to appear.
- Zhang, A. and Han, R. (2018). Optimal Denoising and Singular Value Decomposition for Sparse High-dimensional High-order Data. *Journal of the American Statistical Association*, to appear.
- Cai, T. and Zhang, A. (2018). Rate-Optimal Perturbation Bounds for Singular Subspaces with Applications to High-Dimensional Statistics. *Annals of Statistics*, to appear.