Singular Value Decomposition for High-dimensional High-order Data

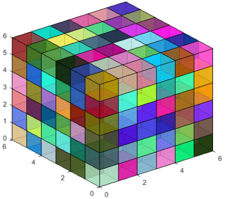
Anru Zhang
Department of Statistics
University of Wisconsin-Madison

CSML @ Princeton University
May 14, 2018

Joint work with Dong Xia

Introduction

- Tensors, or high-order arrays, attract lots of attention recently.

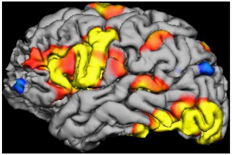


- e.g. an order-d tensor

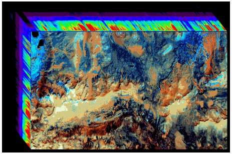
$$\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_d}, \quad \mathcal{X} = \left(X_{i_1 \cdots i_d}\right), \quad 1 \le i_k \le p_k, \quad k = 1, \cdots, d.$$

Example

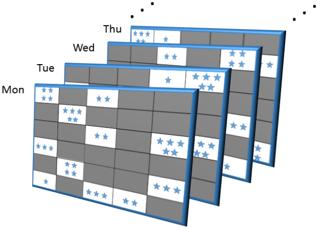
- Brain imaging



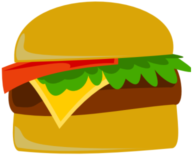
- Hyperspectral imaging



- Recommender system

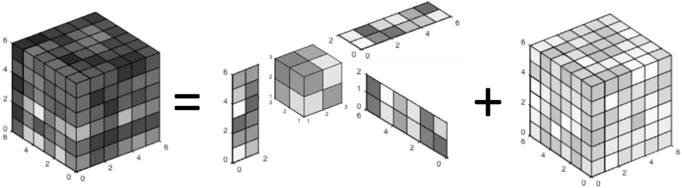


- **Higher order is fancier!**



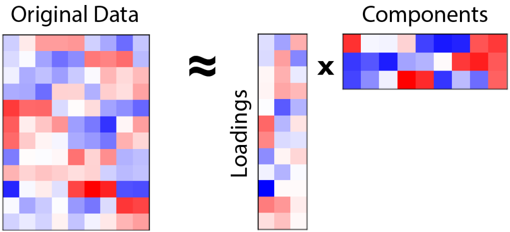

- Higher order tensor problems are **far more than** extension of matrices.
    - More structures
    - High-dimensionality
    - Computational difficulty

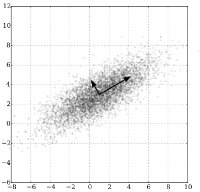
In this talk, we focus on **tensor SVD**.



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**.





- Variations: sparse PCA, robust PCA, sparse SVD, kernel SVD, ...

Related Works

- Rank-1 Tensor SVD: Richard and Montanari, 2014; Hopkins, Shi, Steurer, 2015; Perry, Wein, Bandeira, 2017, Anandkumar, Deng, Ge, Mobahi, 2017.

$$\mathcal{Y} = \lambda \cdot u \otimes v \otimes w + \mathcal{Z}, \quad \mathcal{Z} \overset{iid}{\sim} N(0, \sigma^2).$$

- Methods:
  - power methods, sum-of-squares, approximate message passing, homotopy...
  - MLE, warm-start power iterations...
- Statistical and computational trade-off & phase transition effects.
- The statistical framework for tensor SVD when $r \geq 2$ is not well defined or solved.

Tensor SVD

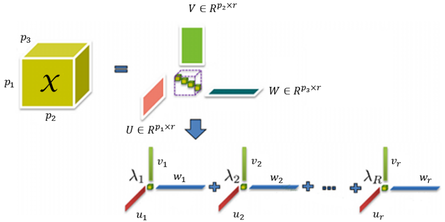
- We propose a general framework for tensor SVD.
- $$\mathcal{Y} = \mathcal{X} + \mathcal{Z},$$

  where
  - $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is the observation;
  - $\mathcal{Z}$ is the noise;
  - $\mathcal{X}$ 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.

Low-rankness for Tensors

- Canonical polyadic (CP) low-rankness is widely used in literature. Definition:

$$r_{cp} = \min_r \quad \text{s.t.} \quad \mathcal{X} = \sum_{i=1}^{r} \lambda_i \cdot u_i \otimes v_i \otimes w_i.$$

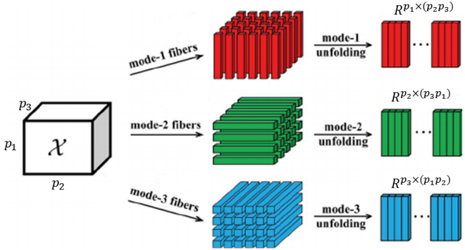


- Disadvantages:
  - possibly $r_{cp} > \max\{p_1, p_2, p_3\}$;
  - the set of rank-r tensors may **not be close**;
    Possible situation: limit of a series of cp rank-2 tensors is of rank 3!
  - u_i's (v_i, w_i's) are usually **not orthogonal**.

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

Tucker Low-rankness

- If X_1, X_2, and X_3 are **matricizations** of $\mathcal{X}$,



- We assume $r_1 = \text{rank}(X_1)$, $r_2 = \text{rank}(X_2)$, $r_3 = \text{rank}(X_3)$, and denote

$$\text{Tucker rank}(\mathcal{X}) = (r_1, r_2, r_3).$$

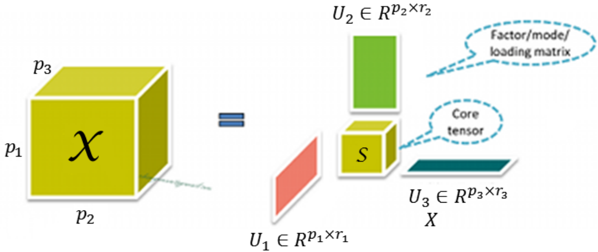
- Note:
  Matrix: dim(row-span) = dim(column-span).
  Tensor: parallel definitions, r_1, r_2, r_3, are not necessarily equal.

Picture Source: Tao, M., Zhou, F., Liu, Y. and Zhang, Z. (2015)

More General Assumption: Tucker Low-rankness

- Equivalent form of definition: Tucker decomposition

$$\mathcal{X} = \mathcal{S} \times_1 U_1 \times_2 U_2 \times_3 U_3$$



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

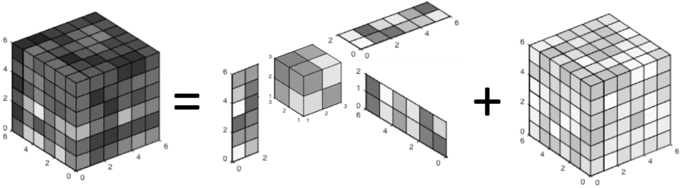
Picture Source: Guoxu Zhou's website. http://www.bsp.brain.riken.jp/ zhougx/tensor.html

Model

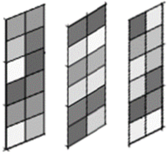
- Observations: $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$,

$$\mathcal{Y} = \mathcal{X} + \mathcal{Z} = \mathcal{S} \times_1 U_1 \times_2 U_2 \times_3 U_3 + \mathcal{Z},$$

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



- Goal: estimate U_1, U_2, U_3, and the original tensor $\mathcal{X}$.

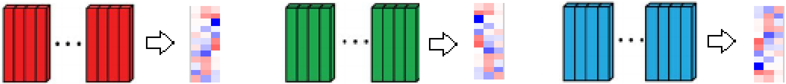


Straightforward Idea 1: Higher order SVD (HOSVD)

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

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

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



Note: $\mathrm{SVD}_r(\cdot)$ 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

- **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

$$\hat{U}_1^{mle}, \hat{U}_2^{mle}, \hat{U}_3^{mle}, \hat{\mathcal{S}}^{mle} = \underset{U_1,U_2,U_3,\mathcal{S}}{\operatorname{argmax}} \|\mathcal{Y} - \mathcal{S} \times_1 U_1 \times_2 U_2 \times_3 U_3\|_F^2$$

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

$$\begin{aligned} \max \quad & \left\|\mathcal{Y} \times_1 V_1^\top \times_2 V_2^\top \times_3 V_3^\top\right\|_F^2 \\ \text{subject to} \quad & V_1 \in \mathbb{O}_{p_1,r_1}, V_2 \in \mathbb{O}_{p_2,r_2}, V_3 \in \mathbb{O}_{p_3,r_3}. \end{aligned}$$

- **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)

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

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

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

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

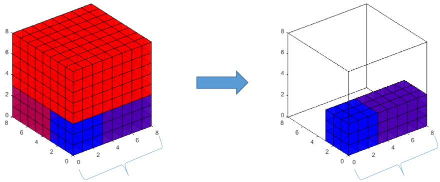
$$\hat{U}_3^{(t)} = \mathrm{SVD}_{r_3}\left(\mathcal{M}_3(\mathcal{Y} \times_1 (\hat{U}_1^{(t)})^\top \times_2 (\hat{U}_2^{(t)})^\top)\right).$$

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

Interpretation

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

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



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

Thus, we update

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

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

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* U_1, U_2, U_3*)*

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

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

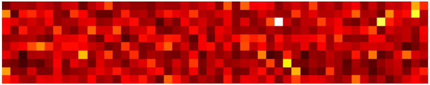
$$\sup_{\mathcal{X} \in \mathcal{F}_{p,r}(\lambda)} \max_{k=1,2,3} \mathbb{E} \left\| \hat{\mathcal{X}} - \mathcal{X} \right\|_F^2 \leq C\left(p_1 r_1 + p_2 r_2 + p_3 r_3\right) \sigma^2,$$

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

Key Tool

- Key tool: one-sided perturbation bound.
- The matricizations $\mathcal{M}_1(\mathcal{Y}) \in \mathbb{R}^{p_1\times(p_2p_3)}$ are flat,

$$\mathcal{M}_1(\mathcal{Y}) = \mathcal{M}_1(\mathcal{X}) + \mathcal{M}_1(\mathcal{Z}) \in \mathbb{R}^{p_1\times(p_2p_3)}$$



$\mathcal{M}_1(\mathcal{Y})$ is observed, $\quad$ rank$(\mathcal{M}_1(\mathcal{X})) \le r_1$, $\quad \mathcal{M}_1(\mathcal{Z}) \overset{iid}{\sim} N(0,\sigma^2)$.

- Let

$$\hat{U}_1 = \text{SVD}_r(\mathcal{M}_1(\mathcal{Y})), \quad U_1 = \text{SVD}_r(\mathcal{M}_1(\mathcal{X})),$$
$$\hat{V}_1 = \text{SVD}_r(\mathcal{M}_1(\mathcal{Y})^\top), \quad V_1 = \text{SVD}_r(\mathcal{M}_1(\mathcal{X})^\top).$$

Naturally, the left singular subspace $\hat{U}_1$ of $\mathcal{M}_1(\mathcal{Y})$ is more "informative" than the right one $\hat{V}_1$.

One-sided Perturbation Analysis

- Traditional perturbation analysis were usually two-sided, e.g., Wedin's lemma (Wedin, 1972)

$$\max\left\{\|\sin\Theta(\hat{U}_1, U_1)\|_F, \|\sin\Theta(\hat{V}_1, V_1)\|_F\right\} \leq \ldots$$

- One-sided perturbation bound (Cai and Z. 2018)

$$Y = X + Z, \quad X, Y, Z \in \mathbb{R}^{p_1\times(p_2p_3)},$$

$$\text{rank}(X) = r, \quad \sigma_r(X) = \lambda, \quad Z \overset{iid}{\sim} N(0,1).$$

Theorem

$$\mathbb{E}\|\sin\Theta(\hat{U}_1, U_1)\|^2 \asymp \frac{p_1}{\lambda^2} + \frac{p_1p_2p_3}{\lambda^4},$$

$$\mathbb{E}\|\sin\Theta(\hat{V}_1, V_1)\|^2 \asymp \frac{p_2p_3}{\lambda^2} + \frac{p_1p_2p_3}{\lambda^4}.$$

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 U_1, U_2, U_3)

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

(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

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

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

- When $\lambda/\sigma \leq cp$, HOOI significantly improves upon HOSVD.
- Analysis of rank-r tensor SVD is more difficult than 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 $\lambda/\sigma < cp^{1/2}$, U_1, U_2, U_3, or $\mathcal{X}$ cannot be stably estimated in general.

Theorem

(Recovery of U_1, U_2, U_3)

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

(Recovery of $\mathcal{X}$)

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

Moderate SNR Case: Statistical Optimality

- First, MLE achieves statistical optimality.

Theorem (Performance of MLE Estimator)

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

- *(Recovery of U_1, U_2, U_3)*

$$\sup_{X \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \min_{O \in \mathbb{O}_r} \left\| \hat{U}_k^{mle} - U_k O \right\|_F \leq C \frac{\sqrt{p_k r_k}}{\lambda/\sigma}, \quad k = 1, 2, 3;$$

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

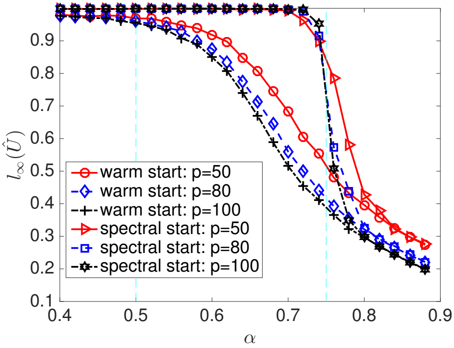
$$\sup_{\mathcal{X} \in \mathcal{F}_{p,r}(\lambda)} \mathbb{E} \left\| \hat{\mathcal{X}}^{mle} - \mathcal{X} \right\|_F^2 \leq C \left(p_1 r_1 + p_2 r_2 + p_3 r_3 \right) \sigma^2,$$

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

- 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 $\hat{U}_1, \hat{U}_2, \hat{U}_3, \hat{\mathcal{X}}$,

(Recovery of U_1, U_2, U_3)

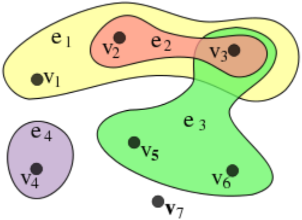
$$\liminf_{p\to\infty} \sup_{\mathcal{X}\in\mathcal{F}_{p,r}(\lambda)} \mathbb{E}\left\| \sin\Theta(\hat{U}_k^{(p)}, U_k)\right\|^2 \geq c_1, \quad k = 1, 2, 3,$$

(Recovery of $\mathcal{X}$)

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

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 U_1, U_2, U_3, and $\mathcal{X}$.

- (Weak SNR) $\lambda/\sigma < cp^{1/2}$,
  $\rightarrow$ no algorithm can stably recover U_1, U_2, U_3, or $\mathcal{X}$.

- (Moderate SNR) $p^{1/2} \ll \lambda/\sigma \ll p^{3/4}$,
  - non-convex MLE stably recovers U_1, U_2, U_3, and $\mathcal{X}$;
  - 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.

References

- Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and Computational Limits. *IEEE Transactions on Information Theory*, to appear.
- Cai, T. and Zhang, A. (2018). Rate-Optimal Perturbation Bounds for Singular Subspaces with Applications to High-Dimensional Statistics. *Annals of Statistics*, 46, 60-89.
- Zhang, A. and Han, R. (2017+). Optimal Denoising and Singular Value Decomposition for Sparse High-dimensional High-order Data. *submitted*.