Beyond matrices: tensor decompositions and applications

Miaoyan Wang

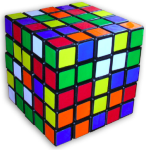
Departments of Statistics and EECS

University of California, Berkeley

February 12, 2017

Department of Statistics and Data Science

Carnegie Mellon University



My research

Statistical machine learning:

- structured tensor decomposition, latent factor models

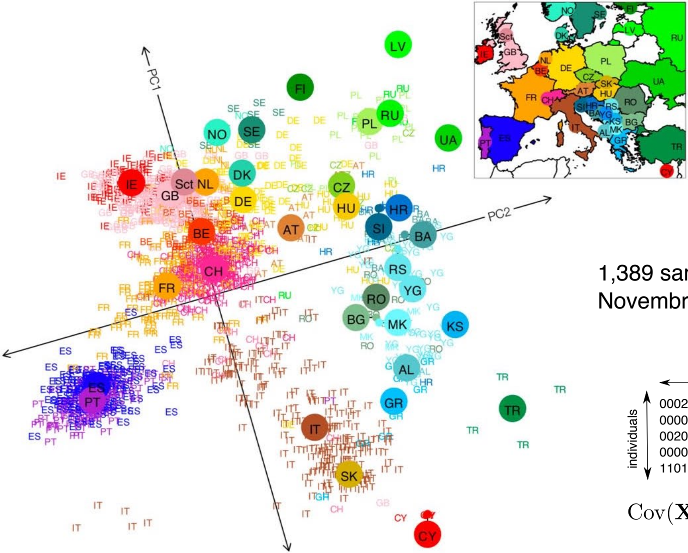
Genetics and genomics:

- gene expression analyses, genetic association studies

Computational foundations of data science:

- algorithm development for big data analytics

A successful story: PCA of Europeans



1,389 samples, ~ 200k SNPs
Novembre et al. (2008)

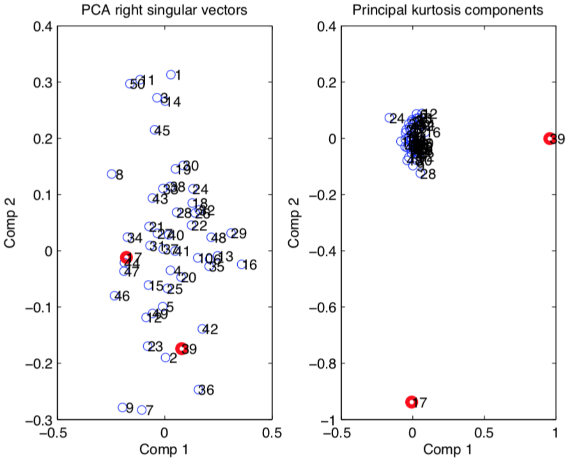
SNPs

individuals

```
00020110000001111110000000...
00001100000001201100000000...
00200111012001010011011 1...
00000000011121010010111 0...
1101101110111101200010 01...
```

$$\mathrm{Cov}(\mathbf{X}) = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

Matrix methods are powerful, however...



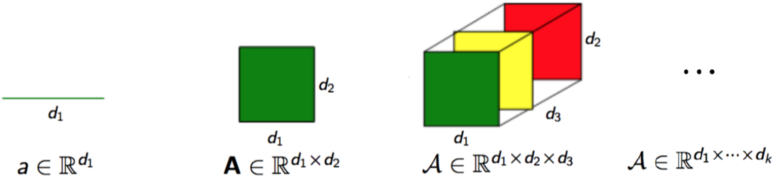
All Gaussian except points 17 and 39.

left: matrix PCA; right: principal components of kurtosis.

Figure credit: Jason Morton and Lek-Heng Lim (2009/2015).

What is a tensor?

- Tensors are generalizations of vectors and matrices:



- An order-k tensor $\mathcal{A} = [\![a_{i_1 \dots i_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ is a hypermatrix with dimensions $(d_1, \dots, d_k)$ and entries $a_{i_1 \dots i_k} \in \mathbb{R}$.
- This talk will focus on tensor of order 3 or greater, also known as higher-order tensors.

Tensors in statistical modeling

"Tensors are the new matrices" that tie together a wide range of areas:

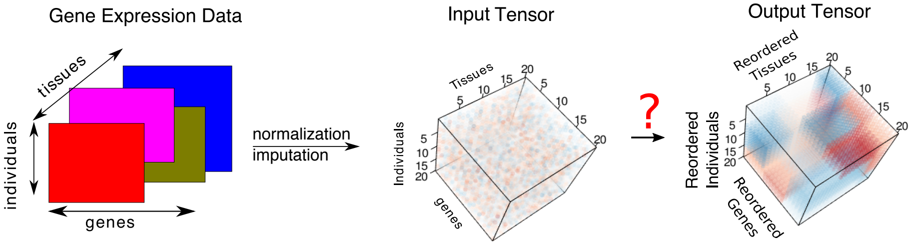
- Longitudinal social network data $\{\boldsymbol{Y}_t : t = 1, ..., n\}$
- Spatio-temporal transcriptome data
- Joint probability table of a set of variables $\mathbb{P}(X_1, X_2, X_3)$
- Higher-order moments in single topic models
- Markov models for the phylogenetic tree $K_{1,3}$

Yuan 2017, Dunson 2016, P. Hoff 2015, Montanari-Richard 2014

Anandkumar et al 2014, Mossel et al 2004, P. McCullagh 1987

Tensors in genomics

- Many biomedical datasets come naturally in a multiway form.
- Multi-tissue multi-individual gene expression measures could be organized as a multiarray dataset $\mathcal{A} = [\![a_{git}]\!] \in \mathbb{R}^{n_G \times n_I \times n_T}$.



Multi-way Clustering

To identify subsets of genes that are similarly expressed within subsets of individuals and tissues, we seek local blocks in the expression tensor.

Talk outline

Prohibitive Computational Complexity

Most higher-order tensor problems are NP-hard [Hillar & Lim, 2013].

Topics I will address:

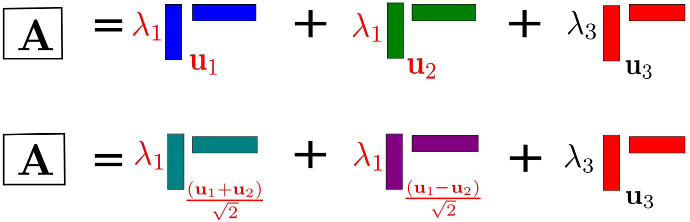
- Tensor decomposition method
- Theoretical results on tensor spectral norm
- Multi-way clustering of gene expression data

Review of matrix eigendecomposition

Matrix perturbation theorem (Davis–Kahan 1970)

Let A and E be symmetric matrices, and $\widetilde{A} = A + E$. Let $\mathbf{u}_i$, $\hat{\mathbf{u}}_i$ denote the ith eigenvectors of A and $\widetilde{A}$, respectively. Then

$$\sin \Theta(\mathbf{u}_i, \hat{\mathbf{u}}_i) \leq \frac{2\,\|E\|_2}{\min_{j \neq i} |\lambda_j - \lambda_i|}.$$

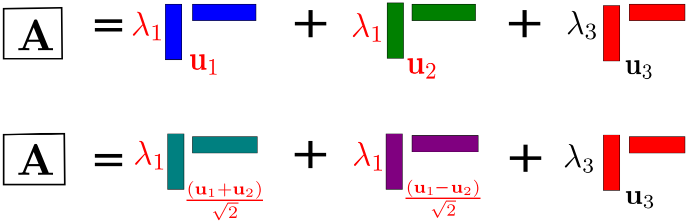


Review of matrix eigendecomposition

Matrix perturbation theorem (Davis–Kahan 1970)

Let A and E be symmetric matrices, and $\widetilde{A} = A + E$. Let $\mathbf{u}_i$, $\hat{\mathbf{u}}_i$ denote the ith eigenvectors of A and $\widetilde{A}$, respectively. Then

$$\sin \Theta(\mathbf{u}_i, \hat{\mathbf{u}}_i) \leq \frac{2\,\|E\|_2}{\min_{j\neq i} |\lambda_j - \lambda_i|}.$$



- Does there exist a tensor analogue of matrix eigendecomposition? How about perturbation analysis?

Symmetric tensors

Definition (Symmetric tensors)

A tensor $\mathcal{A} = [\![a_{i_1 \dots i_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ is called symmetric if $d_1 = \cdots = d_k$ and

$$a_{i_1 i_2 \dots i_k} = a_{\sigma(i_1)\sigma(i_2)\dots\sigma(i_k)},$$

for all permutations σ of $[k]$.

- By the spectral theorem, every symmetric matrix A admits an eigendecomposition,

$$A = \lambda_1 \mathbf{u}_1^{\otimes 2} + \lambda_2 \mathbf{u}_2^{\otimes 2} + \cdots + \lambda_r \mathbf{u}_r^{\otimes 2}.$$

- Does not hold for general symmetric tensors.

example

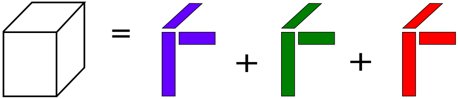
SOD tensors

- A tensor $\mathcal{A}$ is called symmetric and orthogonally decomposable (SOD) if

$$\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i^{\otimes k},$$

where $\{\mathbf{u}_i\}$ are orthonormal vectors in $\mathbb{R}^d$ and $\{\lambda_i\}$ are non-zero scalars.

- For example, $k = 3$ and $r = 3$:



$$\mathcal{A} = \lambda_1 \mathbf{u}_1^{\otimes 3} + \lambda_2 \mathbf{u}_2^{\otimes 3} + \lambda_3 \mathbf{u}_3^{\otimes 3}$$

- Kruskal's theorem implies that $\{\mathbf{u}_i\}$ is unique even in the case of degenerate $\lambda_i s$.
- Eigen-components of a 3rd cumulant tensor are closely related to parameter estimation in latent variable models [Anandkumar et al 2014].

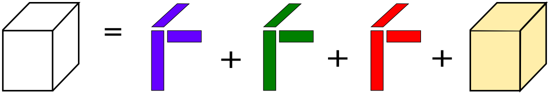
Tensor decomposition

- Nearly SOD tensors:

$$\tilde{\mathcal{A}} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i^{\otimes k} + \mathcal{E},$$

where $\mathcal{E} \in \mathbb{R}^{d\times\cdots\times d}$ is a symmetric but otherwise arbitrary tensor with $\|\mathcal{E}\|_2 \leq \varepsilon$.

- For example, $k = 3$ and $r = 3$:



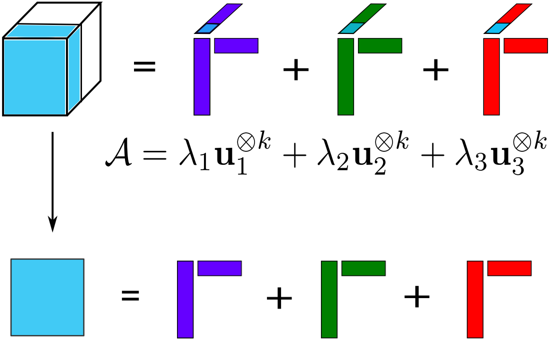
$$\tilde{\mathcal{A}} = \lambda_1 \mathbf{u}_1^{\otimes 3} + \lambda_2 \mathbf{u}_2^{\otimes 3} + \lambda_3 \mathbf{u}_3^{\otimes 3} + \mathcal{E}$$

Key question

Can we recover the vectors $\{\mathbf{u}_i\}$ from the noisy observation $\tilde{\mathcal{A}}$?

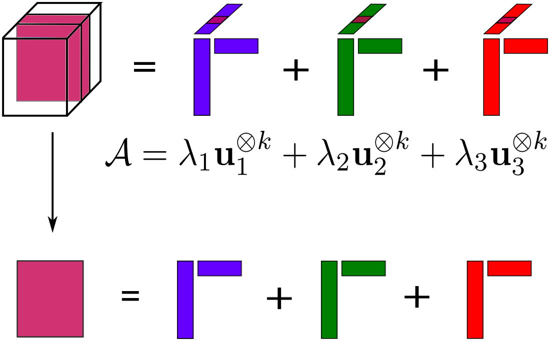
Decomposition of SOD tensors: noiseless case

- The structure of $\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i^{\otimes k}$ implies a common eigenspace for all matrix slices.



Decomposition of SOD tensors: noiseless case

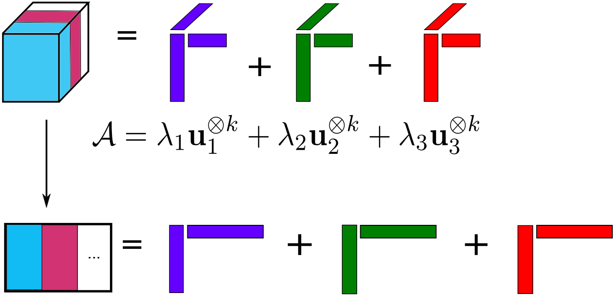
- The structure of $\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i^{\otimes k}$ implies a common eigenspace for all matrix slices.



$$\mathcal{A} = \lambda_1 \mathbf{u}_1^{\otimes k} + \lambda_2 \mathbf{u}_2^{\otimes k} + \lambda_3 \mathbf{u}_3^{\otimes k}$$

Decomposition of SOD tensors: noiseless case

- The structure of $\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i^{\otimes k}$ implies the one-mode unfolding $\mathcal{A}_{(1)(2\ldots k)} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \operatorname{Vec}(\mathbf{u}_i^{\otimes k-1})^T$:

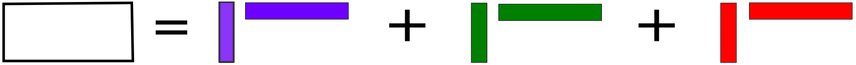


$$\mathcal{A} = \lambda_1 \mathbf{u}_1^{\otimes k} + \lambda_2 \mathbf{u}_2^{\otimes k} + \lambda_3 \mathbf{u}_3^{\otimes k}$$

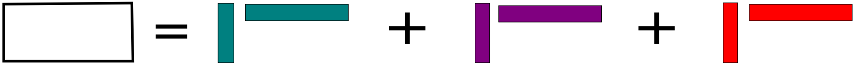
$$\mathcal{A}_{(1)(2\cdots k)} = \lambda_1 \mathbf{u}_1 \mathrm{Vec}(\mathbf{u}_1^{\otimes k-1})^T + \lambda_2 \mathbf{u}_2 \mathrm{Vec}(\mathbf{u}_2^{\otimes k-1})^T + \lambda_3 \mathbf{u}_3 \mathrm{Vec}(\mathbf{u}_3^{\otimes k-1})^T$$

- Is it possible to recover $\{\mathbf{u}_i\}_{i\in[r]}$ using the left singular vectors of the 1-mode unfolding, $\mathcal{A}_{(1)(2\ldots k)}$?

Matrix vs. tensor decompositions

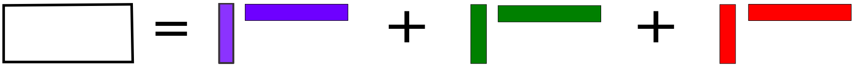


$$\mathcal{A}_{(1)(2\dots k)} = \lambda_1 \mathbf{u}_1 \mathrm{Vec}(\mathbf{u}_1^{\otimes k-1})^T + \lambda_1 \mathbf{u}_2 \mathrm{Vec}(\mathbf{u}_2^{\otimes k-1})^T + \lambda_3 \mathbf{u}_3 \mathrm{Vec}(\mathbf{u}_3^{\otimes k-1})^T$$

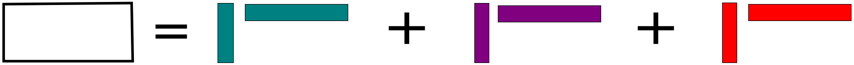


$$\mathcal{A}_{(1)(2\dots k)} = \lambda_1 \frac{\mathbf{u}_1 + \mathbf{u}_2}{\sqrt{2}} \mathbf{a}^T + \lambda_1 \frac{\mathbf{u}_1 - \mathbf{u}_2}{\sqrt{2}} \mathbf{b}^T + \lambda_3 \mathbf{u}_3 \mathrm{Vec}(\mathbf{u}_3^{\otimes k-1})^T$$

Matrix vs. tensor decompositions



$$\mathcal{A}_{(1)(2\ldots k)} = \lambda_1 \mathbf{u}_1 \mathrm{Vec}(\mathbf{u}_1^{\otimes k-1})^T + \lambda_1 \mathbf{u}_2 \mathrm{Vec}(\mathbf{u}_2^{\otimes k-1})^T + \lambda_3 \mathbf{u}_3 \mathrm{Vec}(\mathbf{u}_3^{\otimes k-1})^T$$



$$\mathcal{A}_{(1)(2\ldots k)} = \lambda_1 \frac{\mathbf{u}_1 + \mathbf{u}_2}{\sqrt{2}} \mathbf{a}^T + \lambda_1 \frac{\mathbf{u}_1 - \mathbf{u}_2}{\sqrt{2}} \mathbf{b}^T + \lambda_3 \mathbf{u}_3 \mathrm{Vec}(\mathbf{u}_3^{\otimes k-1})^T$$

Caveats:

- A rank $r > 1$ matrix can be decomposed in multiple ways as a sum of order-product terms in the case of degenerate λ_is.
- Kruskal's theorem guarantees that the set of vectors $\{\mathbf{u}_i\}_{i\in[r]}$ of an SOD tensor is unique up to signs even when some λ_is are degenerate.

Two-mode HOSVD via rank-1 matrix pursuit

Key idea: Instead of $\mathcal{A}_{(1)(2\ldots k)}$, we consider the two-mode unfolding $\mathcal{A}_{(12)(3\ldots k)}$.

Two-mode unfolding

$\mathcal{A}_{(12)(3\ldots k)}$ is a $d^2 \times d^{k-2}$ matrix obtained by grouping the first 2 indices of $\mathcal{A}$ as the row index and the remaining $(k-2)$ indices as the column index.

Two-mode HOSVD via rank-1 matrix pursuit

Key idea: Instead of $\mathcal{A}_{(1)(2\ldots k)}$, we consider the two-mode unfolding $\mathcal{A}_{(12)(3\ldots k)}$.

Two-mode unfolding

$\mathcal{A}_{(12)(3\ldots k)}$ is a $d^2 \times d^{k-2}$ matrix obtained by grouping the first 2 indices of $\mathcal{A}$ as the row index and the remaining $(k-2)$ indices as the column index.

$$\mathcal{A}_{(12)(3\ldots k)} = \lambda_1 \mathrm{Vec}(\mathbf{u}_1^{\otimes 2})\mathrm{Vec}(\mathbf{u}_1^{\otimes k-2})^T + \lambda_1 \mathrm{Vec}(\mathbf{u}_2^{\otimes 2})\mathrm{Vec}(\mathbf{u}_2^{\otimes k-2})^T + \lambda_3 \mathrm{Vec}(\mathbf{u}_3^{\otimes 2})\mathrm{Vec}(\mathbf{u}_3^{\otimes k-2})^T$$

rank-1 matrices

$$\mathcal{A}_{(12)(3\ldots k)} = \lambda_1 \mathrm{Vec}\left(\frac{\mathbf{u}_1^{\otimes 2}+\mathbf{u}_2^{\otimes 2}}{\sqrt{2}}\right)\mathbf{c}^T + \lambda_1 \mathrm{Vec}\left(\frac{\mathbf{u}_1^{\otimes 2}-\mathbf{u}_2^{\otimes 2}}{\sqrt{2}}\right)\mathbf{d}^T + \lambda_3 \mathrm{Vec}(\mathbf{u}_3^{\otimes 2})\mathrm{Vec}(\mathbf{u}_3^{\otimes k-2})^T$$

Our results

Given an order-k nearly SOD tensor $\widetilde{\mathcal{A}} \in \mathbb{R}^{d\times\cdots\times d}$,

$$\widetilde{\mathcal{A}} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i^{\otimes k} + \mathcal{E}, \quad \text{where} \quad \|\mathcal{E}\|_2 \leq \varepsilon.$$

Goal: recover $\{\mathbf{u}_i\}$ from $\widetilde{\mathcal{A}}$.

- Noiseless case:
  Every rank-1 matrix in the left singular space of $A_{(12)(3\ldots k)}$ is (up to a scalar) the Kronecker square of some robust tensor eigenvector $\mathbf{u}_i$.

- Noisy case:
  If $\varepsilon/|\lambda|_{\min} \lesssim d^{-(k-2)/2}$, we can recover $\{\mathbf{u}_i\}$ up to error $O(\varepsilon)$ in polynormial time.

Wang, M. and Song, Y.S., Journal of Machine Learning Research W&CP, Vol. 54 (2017) 614-622.

details

Comparison of tensor decomposition algorithms

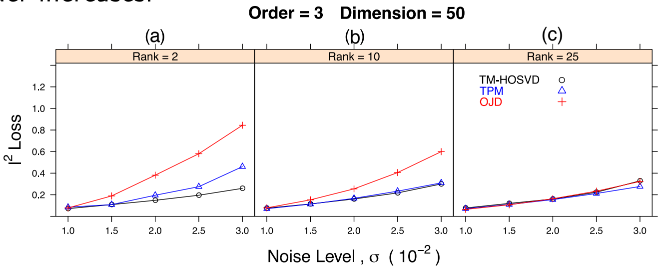
- The error bound in tensor decomposition does not depend on the eigen-value gap $\Rightarrow$ more stable than matrix decomposition.

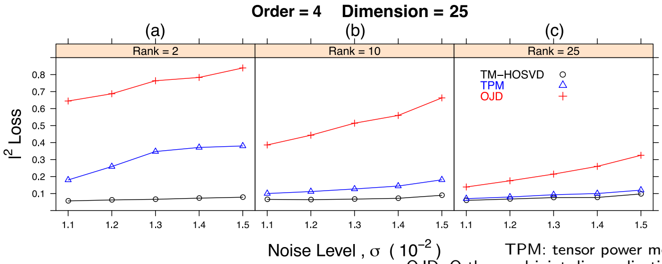
Method	Noise threshold ($\varepsilon/\lvert\lambda\rvert_{\min} \leq$)	Recovery accuracy ($\lVert\hat{\mathbf{u}}_i - \mathbf{u}_i\rVert_2 \leq$)
Power iteration (Anandkumar et al, 2014)	$O(d^{-1})$ for order 3	$\frac{8\varepsilon}{\lambda_i}$
Joint diagonalization (Kuleshov et al, 2015)	–	$\frac{2\varepsilon\sqrt{\lVert\boldsymbol{\lambda}\rVert_1 \lambda_{\max}}}{\lambda_i^2} + o(\varepsilon)$
Our method (W. and Song 2017b)	$\boldsymbol{O(d^{-1/2})}$ **for order 3** $O(d^{-(k-2)/2})$ for order k	$\boldsymbol{\frac{2\varepsilon}{\lambda_i}} + o(\varepsilon)$

Wang, M. and Song, Y.S., Journal of Machine Learning Research W&CP, Vol. 54 (2017) 614-622.

Numerical experiments

Our method achieves a higher estimation accuracy and performs favorably as the order increases.





TPM: tensor power method, JMLR 2014

OJD: Orthogonal joint diagonalization, AISTATS 2015

Example

A symmetric tensor but not orthogonally decomposable:

$$\mathcal{A}(:,:,1) = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix},$$

$$\mathcal{A}(:,:,2) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

back

Orthogonality

Definition (π-orthogonally decomposable)

A tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ is called π-OD with partition $\pi = \{B_1, \ldots, B_\ell\}$ if it admits the decomposition

$$\mathcal{A} = \lambda_1 \underbrace{\boldsymbol{a}_1^{(1)} \otimes \boldsymbol{a}_2^{(1)}}_{B_1} \otimes \cdots \underbrace{\otimes \boldsymbol{a}_k^{(1)}}_{B_\ell} + \cdots + \lambda_r \underbrace{\boldsymbol{a}_1^{(r)} \otimes \boldsymbol{a}_2^{(r)}}_{B_1} \otimes \cdots \underbrace{\otimes \boldsymbol{a}_k^{(r)}}_{B_\ell},$$

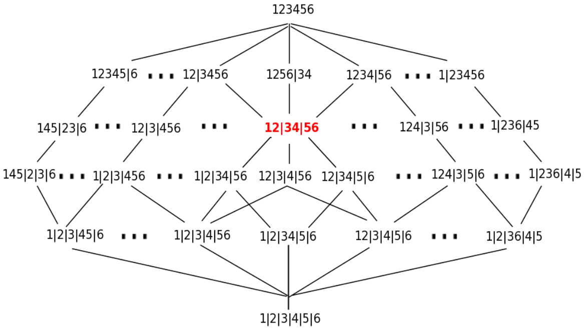
where the set of vectors $\{\boldsymbol{a}_i^{(n)}\}$ satisfies

$$\langle \otimes_{i \in B} \boldsymbol{a}_i^{(n)}, \ \otimes_{i \in B} \boldsymbol{a}_i^{(m)} \rangle = \delta_{nm},$$

for all $B \in \pi$ and all $n, m \in [r]$.

π-OD tensors and norm-preserving cones

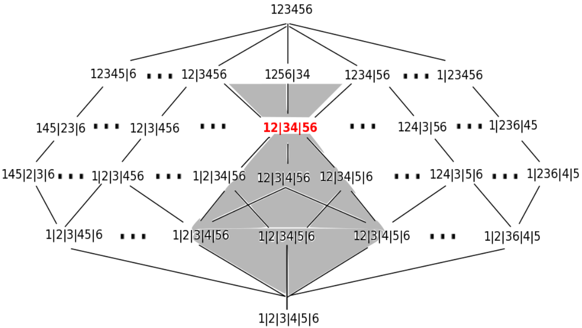
Suppose $\mathcal{A}$ is a π-OD tensor and define $c \stackrel{\text{def}}{=} \|\mathcal{A}\|_2$.



π-OD tensors and norm-preserving cones

Suppose $\mathcal{A}$ is a π-OD tensor and define $c \stackrel{\text{def}}{=} \|\mathcal{A}\|_2$.

- $\|\text{Unfold}_\tau(\mathcal{A})\|_2 = c$ for all $\tau \in C_\pi \stackrel{\text{def}}{=} \{\tau : \tau \geq \pi\} \cup \{\tau : \tau \leq \pi\} \backslash \mathbf{1}_{[k]}$



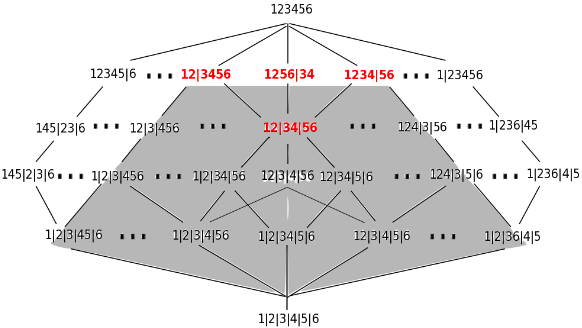
π-OD tensors and norm-preserving cones

Suppose $\mathcal{A}$ is a π-OD tensor and define $c \stackrel{\text{def}}{=} \|\mathcal{A}\|_2$.

- $\|\mathrm{Unfold}_\tau(\mathcal{A})\|_2 = c$ for all $\tau \in C_\pi \stackrel{\text{def}}{=} \{\tau\colon \tau \geq \pi\} \cup \{\tau\colon \tau \leq \pi\} \backslash \mathbf{1}_{[k]}$
- π-OD implies π'-OD for all $\pi' \geq \pi$. $\Rightarrow$

$$\|\mathrm{Unfold}_\tau(\mathcal{A})\|_2 = c \text{ for all } \tau \in C_{\pi_1} \cup \cdots \cup C_{\pi_s},$$

where $\pi_1, \ldots, \pi_s$ are matricizations obtained by merging blocks of π.



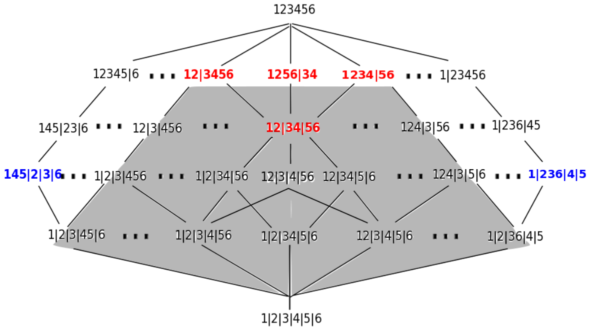
π-OD tensors and norm-preserving cones

Suppose $\mathcal{A}$ is a π-OD tensor and define $c \stackrel{\text{def}}{=} \|\mathcal{A}\|_2$.

- $\|\text{Unfold}_\tau(\mathcal{A})\|_2 = c$ for all $\tau \in C_\pi \stackrel{\text{def}}{=} \{\tau \colon \tau \geq \pi\} \cup \{\tau \colon \tau \leq \pi\} \backslash \mathbf{1}_{[k]}$
- π-OD implies π'-OD for all $\pi' \geq \pi$. $\Rightarrow$

$$\|\text{Unfold}_\tau(\mathcal{A})\|_2 = c \text{ for all } \tau \in C_{\pi_1} \cup \cdots \cup C_{\pi_s},$$

  where $\pi_1, \ldots, \pi_s$ are matricizations obtained by merging blocks of π.
- Can obtain sharper bounds for spectral norm landscape.



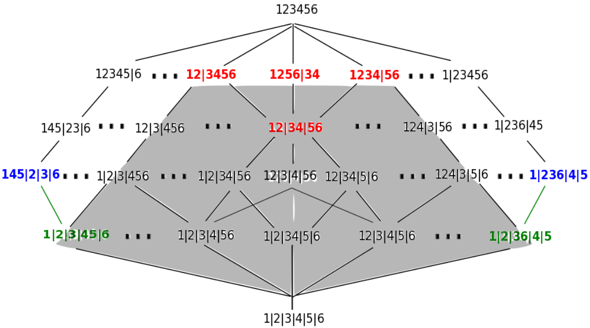
π-OD tensors and norm-preserving cones

Suppose $\mathcal{A}$ is a π-OD tensor and define $c \overset{\text{def}}{=} \|\mathcal{A}\|_2$.

- $\|\text{Unfold}_\tau(\mathcal{A})\|_2 = c$ for all $\tau \in C_\pi \overset{\text{def}}{=} \{\tau : \tau \geq \pi\} \cup \{\tau : \tau \leq \pi\} \backslash \mathbf{1}_{[k]}$
- π-OD implies π'-OD for all $\pi' \geq \pi$. $\Rightarrow$

$$\|\text{Unfold}_\tau(\mathcal{A})\|_2 = c \text{ for all } \tau \in C_{\pi_1} \cup \cdots \cup C_{\pi_s},$$

  where $\pi_1, \ldots, \pi_s$ are matricizations obtained by merging blocks of π.
- Can obtain sharper bounds for spectral norm landscape.



Unfolding of an order-k tensor

- **General Unfolding**. The set of all possible unfoldings of an order-k tensor is in one-to-one correspondence with the set $\mathcal{P}_{[k]}$ of all partitions of $[k] = \{1, \ldots, k\}$.
- For $\pi = \{B_1, \ldots, B_\ell\} \in \mathcal{P}_{[k]}$, $\mathrm{Unfold}_\pi(\mathcal{A})$ is obtained by combining the modes in each block B_n into a single mode.

Example. An order-4 tensor $\mathcal{A} = [\![a_{i_1 i_2 i_3 i_4}]\!] \in \mathbb{R}^{2\times 2\times 2\times 2}$ with $a_{i_1 i_2 i_3 i_4} = \begin{cases} 1 & \text{if } i_1 = i_2 = i_3 = i_4 \\ 0 & \text{otherwise} \end{cases}$ can be matricized into

- 2×2^3 matrix: $\mathrm{Unfold}_{[1|234]}(\mathcal{A}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$.
- $2^2 \times 2^2$ matrix: $\mathrm{Unfold}_{[12|34]}(\mathcal{A}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

Definition (Inner product)

For any two tensors $\mathcal{A} = [\![a_{i_1 \dots i_k}]\!]$, $\mathcal{B} = [\![b_{i_1 \dots i_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ of identical order and dimensions, their inner product is defined as

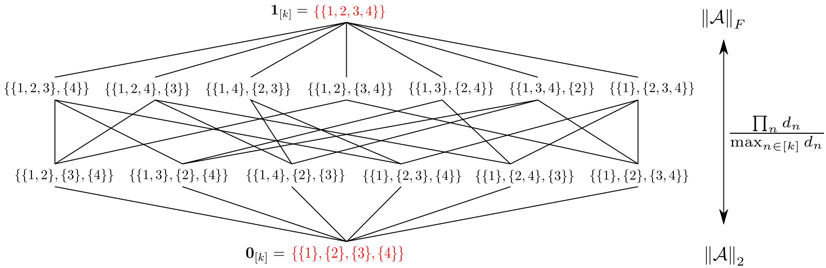
$$\langle \mathcal{A},\ \mathcal{B} \rangle = \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} b_{i_1 \dots i_k}.$$

The tensor Frobenius norm of $\mathcal{A}$ is defined as $\|\mathcal{A}\|_F = \sqrt{\langle \mathcal{A},\ \mathcal{A} \rangle}$.

Frobenius norm vs. spectral norm

$$\|\mathcal{A}\|_F = \max_{\pi \in \mathcal{P}_{[k]}} \|\mathrm{Unfold}_\pi(\mathcal{A})\|_2 \,, \quad \|\mathcal{A}\|_2 = \min_{\pi \in \mathcal{P}_{[k]}} \|\mathrm{Unfold}_\pi(\mathcal{A})\|_2 \,,$$

$$\|\mathcal{A}\|_F \leq \left[\frac{\prod_n d_n}{\max_{n\in[k]} d_n} \right]^{1/2} \|\mathcal{A}\|_2 \,.$$



This bound improves over the recent result found by Friedland and Lim [Lemma 5.1, 2016], namely, $\|\mathcal{A}\|_F \leq \left(\prod_n d_n\right)^{1/2} \|\mathcal{A}\|_2$.

Norm inequalities between any two tensor unfoldings

Given $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$, we define the map $\dim_{\mathcal{A}} \colon \mathcal{P}_{[k]} \times \mathcal{P}_{[k]} \to \mathbb{N}_+$ as :

$$\dim_{\mathcal{A}}(\pi_1, \pi_2) = \prod_{B \in \pi_1} \left[\max_{B' \in \pi_2} \left(\prod_{n \in B \cap B'} d_n \right) \right], \quad \text{where } \pi_1, \pi_2 \in \mathcal{P}_{[k]}.$$

Theorem (p-norm inequalities)

Let $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ be an arbitrary order-k tensor, and π_1, π_2 any two partitions in $\mathcal{P}_{[k]}$. Define $\dim(\mathcal{A}) = \prod_{i=1}^{k} d_i$. Then,

(a) For any $1 \leq p \leq 2$,

$$\frac{[\dim A]^{-1/p}}{[\dim_{\mathcal{A}}(\pi_1, \pi_2)]^{-1/2}} \left\| \mathrm{Unfold}_{\pi_1}(\mathcal{A}) \right\|_p \leq \left\| \mathrm{Unfold}_{\pi_2}(\mathcal{A}) \right\|_p \leq \frac{[\dim(\mathcal{A})]^{1/p}}{[\dim_{\mathcal{A}}(\pi_2, \pi_1)]^{1/2}} \left\| \mathrm{Unfold}_{\pi_1}(\mathcal{A}) \right\|_p .$$

(b) For any $2 \leq p \leq \infty$,

$$\frac{[\dim(\mathcal{A})]^{\frac{1}{p}-1}}{[\dim_{\mathcal{A}}(\pi_1, \pi_2)]^{-1/2}} \left\| \mathrm{Unfold}_{\pi_1}(\mathcal{A}) \right\|_p \leq \left\| \mathrm{Unfold}_{\pi_2}(\mathcal{A}) \right\|_p \leq \frac{[\dim(\mathcal{A})]^{1-\frac{1}{p}}}{[\dim_{\mathcal{A}}(\pi_2, \pi_1)]^{1/2}} \left\| \mathrm{Unfold}_{\pi_1}(\mathcal{A}) \right\|_p .$$

Two-mode HOSVD algorithm for tensors with noise

Rank-1 matrices in $\mathcal{LS}^{(r)}$ are sufficient to find $\{\boldsymbol{u}_i\}$.

- Define the two-mode left singular space by

  $\mathcal{LS}^{(r)} \stackrel{\text{def}}{=} \text{Span}\{\mathbf{a}_i \in \mathbb{R}^{d^2} : \mathbf{a}_i \text{ is the } i\text{th left singular vector of } \widetilde{\mathcal{A}}_{(12)(3\ldots k)}\}.$

- Look for "nearly" rank-1 matrix $\widehat{\boldsymbol{M}}$ in the linear space $\mathcal{LS}^{(r)}$:

$$\underset{\boldsymbol{M}\in\mathbb{R}^{d\times d}}{\text{maximize}} \, \|\boldsymbol{M}\|_2 \, ,$$

$$\text{subject to } \boldsymbol{M} \in \mathcal{LS}^{(r)} \text{ and } \|\boldsymbol{M}\|_F = 1.$$

  Justification of the optimization: $\|\boldsymbol{M}\|_2 \leq \|\boldsymbol{M}\|_F \leq \sqrt{\text{rank } \boldsymbol{M}}\, \|\boldsymbol{M}\|_2$.

- Apply eigendecomposition on the matrix $\widehat{\boldsymbol{M}}$ to recover $\mathbf{u}_i$.

back

Exact Recovery for SOD Tensors in the noiseless case

Optimization to recover the desired factors $\{\mathbf{u}_i\}$ of $\mathcal{A}$:

$$\begin{aligned} &\underset{\boldsymbol{M}\in\mathbb{R}^{d\times d}}{\text{maximize}} \|\boldsymbol{M}\|_2\,, \\ &\text{subject to } \boldsymbol{M} \in \mathcal{LS}_0 \text{ and } \|\boldsymbol{M}\|_F = 1. \end{aligned} \tag{1}$$

Theorem (W. and Song, 2017)

The optimization problem (**??**) *has exactly r pairs of local maximizers $\{\pm \boldsymbol{M}_i^* : i \in [r]\}$. Furthermore, they satisfy the following three properties:*

1. $\|\boldsymbol{M}_i^*\|_2 = 1$ *for all* $i \in [r]$.
2. $\left|\langle \mathrm{Vec}(\boldsymbol{M}_i^*),\ \mathrm{Vec}(\boldsymbol{M}_j^*)\rangle\right| = \delta_{ij}$ *for all* $i, j \in [r]$, *where* $\langle \cdot,\ \cdot \rangle$ *denotes the inner product.*
3. *There exists a permutation π on $[r]$ such that* $\boldsymbol{M}_i^* = \pm \boldsymbol{u}_{\pi(i)}^{\otimes 2}$ *for all* $i \in [r]$.

back

Two-mode HOSVD algorithm for tensors with noise

Optimization to recover the desired factors $\{\mathbf{u}_i\}$ of $\widetilde{\mathcal{T}}$:

$$\underset{\boldsymbol{M}\in\mathbb{R}^{d\times d}}{\text{maximize}} \|\boldsymbol{M}\|_2 \,,$$

$$\text{subject to } \boldsymbol{M} \in \mathcal{LS}_r \text{ and } \|\boldsymbol{M}\|_F = 1.$$

Algorithm 1 Two-mode HOSVD

Input: Noisy tensor $\widetilde{\mathcal{T}}$ where $\widetilde{\mathcal{T}} = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i^{\otimes k} + \mathcal{E}$, number of factors r.

Output: r pairs of estimators $(\widehat{\boldsymbol{u}}_i, \widehat{\lambda}_i)$.

Two-Mode HOSVD {
Reshape the tensor $\widetilde{\mathcal{T}}$ into a d^2-by-d^{k-2} matrix $\widetilde{\mathcal{T}}_{(12)(3\ldots k)}$;
Find the top r left singular vectors of $\widetilde{\mathcal{T}}_{(12)(3\ldots k)}$, denoted $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r\}$;

Initialize $\mathcal{LS}^{(r)} = \mathrm{Span}\{\boldsymbol{a}_i : i \in [r]\}$;
for i=1 to r **do**

Nearly Rank-1 Matrix {
Solve $\widehat{\boldsymbol{M}}_i = \underset{\boldsymbol{M}\in\mathcal{LS}^{(r)}, \|\boldsymbol{M}\|_F=1}{\arg\max} \|\boldsymbol{M}\|_\sigma$ and $\widehat{\boldsymbol{u}}_i = \underset{\boldsymbol{u}\in\mathbf{S}^{d-1}}{\arg\max} |\boldsymbol{u}^T \widehat{\boldsymbol{M}}_i \boldsymbol{u}|$;

Post-Processing {
Update $\widehat{\boldsymbol{M}}_i \leftarrow \widetilde{\mathcal{T}}_{(1)(2)(3\ldots k)}(\boldsymbol{I}, \boldsymbol{I}, \mathrm{Vec}(\widehat{\boldsymbol{u}}_i^{\otimes(k-2)}))$ and $\widehat{\boldsymbol{u}}_i \leftarrow \underset{\boldsymbol{u}\in\mathbf{S}^{d-1}}{\arg\max} |\boldsymbol{u}^T \widehat{\boldsymbol{M}}_i \boldsymbol{u}|$;
Return $(\widehat{\boldsymbol{u}}_i, \widehat{\lambda}_i) \leftarrow (\widehat{\boldsymbol{u}}_i, \widehat{\mathcal{T}}(\widehat{\boldsymbol{u}}_i, \ldots, \widehat{\boldsymbol{u}}_i))$;

Deflation {
Set $\mathcal{LS}^{(r)} \leftarrow \mathcal{LS}^{(r)} \cap \left[\mathrm{Vec}(\widehat{\boldsymbol{u}}_i^{\otimes 2})\right]^{\perp}$;

end for

Theorem (W. and Song, 2017b)

Let $\widetilde{\mathcal{A}} = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i^{\otimes k} + \mathcal{E} \in \mathbb{R}^{d \times \cdots \times d}$, *where* $\{\boldsymbol{u}_i\}_{i \in [r]}$ *are orthonormal vectors,* $\lambda_i > 0$ *for all* $i \in [r]$, *and* $\|\mathcal{E}\|_2 \leq \varepsilon$. *Suppose* $\varepsilon \leq |\lambda|_{\min} / \left[c_0 d^{(k-2)/2}\right]$, *where* $c_0 > 0$ *is a sufficiently large constant that does not depend on* d. *Let* $\{(\widehat{\boldsymbol{u}}_i, \widehat{\lambda}_i)\}_{i \in [r]}$ *be the output of Algorithm 1 for inputs* $\widetilde{\mathcal{A}}$ *and* r. *Then, there exists a permutation* π *on* $[r]$ *such that for all* $i \in [r]$,

$$\mathrm{Loss}(\widehat{\boldsymbol{u}}_i, \boldsymbol{u}_{\pi(i)}) \leq \frac{2\varepsilon}{\lambda_{\pi(i)}} + o(\varepsilon), \qquad \mathrm{Loss}(\widehat{\lambda}_i, \lambda_{\pi(i)}) \leq 2\varepsilon + o(\varepsilon),$$

and

$$\left\| \widetilde{\mathcal{A}} - \sum_{i=1}^{r} \widehat{\lambda}_i \widehat{\boldsymbol{u}}_i^{\otimes k} \right\|_2 \leq C\varepsilon + o(\varepsilon),$$

where $C = C(k) > 0$ *is a constant that only depends on* k.

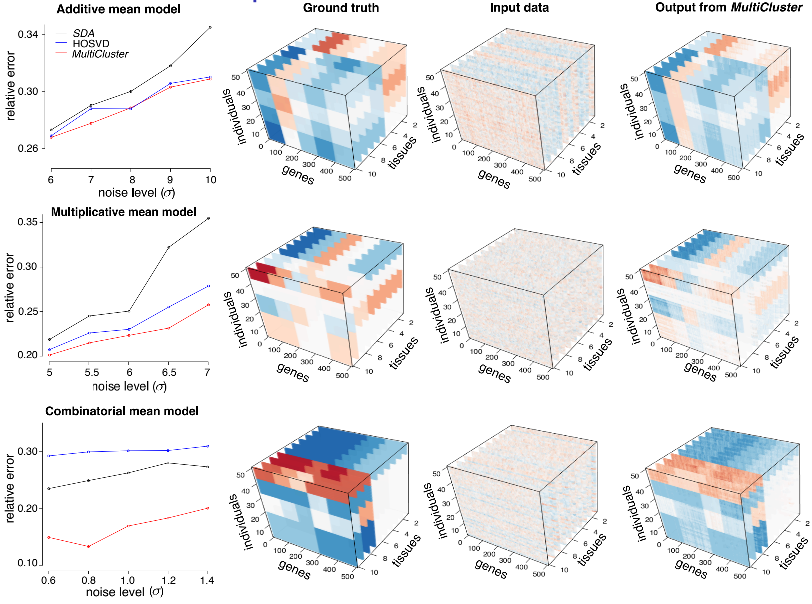
For two unit vectors $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^d$, define

$$\mathrm{Loss}(\mathbf{a}, \boldsymbol{b}) = \min\left(\|\mathbf{a} - \mathbf{b}\|_2, \|\mathbf{a} + \mathbf{b}\|_2\right).$$

If a, b are two scalars in $\mathbb{R}$, we define $\mathrm{Loss}(a, b) = \min(|a - b|, |a + b|)$.

back

Robustness to misspecified models



Run time comparison

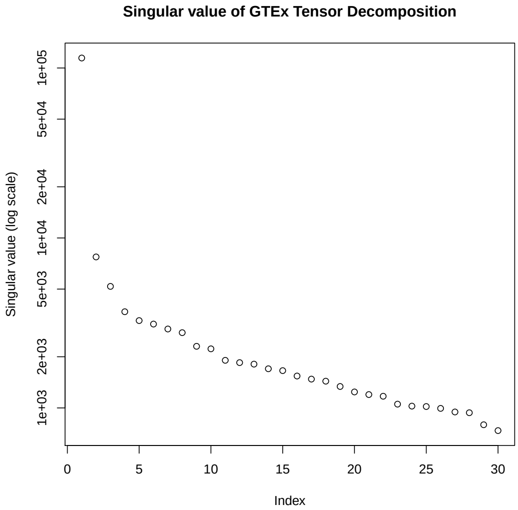
Complexity (for order-3 tensors):

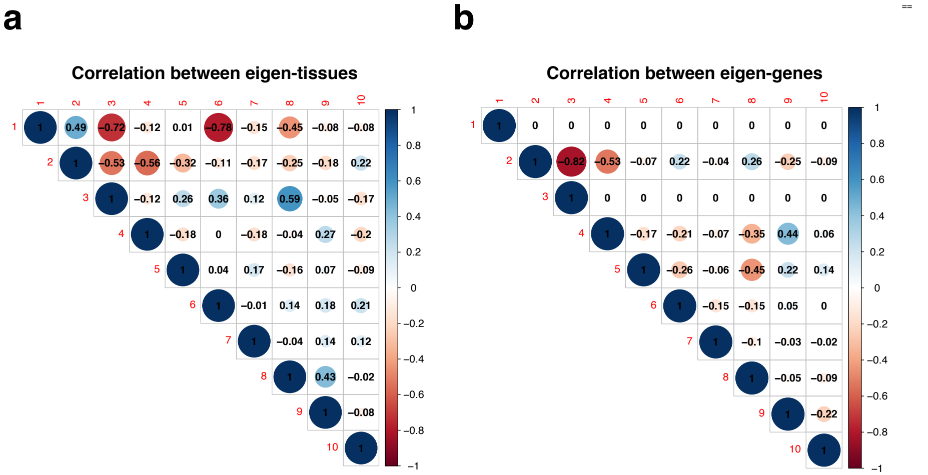
- TPM (Anandkumar et al., 2014): $O(d^3 M)$ per iteration, where M is the number of restarts.
- OJD (Kuleshov et al., 2015): $O(d^3 L)$ per iteration, where L is the number of projected matrices.
- Our method (W. and Song 2017b): $O(d^3)$ per iteration.

Simulation study: decompose $\mathcal{A} \in \mathbb{R}^{18000 \times 500 \times 40}$ into 10 components.

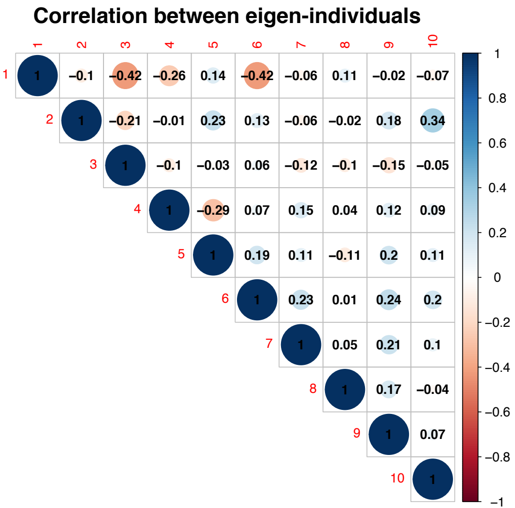
- SDA (Hore et al., 2016): 73,989 seconds ($\sim$ 20.1 hrs)
- HOSVD (Ombert et al., 2007): 5,849 seconds
- Our method (W. el al., 2017c): 6,047 seconds ($\sim$ 1.7 hrs)

Decay of singular values in GTEx




a
Correlation between eigen-tissues
b
Correlation between eigen-genes

c



Correlation between eigen-individuals
1
2
3
4
5
6
7
8
9
10
-0.1
-0.42
-0.26
0.14
-0.42
-0.06
0.11
-0.02
-0.07
-0.21
-0.01
0.23
0.13
-0.06
-0.02
0.18
0.34
-0.1
-0.03
0.06
-0.12
-0.1
-0.15
-0.05
-0.29
0.07
0.15
0.04
0.12
0.09
0.19
0.11
-0.11
0.2
0.11
0.23
0.01
0.24
0.2
0.05
0.21
0.1
0.17
-0.04
0.07
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Theorem

Let $\boldsymbol{A} \in \otimes^k \mathbb{R}^d$ *be an order-k dim-d random tensor with i.i.d. standard Gaussian entries, then*

$$d^{1/2} < \mathbb{E}\,\|\boldsymbol{A}\|_2 < kd^{1/2}.$$

Further, $\|\boldsymbol{A}\|_2$ *concentrates tightly around its expectation. Namely, for any* $s \geq 0$,

$$\mathbb{P}(\big|\,\|\boldsymbol{A}\|_2 - \mathbb{E}\,\|\boldsymbol{A}\|_2\,\big| \geq s) \leq 2e^{-s^2/2}.$$

With little modification, the above result can be generalized to order-k, dimensional-$(d_1, \ldots, d_k)$ tensors. Specifically, we have

$$\sqrt{d_{\max}} < \mathbb{E}\,\|\boldsymbol{A}\|_2 < \sum_{i=1}^{k} \sqrt{d_i}.$$

This implies, $\|\boldsymbol{A}\|_2 \asymp \mathcal{O}_p(\sqrt{d_{\max}})$ asymptotically for large d and fixed k.

Theorem (Non-Asymptotic Chain)

Let $\boldsymbol{A} \in \otimes^k \mathbb{R}^d$ be an order-k dim-d random tensor with i.i.d. standard Gaussian entries. Then for any $d \geq 4$ and $k \geq 2$,

$$\mathbb{E} \left\| Mat_1(\boldsymbol{A}) \right\|_2 > \mathbb{E} \left\| Mat_2(\boldsymbol{A}) \right\|_2 > \cdots > \mathbb{E} \left\| Mat_{\lfloor k/2 \rfloor}(\boldsymbol{A}) \right\|_2 .$$

Further, for any $1 \leq p \leq \lfloor k/2 \rfloor$,

$$d^{(k-p)/2} < \mathbb{E} \left\| Mat_p(\boldsymbol{A}) \right\| < d^{(k-p)/2} + d^{p/2}.$$

The following inequality chain holds almost surely as $d \to \infty$ at any fixed k:

$$\left\| \mathrm{Mat}_1(\boldsymbol{A}) \right\|_2 > \left\| \mathrm{Mat}_2(\boldsymbol{A}) \right\|_2 > \cdots > \left\| \mathrm{Mat}_{\lfloor k/2 \rfloor}(\boldsymbol{A}) \right\|_2 ,$$

Further, for any $1 \leq p \leq \lfloor k/2 \rfloor$,

$$\left\| \mathrm{Mat}_p(\boldsymbol{A}) \right\|_2 \to_{\text{a.s.}} (1 + \mathbf{1}_{\{p=k-p\}}) d^{(k-p)/2} \quad \text{as } d \to \infty.$$