Beyond matrices: Statistical learning with tensor data

Jiaxin Hu
Advisor: Prof. Miaoyan Wang

Department of Statistics
University of Wisconsin-Madison

Applied Algebra Seminar
Nov 2022

Outline

1. **Backgrounds**
2. **Overview of low-dimensional tensor methods**
3. **My research: tensor decomposition with multiple features**
4. **My research: degree-corrected tensor block model**

Outline

What is tensor?

- Tensors are generalizations of vectors and matrices.
- An order-K (real) tensor $\mathcal{X} = [\![x_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a hypermatrix with dimensions $(d_1, \ldots, d_K)$ and entries $x_{i_1,\ldots,i_K} \in \mathbb{R}$.

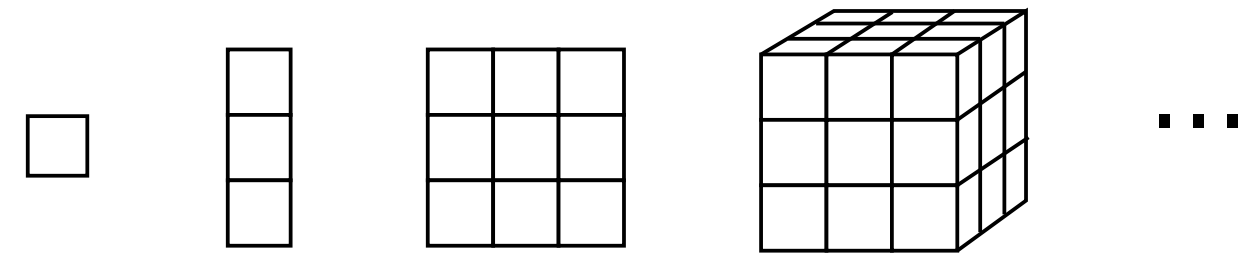


Figure 1: Order-k tensors.

Real-life tensors: social networks

- Higher-order connections among multiple individuals in social networks are represented as tensors.
- For example, the order-3 co-authorship tensor for p authors, $\mathcal{Y} \in \{0,1\}^{p \times p \times p}$, where

$$\mathcal{Y}(i,j,k) = \begin{cases} 1 & \text{authors } i,j,k \text{ co-author one paper;} \\ 0 & \text{otherwise.} \end{cases}$$

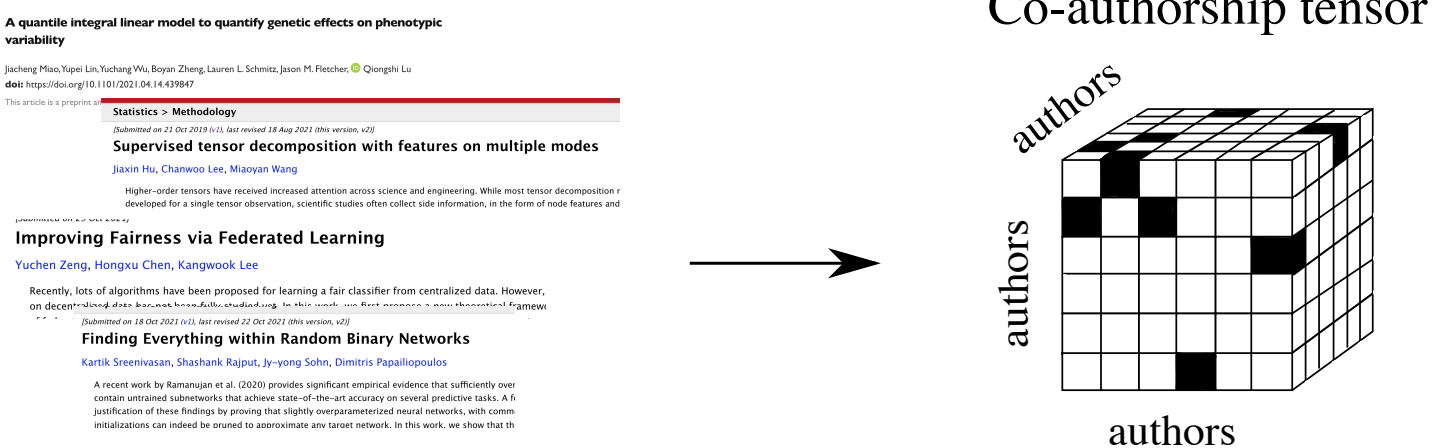


Figure 2: Order-3 co-authorship tensor, where black refers to the entry with value 1 and white refers to the entry with value 0.

Real-life tensors: biological studies

- Multi-tissue, multi-individual gene expression measures could be organized as an order-3 tensor $\mathcal{Y} \in \mathbb{R}^{n_{\text{gene}} \times n_{\text{individual}} \times n_{\text{tissue}}}$.
- Multi-modal brain image data could be represented by an order-3 tensor $\mathcal{Y} \in \mathbb{R}^{n_{\text{nodes}} \times n_{\text{nodes}} \times n_{\text{modality}}}$.

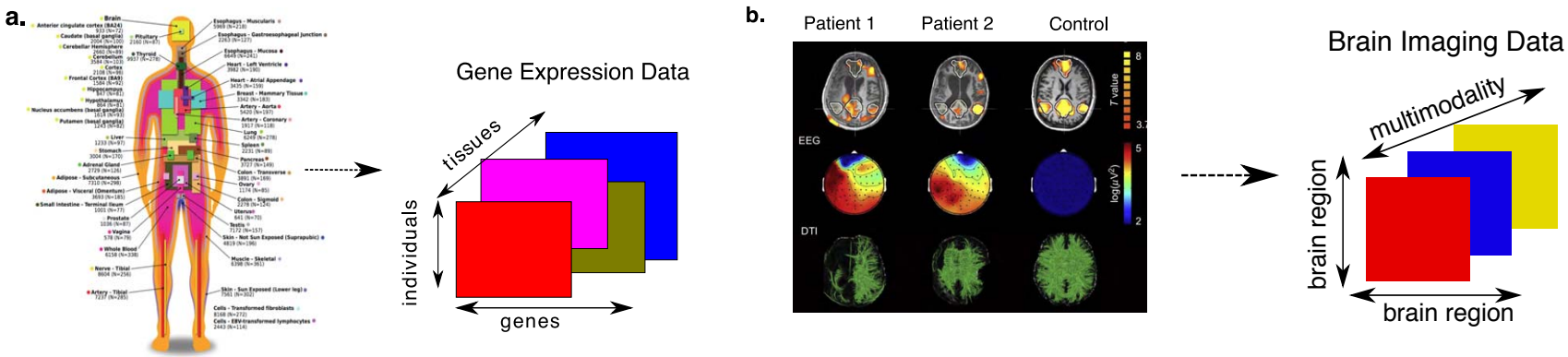


Figure 3: (a) Multi-tissue, multi-individual gene expression data. (b) Multi-modal brain image data.

Tensors applications

In science:

- Computational biology [Cartwright et al. 2009]
- Neuroimaging [Schultz and Seidel 2008]
- Quantum computing [Miyake and Wadati2002]
- Social networks [Ke et al. 2019]
- ...

In statistical modelling:

- Longitudinal social network data
- Spatio-temporal transcriptome or brain connection data
- Joint probability table and higher-order moments of a set of variables $X_1, X_2, X_3, ...$
- Markov models for the phylogenetic tree
- ...

Notations

Lowercase (e.g. c) for scalar; bold lowercase (e.g. $\boldsymbol{v}$) for vector; bold uppercase (e.g. $\boldsymbol{X}$) for matrix; calligraphy letter (e.g. $\mathcal{X}$) for higher-order tensor.

- $[n]$: the set $\{1, 2, \ldots, n\}$;
- $\circ$: vector outer product;
- $\|\cdot\|$: ℓ_2 norm for vectors;

For an order-K tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$

- $\mathrm{Vec}(\mathcal{X}) \in \mathbb{R}^{n_1 \cdots n_K}$: vectorization of $\mathcal{X}$;
- $\mathrm{Mat}_k(\mathcal{X}) \in \mathbb{R}^{n_k \times \prod_{l \neq k} n_l}$: matricization of $\mathcal{X}$ on the k-th mode;
- $\times_k$: tensor-to-matrix product on the k-th mode; for a matrix $\boldsymbol{M} \in \mathbb{R}^{p \times n_k}$, we have $\mathcal{X} \times_k \boldsymbol{M} \in \mathbb{R}^{n_1 \times \cdots \times p \times \cdots \times n_K}$;
- $\langle \cdot, \cdot \rangle$: tensor inner-product;
- $\|\cdot\|_F$: tensor Frobenius norm.

Review: Matrix decomposition

- Singular value decomposition (SVD) for matrix $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T = \sum_{r=1}^{R} \lambda_r \mathbf{u}_r \boldsymbol{v}_r^T,$$

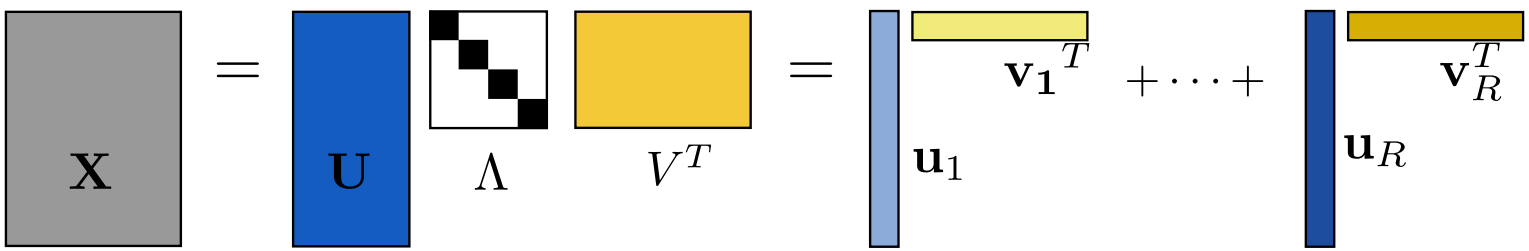


Figure 4: Illustration of matrix SVD.

- Tensor analogues of matrix SVD?

CANDECOMP/PARAFAC (CP) decomposition

- CP decomposition for an order-K tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$:

$$\mathcal{X} = \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r^{(1)} \circ \cdots \circ \boldsymbol{a}_r^{(K)},$$

where $\circ$ refers to vector outer product, $\lambda_1 \geq \cdots \geq \lambda_R$, $\boldsymbol{a}_r^{(k)} \in \mathbb{R}^{n_k}$, and $\|\boldsymbol{a}_r^{(k)}\| = 1$.

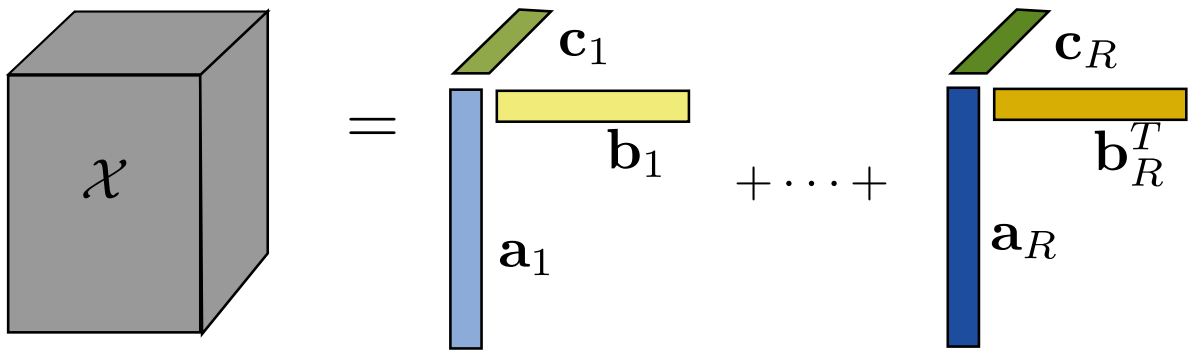


Figure 5: CP decomposition for order-3 tensor.

CANDECOMP/PARAFAC (CP) decomposition

- CP decomposition for an order-K tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$:

$$\mathcal{X} = \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r^{(1)} \circ \cdots \circ \boldsymbol{a}_r^{(K)},$$

where $\circ$ refers to vector outer product, $\lambda_1 \geq \cdots \geq \lambda_R$, $\boldsymbol{a}_r^{(k)} \in \mathbb{R}^{n_k}$, and $\|\boldsymbol{a}_r^{(k)}\| = 1$.

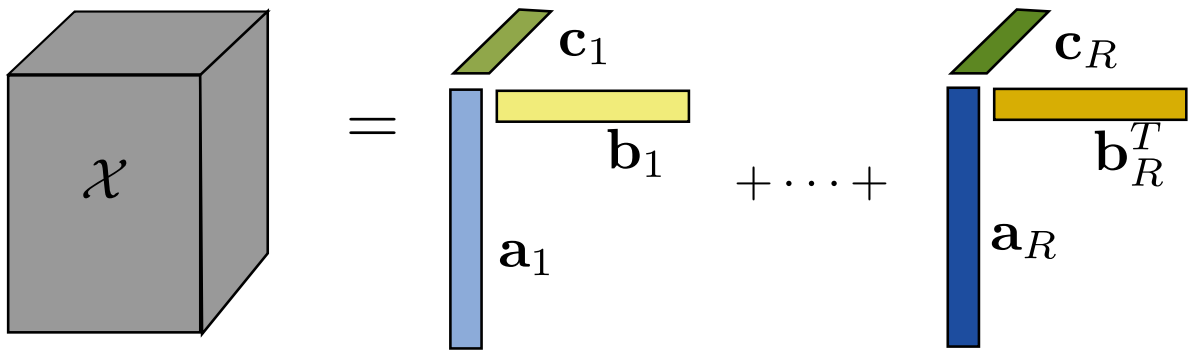


Figure 5: CP decomposition for order-3 tensor.

- Minimal R is call the *rank* of tensor $\mathcal{X}$.

Matrix decomposition vs CP decomposition

Rank properties [Kolda & Bader, 2009]:

- The tensor rank of a real-valued tensor may be different over $\mathbb{R}$ and $\mathbb{C}$.
- No straightforward algorithm to determine the rank of a specific given tensor; the problem is NP-hard.
- The tensor rank may exceeds the dimension; i.e., $R > n_k$ is possible. For order-3 tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we have a weak upper bound

$$\operatorname{rank}(\mathcal{X}) \leq \min\{n_1 n_2, n_1 n_3, n_2 n_3\}.$$

Matrix decomposition vs CP decomposition

Uniqueness (up to permutation and scaling) [Kolda & Bader, 2009]:

- A rank $R > 1$ matrix can be decomposed in multiple ways in the case of degenerate λ_is with multiplicity larger than 1;
- Kruskal's result guarantees the uniqueness of CP decomposition under a weak condition.

Kruskal's theorem

Let $\boldsymbol{A}^k \in \mathbb{R}^{n_k \times R}$ denote the factor matrix with columns $\boldsymbol{a}_r^{(k)}$ for all $k \in [K]$. A sufficient condition for CP decomposition uniqueness is

$$\sum_{k \in [K]} \operatorname{rank}(\boldsymbol{A}^{(k)}) \geq 2R + (K - 1).$$

This is also the necessary condition for tensors of rank $R = 2, 3$, but not for $R > 3$.

Tucker decomposition

- Tucker decomposition for $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$:

$$\mathcal{X} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K,$$

where $\mathcal{S} \in \mathbb{R}^{R_1 \times \ldots \times R_K}$ is the core tensor, $\boldsymbol{M}_k \in \mathbb{R}^{n_k \times R_k}$ are the factor matrices.

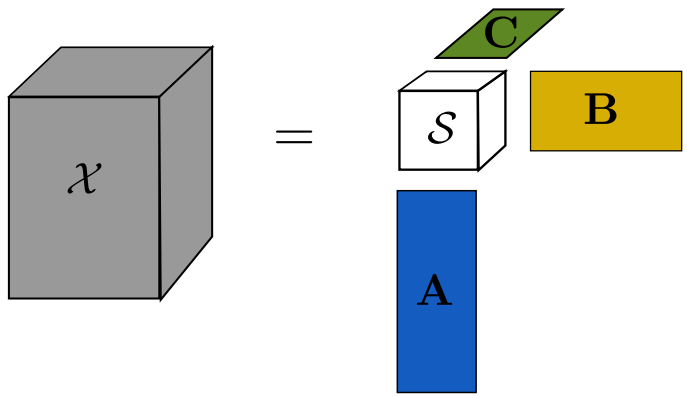


Figure 6: Tucker decomposition for order-3 tensor.

Tucker decomposition

- Tucker decomposition for $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$:

$$\mathcal{X} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K,$$

where $\mathcal{S} \in \mathbb{R}^{R_1 \times \ldots \times R_K}$ is the core tensor, $\boldsymbol{M}_k \in \mathbb{R}^{n_k \times R_k}$ are the factor matrices.

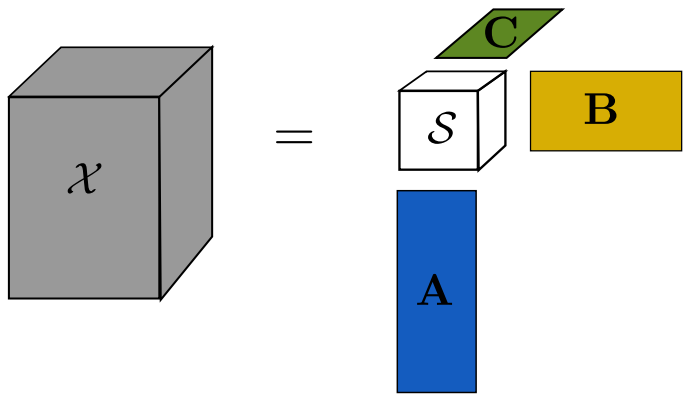


Figure 6: Tucker decomposition for order-3 tensor.

- Tucker rank $\boldsymbol{r} = (R_1, \cdots, R_K)$, where $R_k = \mathrm{rank}(\mathrm{Mat}_k(\mathcal{X}))$.
- CP decomposition is a special case of Tucker decomposition with $R_1 = \cdots = R_K$ and only diagonal $\mathcal{S}_{r \ldots r}$ non-zero.

Outline

General goal

- **Identify the low-dimensional representation from the noisy, high-dimensional, and higher-order tensor data.**

General goal

- **Identify the low-dimensional representation from the noisy, high-dimensional, and higher-order tensor data.**
- Specifically, we aim to recover the low-dimensional signal $\mathcal{X}$ in the signal-plus-noise model

$$\underbrace{\mathcal{Y}}_{\text{noisy tensor observation}} = \underbrace{\mathcal{X}}_{\text{low-dimensional signal tensor}} + \underbrace{\mathcal{E}}_{\text{noise}},$$

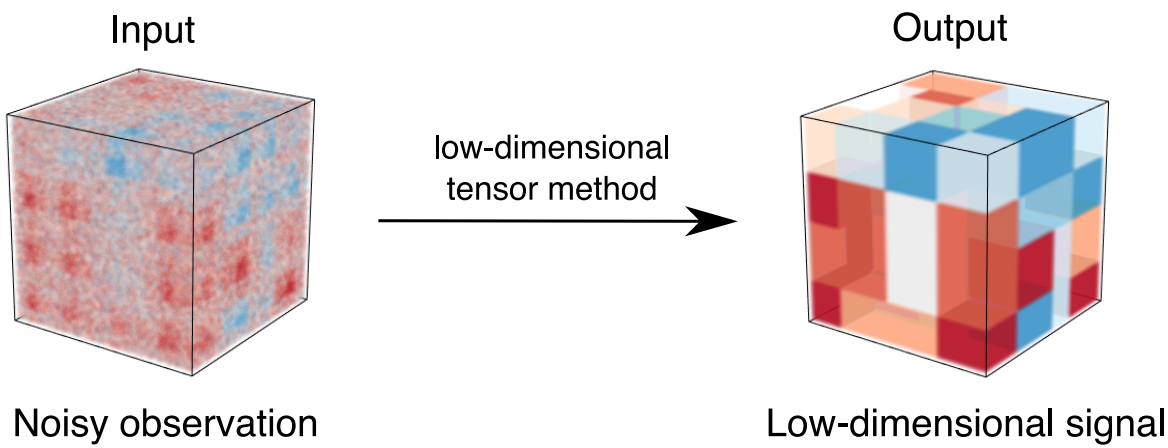


Figure 7: Illustration of the example low-dimensional tensor method.

Challenges

- Conceptual: What kind of low-dimensional structure should we use for tensors?
- Computational: How to develop efficient algorithms to solve the tensor problem?

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

Low rank approximation

- Goal: Recover the low-rank signal from noisy observations.
- CP Model:

$$\mathcal{Y} = \sum_{r \in [R]} \lambda_r \boldsymbol{a}_r^{(1)} \circ \cdots \boldsymbol{a}_r^{(K)} + \mathcal{E}.$$

Low rank approximation

- Goal: Recover the low-rank signal from noisy observations.
- CP Model:

$$\mathcal{Y} = \sum_{r \in [R]} \lambda_r \boldsymbol{a}_r^{(1)} \circ \cdots \boldsymbol{a}_r^{(K)} + \mathcal{E}.$$

- Best low rank approximation? **May not exist!**

$$\hat{\mathcal{X}}_{\text{MLE}} = \underset{\mathcal{X}:\text{rank}(\mathcal{X}) \leq R}{\arg\min} \|\mathcal{Y} - \mathcal{X}\|_F^2.$$

This is an ill-posed problem. Optimizer may not exist.

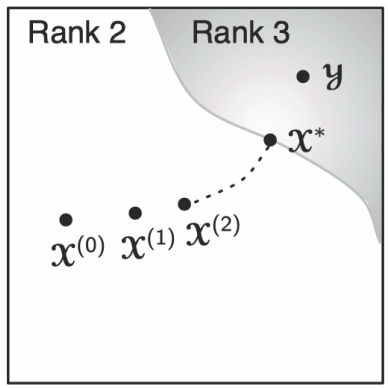


Figure 8: A sequence of rank-2 tensors converging to one of higher rank [Kolda & Bader, 2009].

Low rank approximation

- Solution: impose manageable structure to the signal:

  - *orthogonally decomposable tensors:*
  assume $\{\boldsymbol{a}_r^{(k)}\}_r$ are orthonormal vectors; i.e., $(\boldsymbol{a}_r^{(k)})^T \boldsymbol{a}_{r'}^{(k)} = 0$
  [Kolda, 2001], [Chen & SAAD, 2009], [Auddy & Ming, 2020], ...
  - *symmetric and orthogonally decomposable tensors:*
  further assume $\boldsymbol{a}_r^{(1)} = \cdots = \boldsymbol{a}_r^{(K)}$
  [Robeva, 2016], [Wang & Song, 2017], ...

Low rank approximation

- Solution: impose manageable structure to the signal:

  - *orthogonally decomposable tensors:*
  assume $\{\boldsymbol{a}_r^{(k)}\}_r$ are orthonormal vectors; i.e., $(\boldsymbol{a}_r^{(k)})^T \boldsymbol{a}_{r'}^{(k)} = 0$
  [Kolda, 2001], [Chen & SAAD, 2009], [Auddy & Ming, 2020], ...
  - *symmetric and orthogonally decomposable tensors:*
  further assume $\boldsymbol{a}_r^{(1)} = \cdots = \boldsymbol{a}_r^{(K)}$
  [Robeva, 2016], [Wang & Song, 2017], ...

- Not all tensors are orthogonally decomposable!

Tensor PCA/SVD

- Goal: Recover the low-rank signal from noisy observations.
- Tucker model:

$$\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K + \mathcal{E}.$$

Tensor PCA/SVD

- Goal: Recover the low-rank signal from noisy observations.
- Tucker model:

$$\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K + \mathcal{E}.$$

- Best low rank approximation?

$$\hat{\mathcal{X}}_{\mathrm{MLE}} = \underset{\mathcal{X}:\boldsymbol{r}(\mathcal{X})\leq(R_1,\ldots,R_K)}{\arg\min} \|\mathcal{Y} - \mathcal{X}\|_F^2,$$

$$\hat{\boldsymbol{M}}_{k,\mathrm{MLE}} = \mathrm{SVD}_{R_k}(\mathrm{Mat}_k(\hat{\mathcal{X}}_{\mathrm{MLE}}))$$

May not have unique solution!

Tensor PCA/SVD

- Goal: Recover the low-rank signal from noisy observations.
- Tucker model:

$$\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K + \mathcal{E}.$$

- Best low rank approximation?

$$\hat{\mathcal{X}}_{\mathrm{MLE}} = \underset{\mathcal{X}:\boldsymbol{r}(\mathcal{X})\leq(R_1,\ldots,R_K)}{\arg\min} \|\mathcal{Y} - \mathcal{X}\|_F^2,$$
$$\hat{\boldsymbol{M}}_{k,\mathrm{MLE}} = \mathrm{SVD}_{R_k}(\mathrm{Mat}_k(\hat{\mathcal{X}}_{\mathrm{MLE}}))$$

May not have unique solution!

- Action: Measure the singular subspaces difference

$$\sin\Theta(\hat{\boldsymbol{M}}, \boldsymbol{M}) = \max\{\cos(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x} \in \mathrm{span}(\hat{\boldsymbol{M}}), \boldsymbol{y} \in \mathrm{span}(\boldsymbol{M}^{\perp})\}.$$

Statistical and computational limits

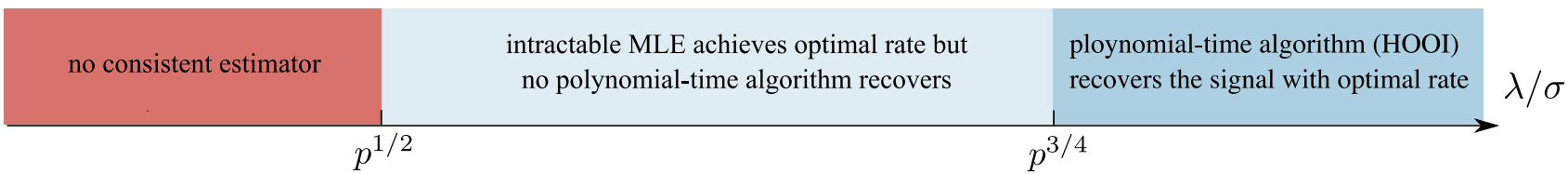
- MLE is NP-hard to compute.
- When can we develop a polynomial-time algorithm to solve tensor SVD?

Statistical and computational limits

- MLE is NP-hard to compute.
- When can we develop a polynomial-time algorithm to solve tensor SVD?

[Zhang & Xia, 2018]

Consider the Tucker model for order-3 tensor $\mathcal{X} \in \mathbb{R}^{p \times p \times p}$. Define the signal-to-noise ratio λ/σ where $\lambda = \min_{k=1,2,3} \lambda_{R_k}(\mathrm{Mat}_k(\mathcal{X}))$. We have phase transition



Tensor clustering

- Goal: Recover community assignments $z_k : [n_k] \mapsto [R_k]$ on every mode of the tensor.

Tensor clustering

- Goal: Recover community assignments $z_k : [n_k] \mapsto [R_k]$ on every mode of the tensor.
- Tensor block model (TBM):

$$\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K + \mathcal{E},$$

- $\mathcal{S} \in \mathbb{R}^{R_1 \times \cdots \times R_K}$ collects the block means;
- $\boldsymbol{M}_k \in \{0,1\}^{n_k \times R_k}$ are membership matrices of our main interest, where $\boldsymbol{M}_{k,ij} = \mathbb{1}\{z_k(i) = j\}$.

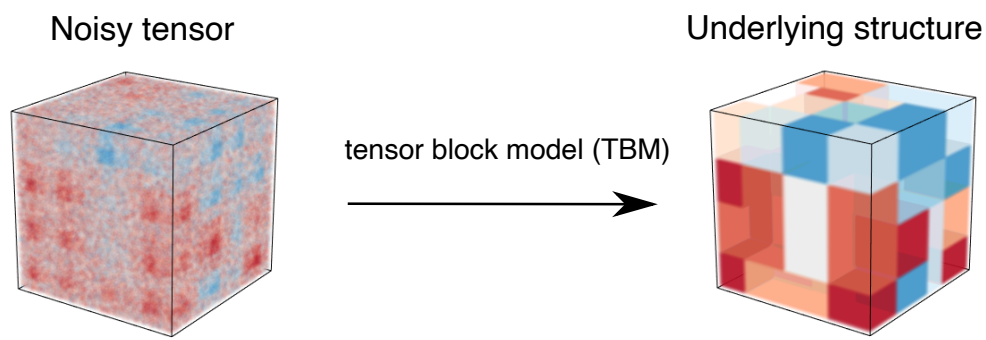


Figure 9: Order-3 tensor block model.

Other problems

- Tensor regression:
  - Goal: Recover low-dimensional tensor coefficient.
  - Model: For data (y_i, X_i),

$$y_i = \langle \mathcal{X}_i, \mathcal{T} \rangle + \mathcal{E}, \quad \mathcal{T} \text{ has low-rank structure}$$

Other problems

- Tensor regression:
  - Goal: Recover low-dimensional tensor coefficient.
  - Model: For data (y_i, X_i),

$$y_i = \langle \mathcal{X}_i, \mathcal{T} \rangle + \mathcal{E}, \quad \mathcal{T} \text{ has low-rank structure}$$

- Tensor completion:
  - Goal: Recover low-dimensional signal given observations with missing entries.
  - Model: For given incomplete data $\mathcal{Y}$,

$$\mathcal{Y} = \mathcal{X} + \mathcal{E}, \quad \text{only } \mathcal{Y}_\omega \text{ for } \omega \in \Omega \text{ are observed,}$$

where $\Omega \subset [n_1] \times \cdots \times [n_K]$ is the set of observed indices.

Outline

Motivation: Side information

- Scientific studies (e.g. neuroimaging, social network analysis) often collect the tensor observations with side information.

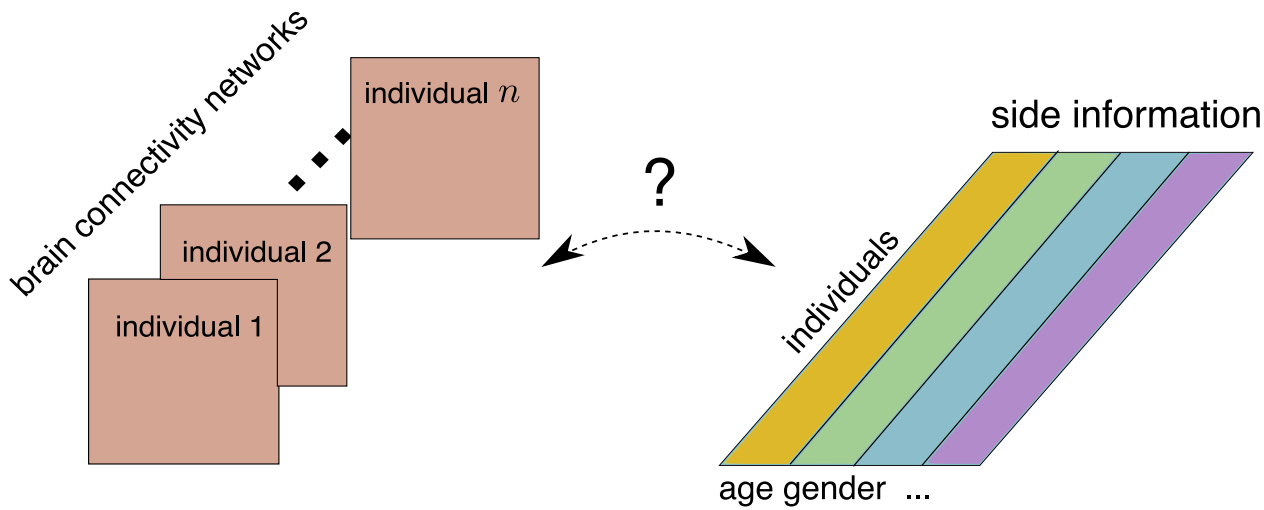


Figure 10: Brain connectivity networks (binary adjacency matrices) with side information.

Motivation: Side information

- Scientific studies (e.g. neuroimaging, social network analysis) often collect the tensor observations with side information.

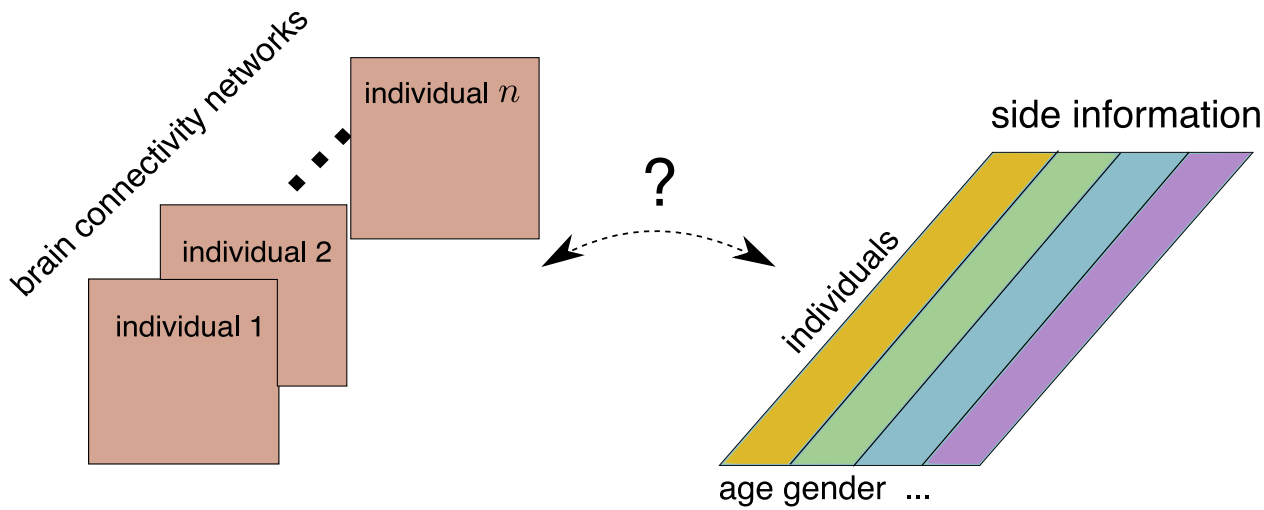


Figure 10: Brain connectivity networks (binary adjacency matrices) with side information.

Can we identify low-rank structure in the data tensor affected by side information?

Supervised tensor decomposition

Our model:

$$\begin{aligned}\mathbb{E}[\mathcal{Y}|\boldsymbol{X}_1,...,\boldsymbol{X}_K] &= f\left(\mathcal{B}\times_1\boldsymbol{X}_1\times_2\cdots\times_K\boldsymbol{X}_K\right)\\ &\overset{\mathcal{B}\sim Tucker(\boldsymbol{r})}{=} f\left(\mathcal{C}\times_1\boldsymbol{X}_1\boldsymbol{M}_1\times_2\cdots\times_K\boldsymbol{X}_K\boldsymbol{M}_K\right),\end{aligned}$$

where

- $\mathcal{Y}\in\mathbb{R}^{d\times\cdots\times d}$: an order-$K$ tensor observation;
- $\boldsymbol{X}_i\in\mathbb{R}^{d\times p}$: feature matrices contain the side information, for $i=1,...,K$.
- $f(\cdot)$: known link function depending on the data type of $\mathcal{Y}$;
- $\mathcal{B}\sim Tucker(\boldsymbol{r}), \mathcal{B}\in\mathbb{R}^{p\times\cdots\times p}$: unknown coefficient tensor.

Supervised tensor decomposition

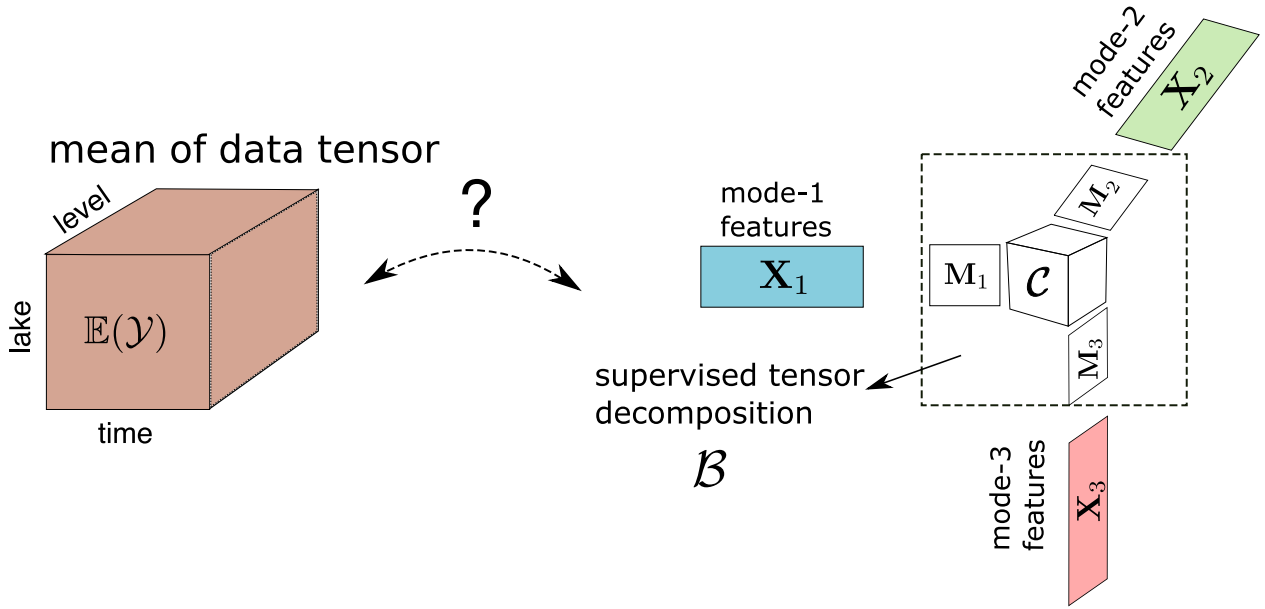


Figure 11: Supervised tensor decomposition for an order-3 tensor with side information.

Supervised tensor decomposition

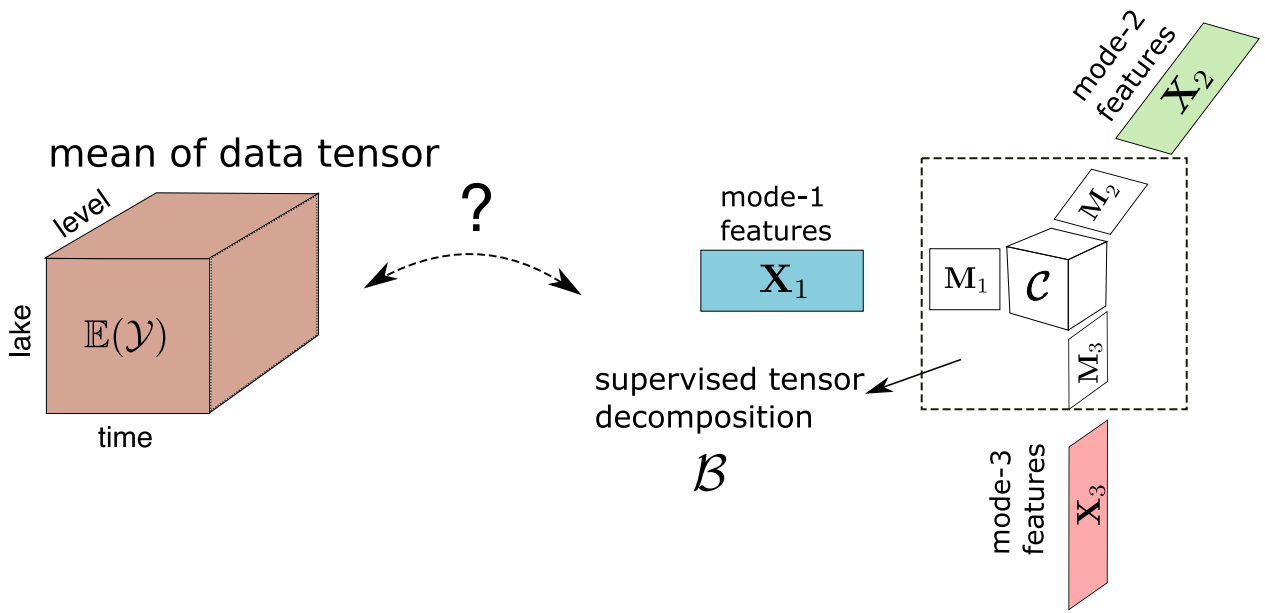


Figure 11: Supervised tensor decomposition for an order-3 tensor with side information.

- **Why "supervised"?** The factor matrices of $\mathbb{E}[\mathcal{Y}|\boldsymbol{X}_k]$, $\boldsymbol{X}_k\boldsymbol{M}_k$, are restricted in the column space of $\boldsymbol{X}_k$.
- The coefficient $\mathcal{B}$ is identifiable, but ingredients $\mathcal{C}, \boldsymbol{M}_1, ..., \boldsymbol{M}_K$ are not identifiable with orthogonal transformation.

Global estimator

- We propose a rank-constrained likelihood-based estimator and its convergence rate.

$$(\hat{\mathcal{C}}_{MLE}, \hat{\boldsymbol{M}}_{1,MLE}, ..., \hat{\boldsymbol{M}}_{K,MLE}) = \underset{(\mathcal{C},\boldsymbol{M}_1,...,\boldsymbol{M}_K)\in\mathcal{P}}{\arg\max} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, ..., \boldsymbol{M}_K),$$

Global estimator

- We propose a rank-constrained likelihood-based estimator and its convergence rate.

$$(\hat{\mathcal{C}}_{MLE}, \hat{\boldsymbol{M}}_{1,MLE}, ..., \hat{\boldsymbol{M}}_{K,MLE}) = \underset{(\mathcal{C},\boldsymbol{M}_1,...,\boldsymbol{M}_K)\in\mathcal{P}}{\arg\max} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, ..., \boldsymbol{M}_K),$$

Theorem (Global optimizer convergence)

Under mild technical conditions, with probability at least $1 - \exp(-p)$, *we have*

$$\max_{k\in[K]} \sin^2\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_{k,MLE}) = \mathcal{O}\left(\frac{Kp}{\lambda^2 d^K}\right), \left\|\mathcal{B}_{true} - \hat{\mathcal{B}}_{MLE}\right\|_F^2 = \mathcal{O}\left(\frac{Kp}{d^K}\right)$$

where $\sin\Theta(\boldsymbol{M}, \boldsymbol{M}')$ *is the angle distance between the column spaces of* $\boldsymbol{M}$ *and* $\boldsymbol{M}'$.

Efficient alternative algorithm

- We propose an alternating algorithm that updates one block of parameters when fixes other blocks of parameters.

Efficient alternative algorithm

- We propose an alternating algorithm that updates one block of parameters when fixes other blocks of parameters.

Theorem (Algorithm convergence)

Suppose we have a good initialization with signal-to-noise ratio larger than $\mathcal{O}(p^{K/2}d^{-K})$. *Under mild technical conditions, with probability at least* $1-\exp(-p)$, *there exists a contraction parameters* $\rho \in [0,1]$,

$$\max_{k\in[K]} \sin^2\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(t)}) \lesssim \mathcal{O}\left(\frac{p}{\lambda^2 d^K}\right) + \rho^t \max_{k\in[K]} \sin\Theta^2(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(0)}).$$

The first term represents statistical error and the second term is represents algorithmic error.

Real data

The Human Connectome Project (HCP) data contains 68-by-68 binary symmetric brain network matrices from 136 individuals, denoted $\mathcal{Y} \in \{0,1\}^{68\times 68\times 136}$. We apply our method to the HCP with age and gender information of the individuals.

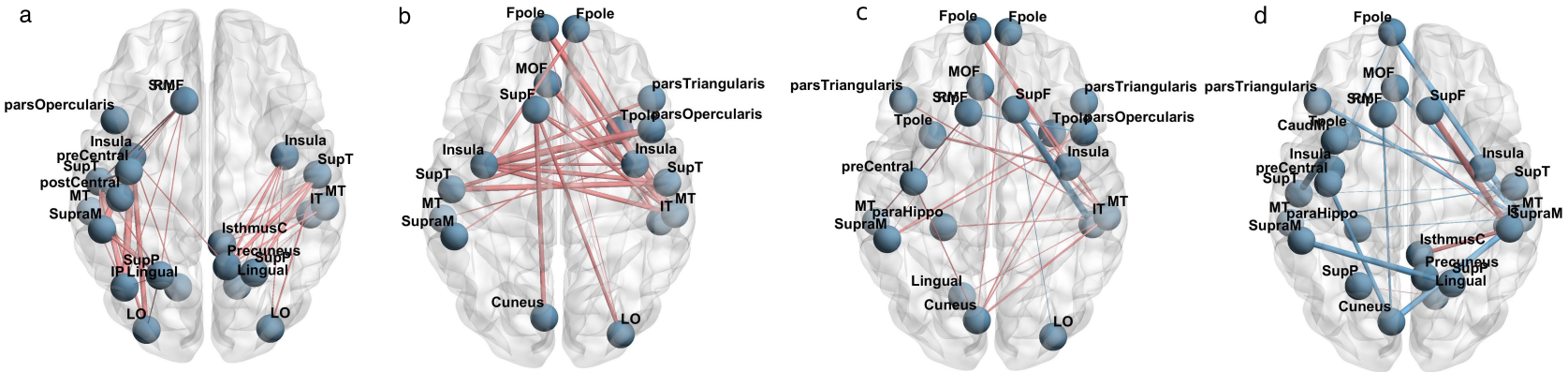


Figure 12: Top edges with large effects. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+. Red (blue) edges represent positive (negative) effects. Edge-widths are proportional to the magnitude of effect sizes.

Outline

Recall: multiway clustering

- Multiway clustering detects the community structure on each mode of the tensor.

Recall: multiway clustering

- Multiway clustering detects the community structure on each mode of the tensor.
- Tensor block model (TBM) [Han, Luo, Wang and Zhang, 2021] is a common model for multiway clustering.

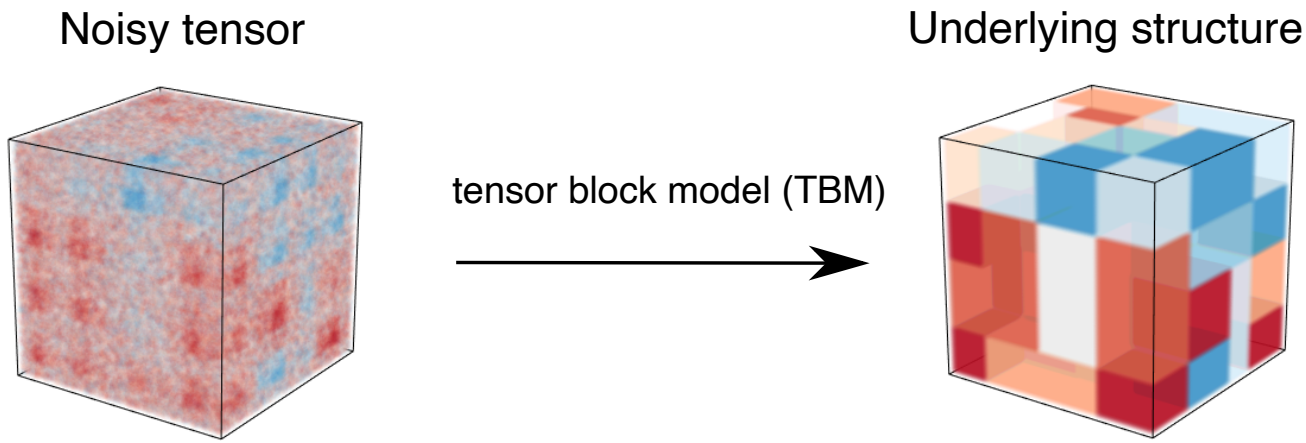


Figure 13: Order-3 tensor block model.

Motivation: Individual heterogeneity

- However, TBM unrealistically assumes no individual-specific effects of entries.

Motivation: Individual heterogeneity

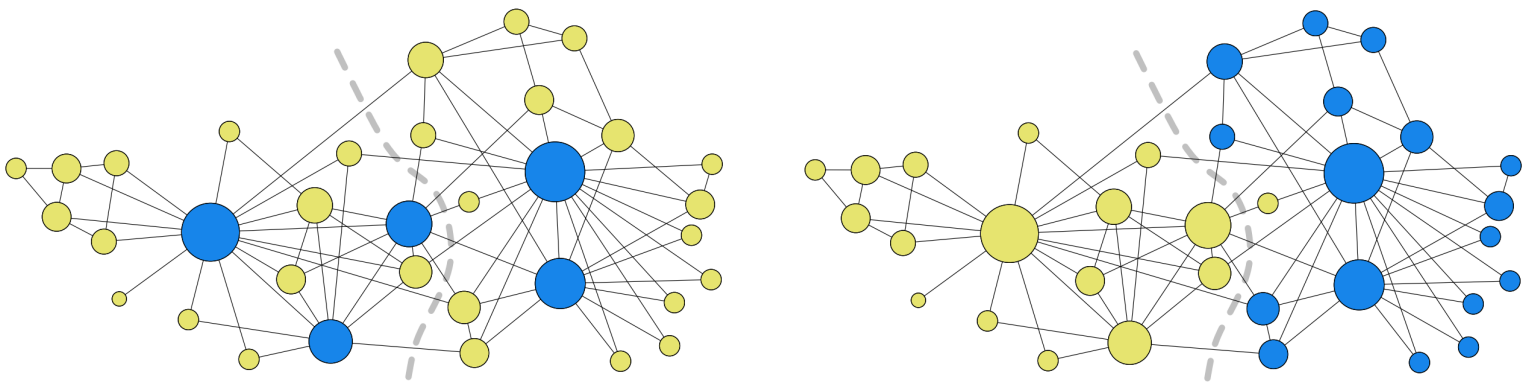
- However, TBM unrealistically assumes no individual-specific effects of entries.
- For example, TBM treats students and advisors in the same way when they are assigned to one co-authorship cluster.

Motivation: Individual heterogeneity

- However, TBM unrealistically assumes no individual-specific effects of entries.
- For example, TBM treats students and advisors in the same way when they are assigned to one co-authorship cluster.
- Ignoring the unequal individual heterogeneity may seriously mislead the clustering results.

Motivation: Individual heterogeneity

- However, TBM unrealistically assumes no individual-specific effects of entries.
- For example, TBM treats students and advisors in the same way when they are assigned to one co-authorship cluster.
- Ignoring the unequal individual heterogeneity may seriously mislead the clustering results.



(a) Without degree correction

(b) With degree-correction

Figure 14: Clustering results of karate club (matrix) network with (a) uncorrected and (b) degree-corrected block model [Karrer and Newman, 2010]. Degree correction accounts for the individual heterogeneity. Grey dashed line indicates true split.

Degree-corrected tensor block model

We propose the **degree-corrected tensor block model (dTBM)**. Assume there exist r communities for p nodes on each mode.

$$\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{\Theta}\boldsymbol{M} \times_2 \cdots \times_K \boldsymbol{\Theta}\boldsymbol{M} + \mathcal{E}, \tag{1}$$

where

- $\mathcal{Y} \in \mathbb{R}^{p \times \cdots \times p}$: tensor collecting order-K connections among p nodes;
- $\mathcal{S} \in \mathbb{R}^{r \times \cdots \times r}$: core tensor collecting block means;
- $\boldsymbol{M} \in \{0,1\}^{p \times r}$: membership matrix indicating the clustering assignment of our main interest;
- $\boldsymbol{\Theta} \in \mathbb{R}^{p \times p}$: diagonal matrix collecting the heterogeneity;
- $\mathcal{E} \in \mathbb{R}^{p \times \cdots \times p}$: zero-mean sub-Gaussian noise tensor with entrywise variance bounded by σ^2;
- $\times_k$: matrix-by-tensor product on the k-th mode.

Degree-corrected tensor block model

The dTBM allows a richer structure with heterogeneity.

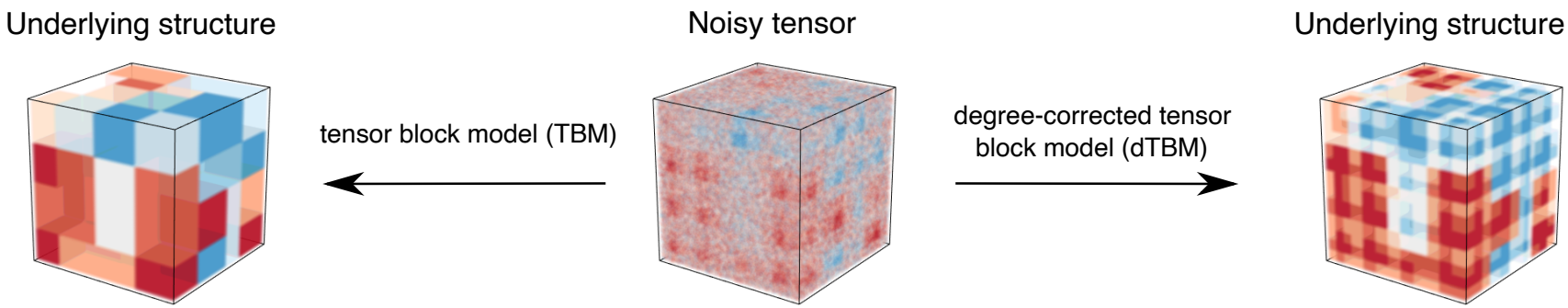


Figure 15: Examples for order-3 TBM and dTBM with 4 clusters on each mode.

Our results: Identifiability

Angle gap assumption:

We find the clustering performance is characterized by

$$\Delta_{\min}^2 = \min_{a \neq b \in \{1,\ldots,r\}} \left\| \frac{\mathrm{Mat}(\mathcal{S})_{a:}}{\|\mathrm{Mat}(\mathcal{S})_{a:}\|} - \frac{\mathrm{Mat}(\mathcal{S})_{b:}}{\|\mathrm{Mat}(\mathcal{S})_{b:}\|} \right\| > 0,$$

where $\|\cdot\|$ is Frobenius norm, $\mathrm{Mat}(\cdot)$ unfold tensors to matrices, $\boldsymbol{X}_{a:}$ represents the a-th row of matrix $\boldsymbol{X}$.

Our results: Identifiability

Angle gap assumption:

We find the clustering performance is characterized by

$$\Delta_{\min}^2 = \min_{a \neq b \in \{1,\ldots,r\}} \left\| \frac{\mathrm{Mat}(\mathcal{S})_{a:}}{\|\mathrm{Mat}(\mathcal{S})_{a:}\|} - \frac{\mathrm{Mat}(\mathcal{S})_{b:}}{\|\mathrm{Mat}(\mathcal{S})_{b:}\|} \right\| > 0,$$

where $\|\cdot\|$ is Frobenius norm, $\mathrm{Mat}(\cdot)$ unfold tensors to matrices, $\boldsymbol{X}_{a:}$ represents the a-th row of matrix $\boldsymbol{X}$. Equivalently, we have

$$\max_{a \neq b \in \{1,\ldots,r\}} \cos(\mathrm{Mat}(\mathcal{S})_{a:}, \mathrm{Mat}(\mathcal{S})_{b:}) < 1.$$

Our results: Identifiability

Angle gap assumption:

We find the clustering performance is characterized by

$$\Delta_{\min}^2 = \min_{a \neq b \in \{1,\ldots,r\}} \left\| \frac{\mathrm{Mat}(\mathcal{S})_{a:}}{\|\mathrm{Mat}(\mathcal{S})_{a:}\|} - \frac{\mathrm{Mat}(\mathcal{S})_{b:}}{\|\mathrm{Mat}(\mathcal{S})_{b:}\|} \right\| > 0,$$

where $\|\cdot\|$ is Frobenius norm, $\mathrm{Mat}(\cdot)$ unfold tensors to matrices, $\boldsymbol{X}_{a:}$ represents the a-th row of matrix $\boldsymbol{X}$. Equivalently, we have

$$\max_{a \neq b \in \{1,\ldots,r\}} \cos(\mathrm{Mat}(\mathcal{S})_{a:}, \mathrm{Mat}(\mathcal{S})_{b:}) < 1.$$

Identifiability Theorem [Hu and Wang, 2021+]

The parameterization of dTBM (1) is unique up to label permutation if and only if the Angle Gap Assumption holds, under a few technical constraints.

Our results: Statistical and computational limits

Phase-transition Theorem [Hu and Wang, 2021+]

Define the signal-to-noise ratio (SNR):$= \Delta_{\min}^2/\sigma^2 = p^\gamma$. Under order-$K$ dTBM (1) with technical conditions and conjectures,

Our results: Statistical and computational limits

Phase-transition Theorem [Hu and Wang, 2021+]

Define the signal-to-noise ratio (SNR):$= \Delta_{\min}^2/\sigma^2 = p^\gamma$. Under order-$K$ dTBM (1) with technical conditions and conjectures,

- **Statistical limit:**
  - Impossibility: every estimator fails to fully recover with unlimited computational cost when $\gamma < -(K-1)$;
  - Achievability: MLE achieves exact recovery with error $\exp\left(-\frac{p^{K-1}}{r^{K-1}}\text{SNR}\right)$ when $\gamma > -(K-1)$.

Our results: Statistical and computational limits

Phase-transition Theorem [Hu and Wang, 2021+]

Define the signal-to-noise ratio (SNR):$= \Delta_{\min}^2/\sigma^2 = p^\gamma$. Under order-$K$ dTBM (1) with technical conditions and conjectures,

- **Statistical limit:**
  - Impossibility: every estimator fails to fully recover with unlimited computational cost when $\gamma < -(K-1)$;
  - Achievability: MLE achieves exact recovery with error $\exp\left(-\frac{p^{K-1}}{r^{K-1}}\text{SNR}\right)$ when $\gamma > -(K-1)$.
- **Computational limit:**
  - Impossibility: every polynomial-time estimator fails to fully recover when $\gamma < -K/2$;
  - Achievability: there exists a polynomial-time algorithm achieves exact recovery with error $\exp\left(-\frac{p^{K-1}}{r^{K-1}}\text{SNR}\right)$ when $\gamma > -K/2$.

Our results: Statistical and computational limits

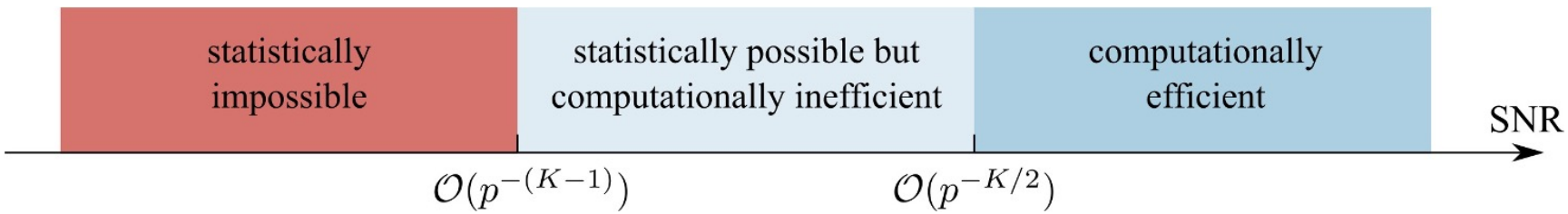


Figure 16: SNR thresholds for statistical and computational limits with $K \geq 3$.

Statistical-computational gap emerges for higher-order tensors with $K \geq 3$ modes.

Our results: Statistical and computational limits

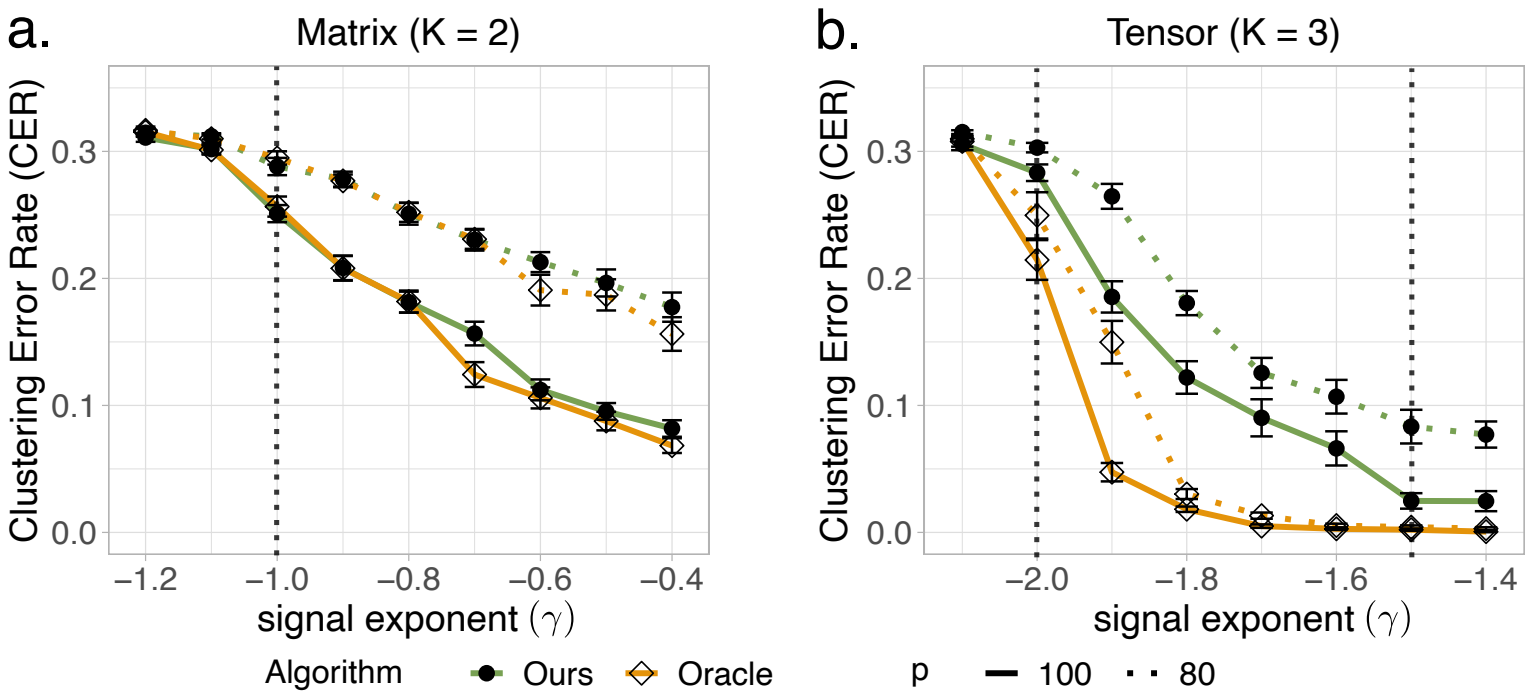


Figure 17: SNR phase transitions for clustering in dTBM with $r = 5$ under (a) matrix case and (b) tensor case. Ours method is a polynomial-time algorithm, and Oracle is an approximation of maximum likelihood estimation (MLE).

Our results: Polynomial-time Algorithm

Global-to-local algorithm:

- *Global*: Spectral-based initialization with polynomial error;
- *Local*: Angle-based iteration to achieve exact clustering.

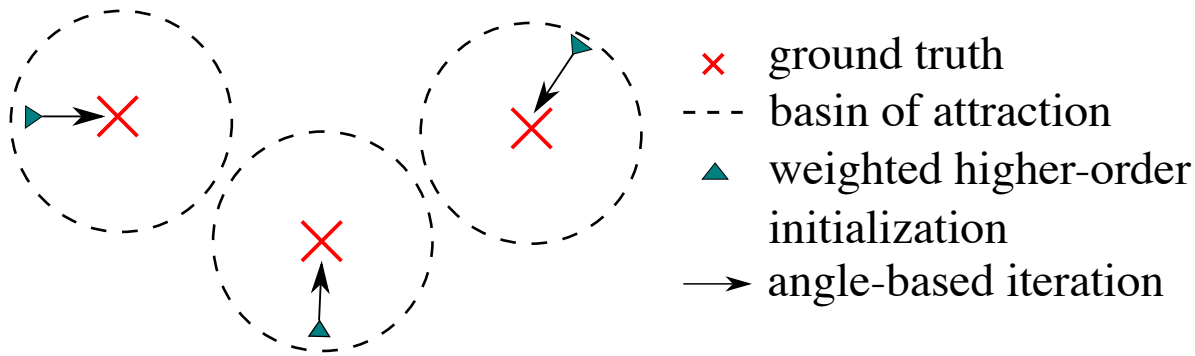


Figure 18: Illustration of our two-stage global-to-local algorithm.

Our results: Polynomial-time Algorithm

Algorithm Accuracy Theorem [Hu and Wang, 2021+]

Let $\ell^{(t)}$ denote the misclassification error in the t-th step. Under order-K dTBM (1) and technical conditions, with high probability

- initialization error:

$$\ell^{(0)} \lesssim \frac{r^K p^{-K/2}}{\mathrm{SNR}};$$

Our results: Polynomial-time Algorithm

Algorithm Accuracy Theorem [Hu and Wang, 2021+]

Let $\ell^{(t)}$ denote the misclassification error in the t-th step. Under order-K dTBM (1) and technical conditions, with high probability

- initialization error:

$$\ell^{(0)} \lesssim \frac{r^K p^{-K/2}}{\mathrm{SNR}};$$

- iteration error: when $\mathrm{SNR} \gtrsim p^{-K/2} \log p$,

$$\ell^{(t+1)} \lesssim \underbrace{\mathrm{SNR}^{-1} \exp\left(-\frac{p^{K-1}\mathrm{SNR}}{r^{K-1}}\right)}_{\text{statistical error}} + \underbrace{\rho^t \ell^{(0)}}_{\substack{\text{computational} \\ \text{error}}}, \ \rho \in (0,1).$$

Our results: Polynomial-time Algorithm

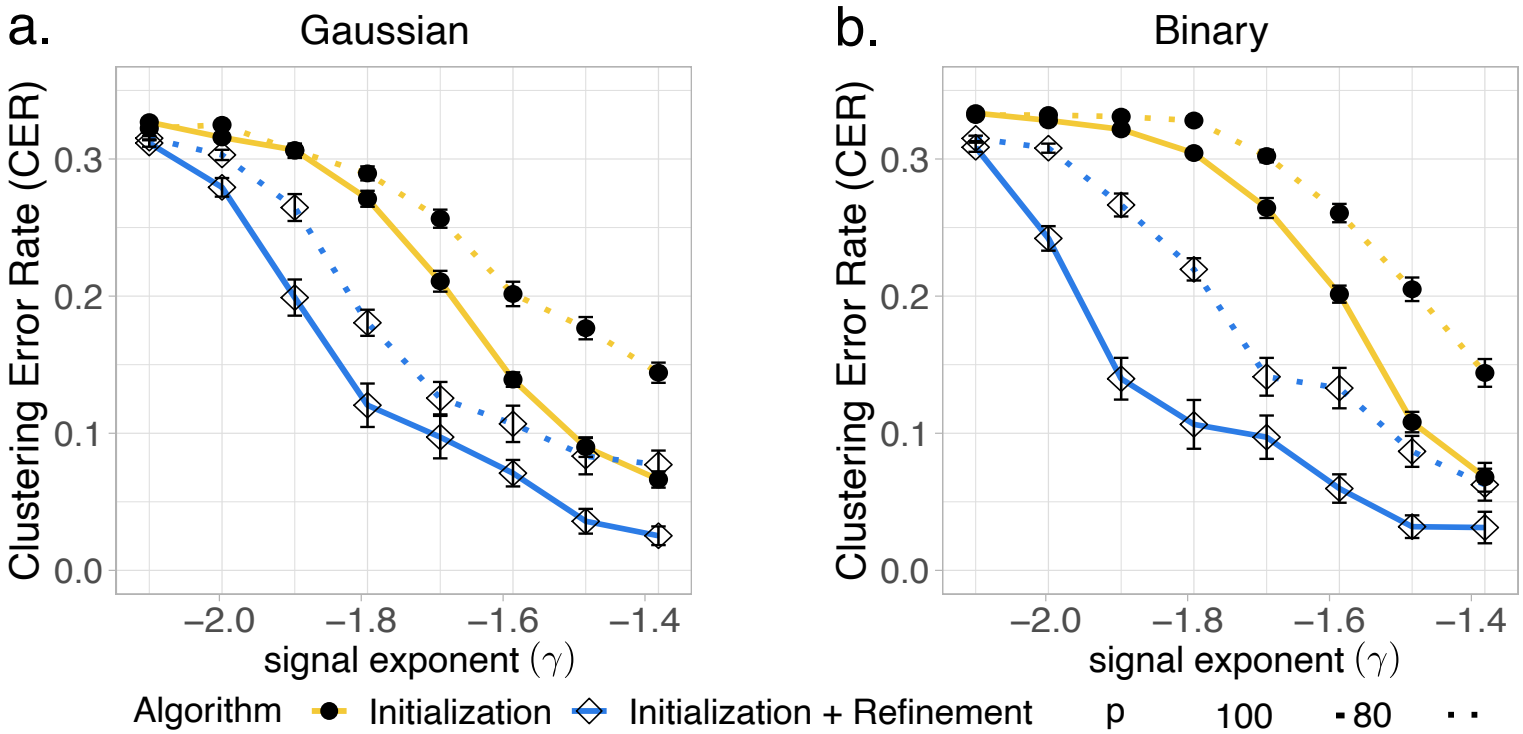


Figure 19: Clustering error versus signal exponent (γ) for initialization only and for combined algorithm under (a) Gaussian models and (b) Bernoulli models with $r = 5$.

Real data

We apply our algorithm to HCP data, with $r = 4$ blocks.

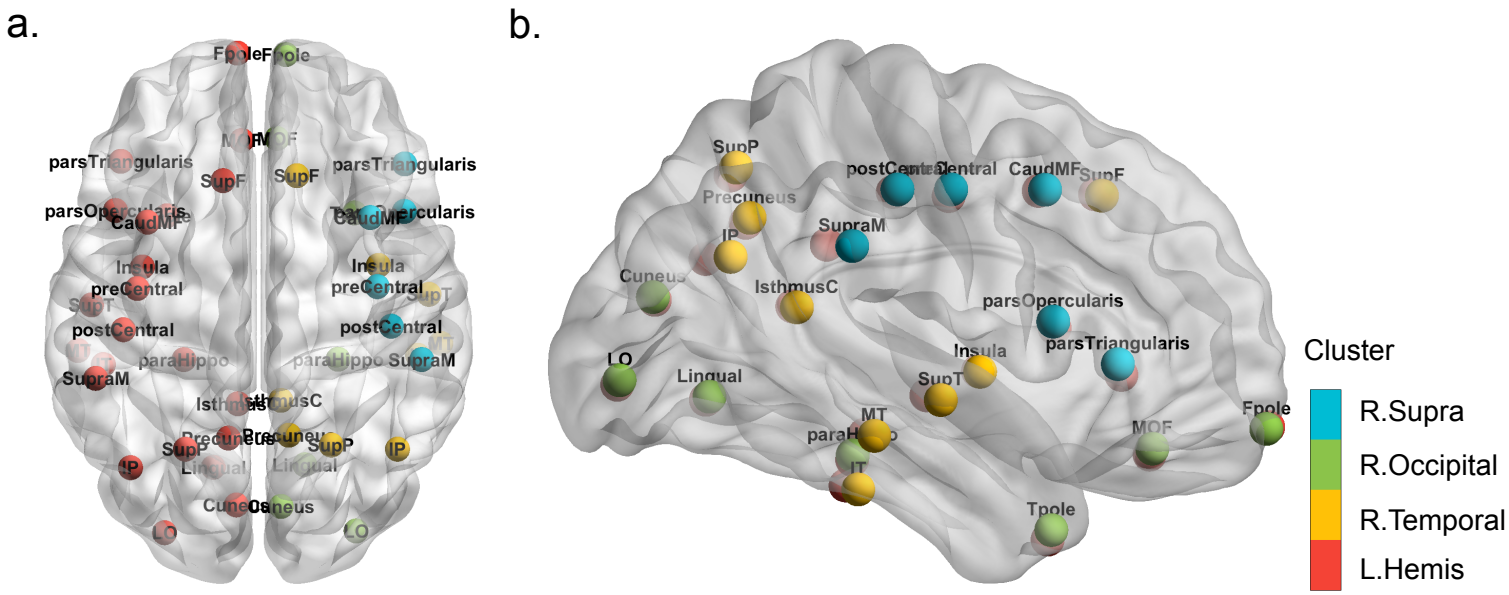


Figure 20: Illustration of brain node clustering results for HCP data with (a) top and (b) side views.

Summary

- Tensor methods provide a rich source of:
  - fundamental problem in data science;
  - new tools for long-standing higher-order problems;
  - potentials for new applications;
- Challenges:
  - conceptual: tensor algebra, various low-structure for tensors
  - computational: design efficient tensor algorithms

Acknowledgement: This research is supported in part by NSF CAREER DMS-2141865, DMS-1915978, DMS2023239, and funding from the Wisconsin Alumni Research foundation.