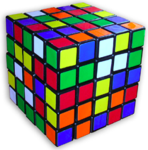
Beyond the signs: Nonparametric tensor completion via sign series

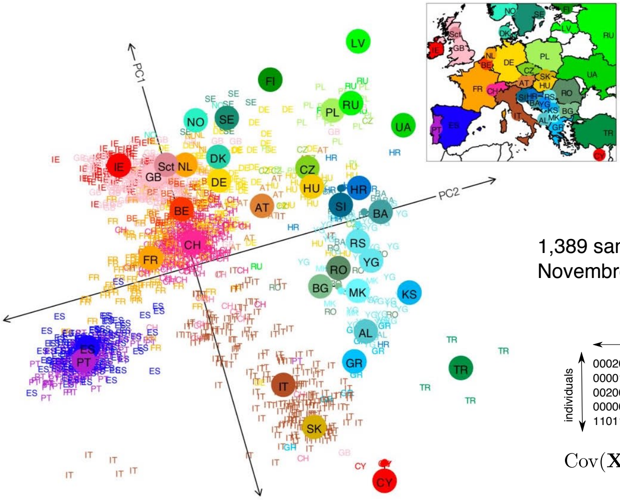
Miaoyan Wang

Department of Statistics, UW-Madison

Joint work with Chanwoo Lee (3rd-year PhD student)



A successful story: PCA of Europeans

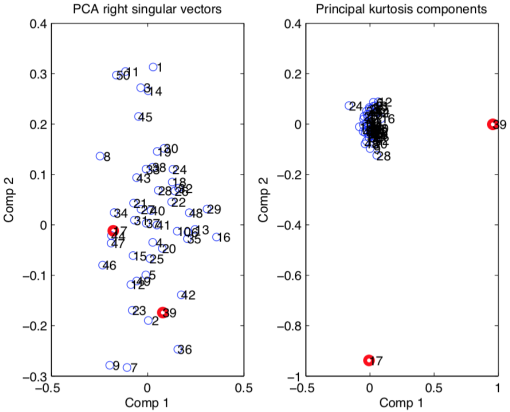


1,389 samples, ~ 200k SNPs
Novembre et al. (2008)

$$
\begin{array}{c}
\text{SNPs} \\
\begin{array}{l}
000201100000111110000000\ldots \\
000011000000120110000000\ldots \\
002001110120010100110111\ldots \\
000000000111210100101110\ldots \\
110110111011110120001001\ldots
\end{array}
\end{array}
\quad \text{(rows: individuals)}
$$

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

Matrix methods are powerful, however...



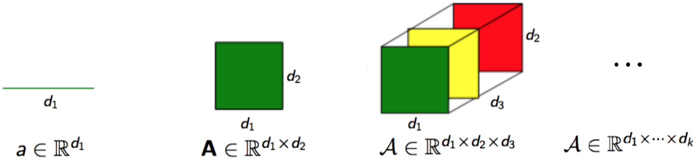
All Gaussian except points 17 and 39.

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 $\mathcal{A} = [\![a_{i_1 \ldots i_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ is a hypermatrix with dimensions $(d_1, \ldots, d_k)$ and entries $a_{i_1 \ldots i_k} \in \mathbb{R}$.
- 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:

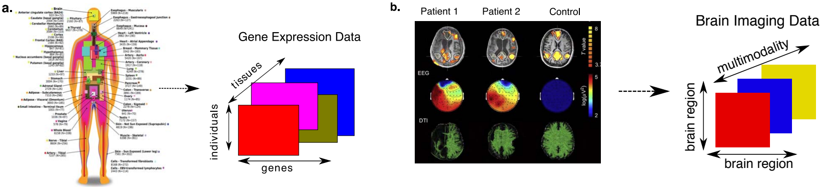
- Longitudinal social network data $\{\boldsymbol{Y}_t : t = 1, ..., n\}$
- Spatio-temporal transcriptome data
- Joint probability table of a set of variables $\mathbb{P}(X_1, X_2, X_3)$
- Higher-order moments in topic models
- Markov models for the phylogenetic tree

W & Song 2017, P. Hoff 2015, Montanari & Richard 2014

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

Tensors in biomedical science

- Many datasets come naturally in a multiway form.
- Multi-tissue, multi-individual gene expression data could be organized as an order-3 tensor $\mathcal{Y} = [\![y_{ijk}]\!] \in \mathbb{R}^{n_I \times n_G \times n_T}$.
- Multi-individual, multi-modal brain connectivity data could be organized as an order-4 tensor $\mathcal{Y} = [\![y_{ijkl}]\!] \in \mathbb{R}^{n_I \times n_b \times n_b \times n_m}$.

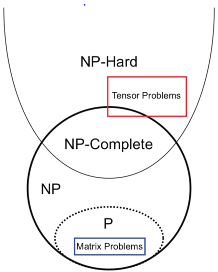


W., Fischer, et al, 2019, W. & Li, 2020

Talk outline

Prohibitive Computational Complexity

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



Fortunately, tensors sought in statistical and machine learning applications are often specially structured:

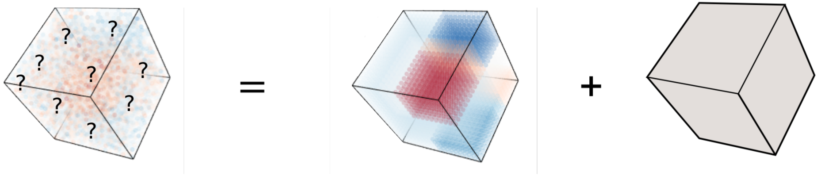
- Low-rankness
- Sparsity
- Non-negativity
- ...

This talk is based on

Beyond the Signs: Nonparametric Tensor Completion via Sign Series. Lee and W. 2021

https://arxiv.org/pdf/2102.00384.pdf

Setup: signal plus noise model



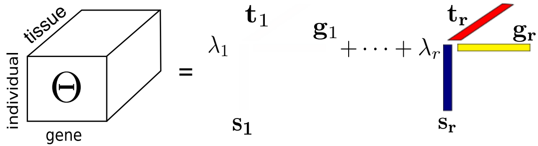
$$\overbrace{\mathcal{Y}}^{}_{\text{Data tensor}} \qquad = \qquad \overbrace{\Theta}^{}_{\text{Signal tensor}} \qquad + \qquad \overbrace{\mathcal{E}}^{}_{\text{Noise tensor}}$$

We focus on two problems:

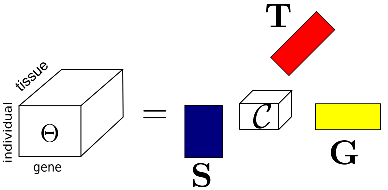
- Nonparametric tensor estimation: How to estimate the signal tensor Θ under a wide range of structures?
- Tensor completion: How many observed tensor entries do we need in order for consistent recovery?

Various notions of low-rankness

- Canonical polyadic (CP) low-rankness [Kolda & Bader '09]: $\Theta = \sum_{r=1}^{R} \lambda_r \boldsymbol{s}_r \otimes \boldsymbol{g}_r \otimes \boldsymbol{t}_r \in \mathbb{R}^{d_1 \times d_2 \times d_3}$.



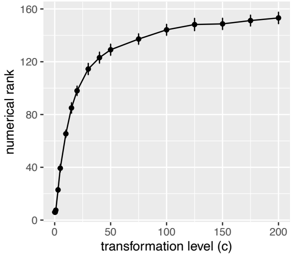
- Tucker low-rankness [Lathauwer '00]: $\Theta = \mathcal{C} \times_1 \boldsymbol{S} \times_2 \boldsymbol{G} \times_3 \boldsymbol{T} \in \mathbb{R}^{d_1 \times d_2 \times d_2}$.



- Others: tensor train model [Oseledet '11], tensor block model [W. & Zeng '19; Han, Luo, W. et al '20], etc.

Inadequacies of low-rank models

- Tensor rank is sensitive to order-preserving transformation.

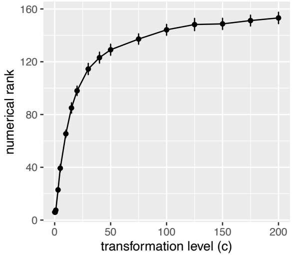


$$\Theta = \frac{1}{1+\exp(-c\mathcal{Z})}, \quad \text{where}$$
$$\mathcal{Z} = \boldsymbol{a}^{\otimes 3} + \boldsymbol{b}^{\otimes 3} + \boldsymbol{c}^{\otimes 3}$$

$\Rightarrow$ Θ is high-rank but $\mathcal{Z}$ is low-rank.

Inadequacies of low-rank models

- Tensor rank is sensitive to order-preserving transformation.

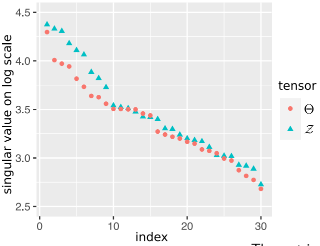


$$\Theta = \frac{1}{1+\exp(-c\mathcal{Z})}, \quad \text{where}$$

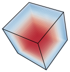
$$\mathcal{Z} = \boldsymbol{a}^{\otimes 3} + \boldsymbol{b}^{\otimes 3} + \boldsymbol{c}^{\otimes 3}$$

$\Rightarrow$ Θ is high-rank but $\mathcal{Z}$ is low-rank.

- Low-rank model fails to address several important structures.



$$\Theta = \text{entrywise polynomial of } \mathcal{Z} \in \mathbb{R}^{d\times d\times d}$$

$$\mathcal{Z} = [\tfrac{1}{d}\max(i,j,k)] =$$ 

$\Rightarrow$ Both Θ and $\mathcal{Z}$ are full rank.

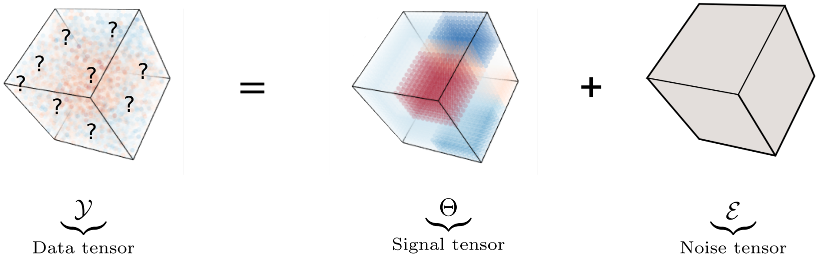
The matrix analogy of Θ was studied in the context of graphon analysis by Chan & Airoldi '14.

Why sign matters?

For a bounded tensor $\Theta \in [-1,1]^{d_1 \times \cdots \times d_K}$,

$$\Theta \approx \frac{1}{|\mathcal{H}|} \sum_{\pi \in \mathcal{H}} \text{sgn}(\Theta - \pi), \quad \text{where} \quad \mathcal{H} = \left\{ -1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1 \right\}.$$

- We do not observe Θ; instead, we observe a noisy incomplete version $\mathcal{Y}$.
- How to estimate the signal tensor Θ given data tensor $\mathcal{Y}$?



Sign rank

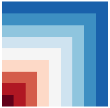
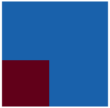
- Key ideas: we use a local (nonparametric) notion of "low-rankness" to allow a broad family of signal tensors.
- Two tensors are sign equivalent, denoted $\Theta \simeq \Theta'$, if $\text{sgn}(\Theta) = \text{sgn}(\Theta')$.
- Define the sign rank by

$$\text{srank}(\Theta) = \min\{\text{rank}(\Theta') \colon \Theta' \simeq \Theta,\ \Theta' \in \mathbb{R}^{d_1 \times \cdots \times d_K}\}.$$

Sign rank

- Key ideas: we use a local (nonparametric) notion of "low-rankness" to allow a broad family of signal tensors.
- Two tensors are sign equivalent, denoted $\Theta \simeq \Theta'$, if $\text{sgn}(\Theta) = \text{sgn}(\Theta')$.
- Define the sign rank by

$$\text{srank}(\Theta) = \min\{\text{rank}(\Theta') \colon \Theta' \simeq \Theta,\ \Theta' \in \mathbb{R}^{d_1 \times \cdots \times d_K}\}.$$

$\Theta =$  $,\quad \text{sgn}(\Theta) =$  $\Longrightarrow \begin{array}{l} \text{rank}(\Theta) = d \\ \text{srank}(\Theta) = 2 \end{array}$

- For any strictly monotonic function $g \colon \mathbb{R} \to \mathbb{R}$ with $g(0) = 0$,

$$\text{srank}(\Theta) \leq \text{rank}(g(\Theta)).$$

Sign representable tensors

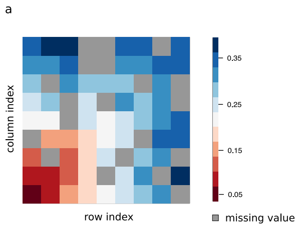
Sign representable tensors

A tensor Θ is called r-sign representable if the tensor $(\Theta - \pi)$ has sign rank bounded by r for all $\pi \in [-1, 1]$.

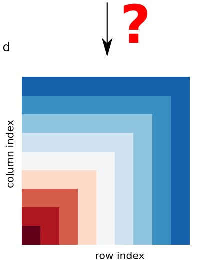
- Most existing structured tensors belong to sign representable family:
  - Low-rank CP tensors, Tucker tensors, stochastic tensor block models.
  - High-rank tensors from GLM, single index models.
  - Tensors with repeating patterns, e.g. earlier max/min hypergraphon model $\Theta(i_1, \ldots, i_K) = \log(1 + \max(i_1, \ldots, i_K))$ is 2-sign representable.
- We propose the signal tensor family

$$\Theta \in \mathscr{P}_{\text{sgn}}(r) := \{\Theta \colon \text{srank}(\Theta - \pi) \le r \text{ for all } \pi \in [-1, 1]\}.$$

Our solution: sign signal helps!

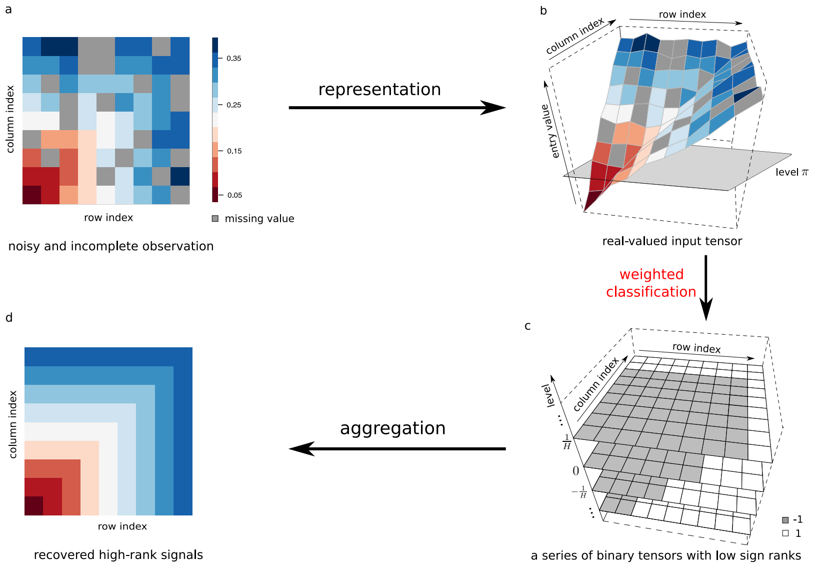


noisy and incomplete observation



recovered high-rank signals

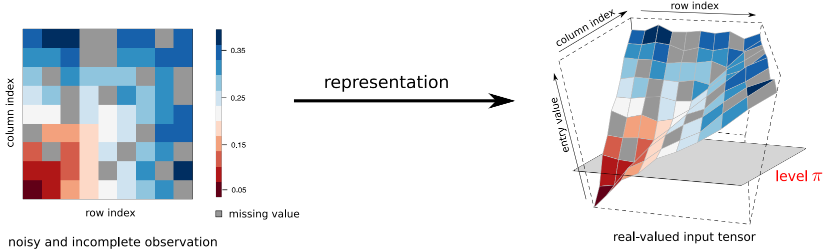
Our solution: sign signal helps!



Sign representation

- We observe a noisy incomplete tensor $\mathcal{Y}_\Omega$ with observed index set $\Omega \in [d_1] \times \cdots \times [d_K]$ under uniform sampling scheme.
- We dichotomize the data into a series of sign tensors:

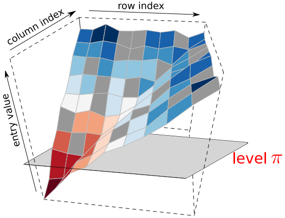
$$\{\text{sgn}(\mathcal{Y}_\Omega - \pi)\}_{\pi \in \mathcal{H}}, \quad \text{where } \mathcal{H} = \left\{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\right\}.$$



Sign estimation via weighted classification

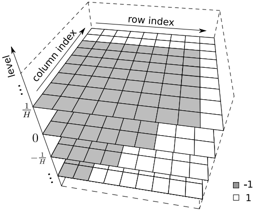
- We estimate $\text{sgn}(\Theta - \pi)$ through $\text{sgn}(\mathcal{Y}_\Omega - \pi)$ via weighted classification.
- Objective function of weighted classification is

$$L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi) = \frac{1}{|\Omega|} \sum_{\pi \in \Omega} \underbrace{|\mathcal{Y}(\omega) - \pi|}_{\text{weight}} \times \underbrace{|\text{sgn}(\mathcal{Z}(\omega)) - \text{sgn}(\mathcal{Y}(\omega) - \pi)|}_{\text{classification loss}}$$



real-valued input tensor

weighted classification $\longrightarrow$



a series of binary tensors with low sign ranks

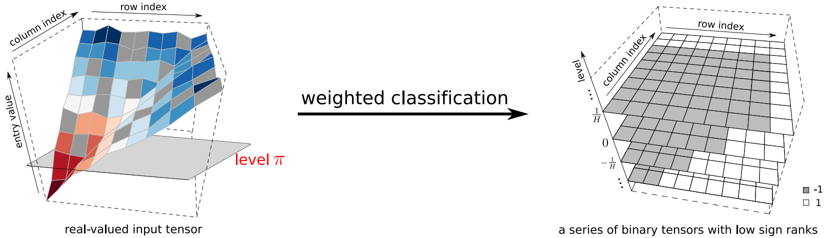
Identification for sign tensor estimation

- If $\Theta \in \mathscr{P}_{\textsf{sgn}}(r)$ is (locally) α-smooth $(\alpha \neq 0)$ at π, we have a unique optimizer such that

$$\textsf{sgn}(\Theta - \pi) = \underset{\mathcal{Z}\colon \textsf{rank}(\mathcal{Z})\leq r}{\arg\min}\ \mathbb{E}_{\mathcal{Y}_\Omega} L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi).$$

- We obtain a series of optimizers $\{\hat{\mathcal{Z}}_\pi\}_{\pi\in\mathcal{H}}$ as

$$\hat{\mathcal{Z}}_\pi = \underset{\mathcal{Z}\colon \textsf{rank}(\mathcal{Z})\leq r}{\arg\min}\ L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi).$$



* Uniqueness up to sign equivalence, meaning the optimizer $\Theta_{\textsf{opt}} \simeq \textsf{sgn}(\Theta - \pi)$.

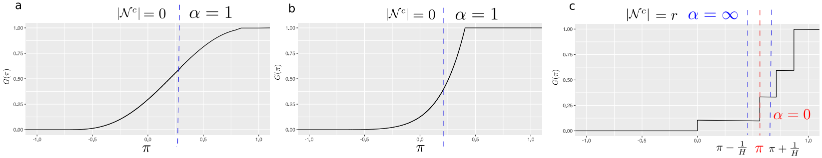
Identification for sign tensor estimation

We quantify smoothness of Θ using CDF $G(\pi) = \mathbb{P}_{\omega\sim\Pi}(\Theta(\omega) \leq \pi)$ and the induced pseudo density (i.e., histogram with bin size $\Delta s = d^{-K}$).

Identification for sign tensor estimation

We quantify smoothness of Θ using CDF $G(\pi) = \mathbb{P}_{\omega\sim\Pi}(\Theta(\omega) \le \pi)$ and the induced pseudo density (i.e., histogram with bin size $\Delta s = d^{-K}$).

Intuition: sign recovery is harder at levels where point mass concentrates.



α-smoothness of signal tensor

- Partition $[-1,1] = \mathcal{N} \cup \mathcal{N}^c$, where $\mathcal{N}$ consists of levels whose pseudo density based on Δs-bin is uniformly bounded, and $\mathcal{N}^c$ otherwise.
- $G(\pi)$ is (globally) α-smooth in that for all $\pi \in \mathcal{N}$,

$$\sup_{\Delta s \le t < \rho(\pi, \mathcal{N}^c)} \frac{G(\pi + t) - G(\pi - t)}{t^{\alpha}} \le c$$

for two constants $\alpha > 0, c > 0$, where $\rho(\pi, \mathcal{N}^c) = \min_{\pi' \in \mathcal{N}^c} |\pi - \pi'| + \Delta s$.

Sign tensor estimation error

- For two tensor Θ_1, Θ_2, define $\text{MAE}(\Theta_1, \Theta_2) = \mathbb{E}_{\omega \in \Pi} |\Theta_1(\omega) - \Theta_2(\omega)|$.

Sign tensor estimation for $\pi \in \mathcal{N}$ (Lee and W. 2021)

Suppose $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ is α-smooth. Let $d_{\max} = \max_{k \in [K]} d_k$. Then, with very high probability over $\mathcal{Y}_\Omega$, we have uniform error bound over $\pi \in \mathcal{N}$,

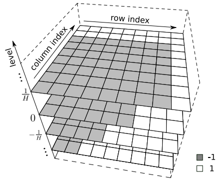
$$\text{MAE}(\text{sgn}\hat{\mathcal{Z}}_\pi, \text{sgn}(\Theta - \pi)) \lesssim C(\pi) \left(\frac{d_{\max} r}{|\Omega|} \right)^{\frac{\alpha}{\alpha+2}}.$$

- Sign estimation error shows a polynomial decay with $|\Omega|$.
- Best rate is attained at $\alpha = \infty$ for stochastic tensor block models.

From sign to signal estimation

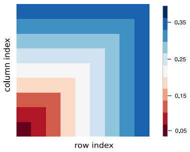
- Aggregation of sign tensors from weighted classification yields our signal tensor estimate:

$$\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \text{sgn}\hat{\mathcal{Z}}_{\pi}.$$



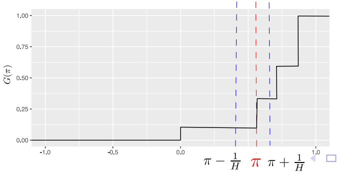
a series of binary tensors with low sign ranks

aggregation



recovered high-rank signals

- Signal tensor estimation is robust to a few off-target classifications.



Signal tensor estimation error

Tensor estimation error (Lee and W. 2021)

Suppose $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ is α-smooth with bounded $|\mathcal{N}^c|$. Then, with very high probability over $\mathcal{Y}_\Omega$,

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim \underbrace{\left(\frac{d_{\max} r}{|\Omega|}\right)^{\frac{\alpha}{\alpha+2}}}_{\text{Error inherited from sign estimation}} + \underbrace{\frac{1}{H}}_{\text{Bias}} + \underbrace{\frac{H d_{\max} r}{|\Omega|}}_{\text{Variance}}.$$

In particular, setting $H \asymp \left(\frac{|\Omega|}{d_{\max} r}\right)^{1/2}$ yields the error bound

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim \left(\frac{d_{\max} r}{|\Omega|}\right)^{\frac{\alpha}{\alpha+2} \wedge \frac{1}{2}}.$$

- See papers for general results that allow unbounded $|\mathcal{N}^c|$ and sub-Gaussian noise.

extension

Comparison to existing results

- Sample requirement for tensor completion:

$$\text{MAE}(\hat{\Theta}, \Theta) \to 0, \text{ as } \frac{|\Omega|}{d_{\max} r} \to \infty.$$

- Special case with full observation and equal dimension $d_1 = \cdots = d_K = d$:

Model	Our rate (power of d)	Previous results
Tensor block model	$-(K-1)/2$	$\alpha = \infty$; minimax rate in W. & Zeng '19
Single index model	$-(K-1)/3$	$\alpha = 1$; conjecture on the optimality; matrix rate $d^{-1/3}$ improves $\mathcal{O}(d^{-1/4})$ by Ganti et al. '18.
Rader type tensor	$-(K-2)/2$	close to parametric rate
α-smooth $\mathscr{P}_{\text{sgn}}(r)$	$-(K-1)\min(\frac{\alpha}{\alpha+2} \wedge \frac{1}{2})$	faster rate as α increases

Our contributions and related work

Our contributions

- We develop a new model called sign representable tensors to fill the gap between parametric (low-rank) and nonparametric (high-rank) tensors.

Our contributions and related work

Our contributions

- We develop a new model called sign representable tensors to fill the gap between parametric (low-rank) and nonparametric (high-rank) tensors.
- Our tensor estimate is provably reducible to a series of classifications. $\Rightarrow$ computationally efficient via a divide-and-conquer algorithm.

Our contributions and related work

Our contributions

- We develop a new model called sign representable tensors to fill the gap between parametric (low-rank) and nonparametric (high-rank) tensors.
- Our tensor estimate is provably reducible to a series of classifications. $\Rightarrow$ computationally efficient via a divide-and-conquer algorithm.

- **Low-rank tensor estimation** (Anandkumar et al. '14; Montanari & Sun '18; Cai et al. '19) $\Rightarrow$ low-rank assumption is often violated in practice.

Our contributions and related work

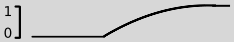
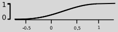
Our contributions

- We develop a new model called sign representable tensors to fill the gap between parametric (low-rank) and nonparametric (high-rank) tensors.
- Our tensor estimate is provably reducible to a series of classifications. $\Rightarrow$ computationally efficient via a divide-and-conquer algorithm.

- **Low-rank tensor estimation** (Anandkumar et al. '14; Montanari & Sun '18; Cai et al. '19) $\Rightarrow$ low-rank assumption is often violated in practice.
- **High-rank matrix estimation** was studied under nonlinear models (Ganti et al. '15), permutation rank (Shah et al. '18), and other shape constraints (Chatterjee et al. '19) $\Rightarrow$ tensors are more challenging because tensor rank may exceed dimension.

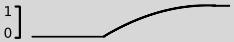
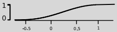
Comparison of estimation error versus dimension

- We simulate signal tensors under a wide range of complexity.

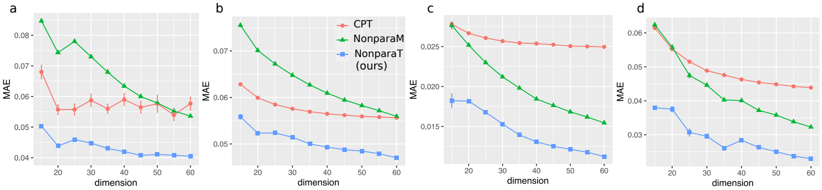
Simulation	Signal Tensor Θ	Rank	Sign Rank	α	$\lvert\mathcal{N}^c\rvert$	CDF	Noise
1	$\mathcal{C} \times \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \boldsymbol{M}_3$	3^3	$\leq 3^3$	∞	$\leq 3^3$		Uniform $[-0.3, 0.3]$
2	$\lvert \boldsymbol{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \boldsymbol{a} \otimes \mathbf{1} \rvert$	d	≤ 3	1	0		Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + Z_{\max})$	$\geq d$	2	∞	d		Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(\mathcal{Z}_{\min}^{1/3})$	$\geq d$	2	∞	d		Normal $\mathcal{N}(0, 0.15)$

Comparison of estimation error versus dimension

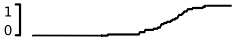
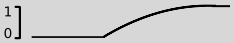
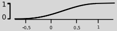
- We simulate signal tensors under a wide range of complexity.

Simulation	Signal Tensor Θ	Rank	Sign Rank	α	$\lvert\mathcal{N}^c\rvert$	CDF	Noise
1	$\mathcal{C}\times \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \boldsymbol{M}_3$	3^3	$\leq 3^3$	∞	$\leq 3^3$		Uniform $[-0.3, 0.3]$
2	$\lvert \boldsymbol{a}\otimes \mathbf{1}\otimes \mathbf{1} - \mathbf{1}\otimes \boldsymbol{a}\otimes \mathbf{1}\rvert$	d	≤ 3	1	0		Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + \mathcal{Z}_{\max})$	$\geq d$	2	∞	d		Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(\mathcal{Z}_{\min}^{1/3})$	$\geq d$	2	∞	d		Normal $\mathcal{N}(0, 0.15)$

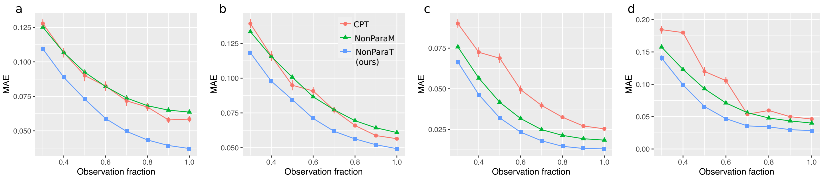
- Our method (NonparaT) achieves the best performance, whereas the second best method is low-rank CP tensor (CPT) for models 1-2, and matrix nonparametric method (NonParaM) for models 3-4.



Estimation error versus observation fraction

Simulation	Signal Tensor Θ	Rank	Sign Rank	α	$\lvert\mathcal{N}^c\rvert$	CDF	Noise
1	$\mathcal{C} \times \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \boldsymbol{M}_3$	3^3	$\leq 3^3$	∞	$\leq 3^3$		Uniform $[-0.3, 0.3]$
2	$\lvert \boldsymbol{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \boldsymbol{a} \otimes \mathbf{1} \rvert$	d	≤ 3	1	0		Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + Z_{\max})$	$\geq d$	2	∞	d		Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(\mathcal{Z}_{\min}^{1/3})$	$\geq d$	2	∞	d		Normal $\mathcal{N}(0, 0.15)$

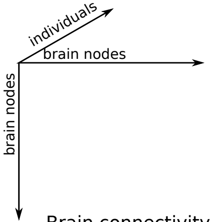
- Our method (NonparaT) achieves the best performance in completion.



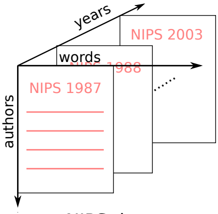
Data application

We apply our method to three datasets:

- The NIPS dataset (Globerson et al '07) consists of word occurrence counts in papers published from 1987 to 2003.
- The human brain connectivity data (Want et al '17) consists of 68 brain regions for 114 individuals along with their IQ scores.
- The 3-channel image data is from licensed google image file.



Brain connectivity



NIPS data

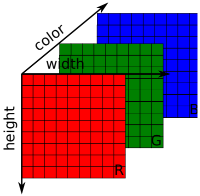
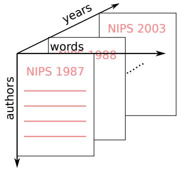


Image data

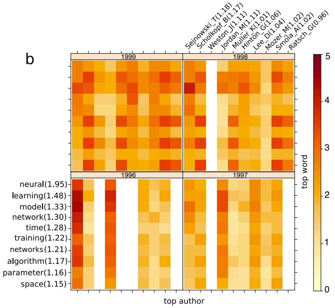
Data application: NIPS



NIPS data

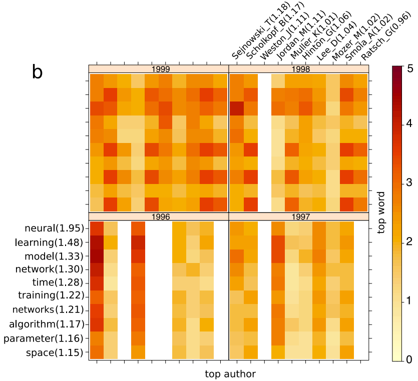
- The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003.
- Data tensor $\mathcal{Y} \in \mathbb{R}^{100\times 200\times 17}$.

- We examine the estimated signal tensor $\hat{\Theta}$.
- Most frequent words are consistent with active topics in NIPS conference.
- Strong heterogeneity among word occurrences across authors and years.



Data application: NIPS

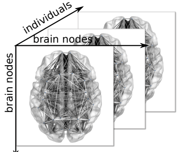
- Top words: *neural* (1.95), *learning* (1.48), *network* (1.21), *training* (1.22), *parameter (1.16)*.
- Top authors: *T. Sejnowski* (1.18), *B. Schölkopf* (1.17), *M. Jordan* (1.11), and *G. Hinton* (1.06),
- Top combinations:
  (*training*, *algorithm*) $\times$ (*B. Schölkopf*, *A. Smola*) $\times$ (1998, 1999),
  (*model*) $\times$ (*M. Jordan*) $\times$ (1996).
- Similar word patterns in 1998-1999: (*B. Schölkopf*, *K-R Möller*, *A. Smola*, *G. Rätsch*, *J. Weston*) $\Rightarrow$ Co-authors.



Numbers in parentheses denote marginal averages based on $\hat{\Theta}$

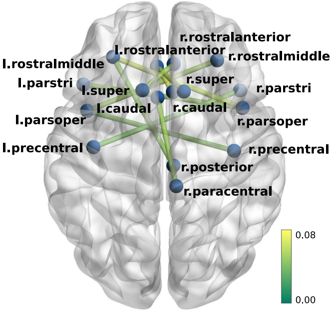
details

Data application: Brain connectivity



- The human brain connectivity data consists of 68 brain regions for 114 individuals along with their IQ scores.
- Data tensor $\mathcal{Y} \in \{0,1\}^{68\times 68\times 114}$.

- We examine the estimated signal tensor $\hat{\Theta}$.
- Top 10 brain edges based on regression analysis show inter-hemisphere connections.



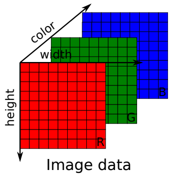
Data application: Brain connectivity + NIPS

- Our method achieves better test performance than low-rank methods.

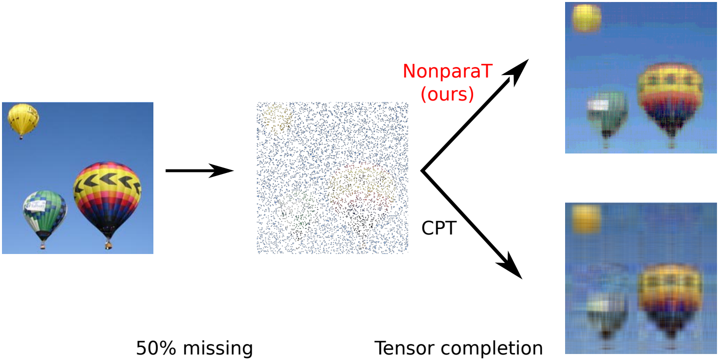
MRN-114 brain connectivity dataset					
Method	$r = 3$	$r = 6$	$r = 9$	$r = 12$	$r = 15$
NonparaT (Ours)	**0.18**(0.001)	**0.14**(0.001)	**0.12**(0.001)	**0.12**(0.001)	**0.11**(0.001)
Low-rank CPT	0.26(0.006)	0.23(0.006)	0.22(0.004)	0.21(0.006)	0.20(0.008)
NIPS word occurrence dataset					
Method	$r = 3$	$r = 6$	$r = 9$	$r = 12$	$r = 15$
NonparaT (Ours)	**0.18**(0.002)	**0.16**(0.002)	**0.15**(0.001)	**0.14**(0.001)	**0.13**(0.001)
Low-rank CPT	0.22(0.004)	0.20(0.007)	0.19(0.007)	0.17(0.007)	0.17(0.007)
Naive imputation (Baseline)	0.32(.001)				

Table: MAE comparison in the brain data and NIPS data based on 5-folded cross-validations. Standard errors are reported in parenthesis.

Data application: Image



- The original data is from licensed google image file.
- Data tensor $\mathcal{Y} \in [0,1]^{217\times 217\times 3}$.
- We assess completion performance by sampling 50% entries in the original image tensor.



Summary

Tensor analysis provides a rich source of

- fundamental problems in data science.
- new tools for long-standing questions.
- potentials for new applications.

Our general strategy is to develop efficient statistical methods for analyzing a broad range of specially-structured tensors that are useful in practice.

References:

- Beyond the signs: nonparametric tensor completion from sign series. https://arxiv.org/pdf/2102.00384.pdf

Acknowledgment: NSF DMS-1915978, DMS-2023239, and Grant from Wisconsin Alumni Research Foundation.

Appendix

Algorithm 1 Nonparametric tensor completion

Input: Noisy and incomplete data tensor $\mathcal{Y}_\Omega$, rank r, resolution parameter H.

for $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$ **do**

Random initialization of tensor factors $\boldsymbol{A}_k = [\boldsymbol{a}_1^{(k)}, \ldots, \boldsymbol{a}_r^{(k)}] \in \mathbb{R}^{d_k \times r}$ for all $k \in [K]$.

while not convergence **do**

for $k = 1, \ldots, K$ **do**

Update $\boldsymbol{A}_k$ while holding others fixed:

$\boldsymbol{A}_k \leftarrow \arg\min_{\boldsymbol{A}_k \in \mathbb{R}^{d_k \times r}} \sum_{\omega \in \Omega} |\mathcal{Y}(\omega) - \pi| F(\mathcal{Z}(\omega) \text{sgn}(\mathcal{Y}(\omega) - \pi))$,

where $F(\cdot)$ is the large-margin loss, and $\mathcal{Z} = \sum_{s \in [r]} \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}$ is a rank-r tensor.

end for

end while

Return $\mathcal{Z}_\pi \leftarrow \sum_{s \in [r]} \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}$.

end for

Output: Estimated signal tensor $\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \text{sgn}(\mathcal{Z}_\pi)$.

Paper titles in NIPS dataset

1998:

- Kernel PCA and De-Noising in Feature Spaces. Sebastian Mika, **B. Schölkopf**, **A. Smola**, **K. Müller**, M. Scholz, **G. Rätsch**
- Shrinking the Tube: A New Support Vector Regression Algorithm. **B. Schölkopf**, P. Bartlett, **A. Smola**, R. C. Williamson
- Semiparametric Support Vector and Linear Programming Machines. **A. Smola**, T-T. Frieb, **B. Schölkopf**
- Regularizing AdaBoost. **G. Rätsch**, T. Onoda, **K. Müller**

1999:

- v-Arc: Ensemble Learning in the Presence of Outliers. **G. Rätsch**, **B. Schölkopf**, **A. Smola**, **K. Müller**, T. Onoda, S. Mika
- Invariant Feature Extraction and Classification in Kernel Spaces. S. Mika, **G. Rätsch**, **J. Weston**, **B. Schölkopf**, **A. Smola**, **K Müller**
- Support Vector Method for Novelty Detection. **B. Schölkopf**, R. C. Williamson, **A. Smola**, J. S. Taylor, J. Platt.

1996:

- A Variational Principle for Model-based Morphing. L. Saul, **M. Jordan**
- Hidden Markov Decision Trees. **M. Jordan**, Z. Ghahramani, L. Saul
- Triangulation by Continuous Embedding. M. Meila, **M. Jordan**
- Recursive Algorithms for Approximating Probabilities in Graphical Models. T. Jaakkola, **M. Jordan**

Back to NIPS analysis

Extension to sub-Gaussian noise

Consider the signal plus noise model

$$\mathcal{Y} = \Theta + \mathcal{E},$$

where $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ is an α-smooth tensor with $\|\Theta\|_\infty \leq \beta$, and $\mathcal{E}$ consists of independent sub-Gaussian noise with variance proxy σ^2.

Nonparametric tensor estimation with sub-Gaussian noise

With high probability over $\mathcal{Y}_\Omega$, we have

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim \left(\frac{(\sigma^2 + \beta^2) r d_{\max} \log d_{\max}}{|\Omega|} \right)^{\frac{\alpha}{\alpha+2} \vee \frac{1}{2}}.$$

Back to bounded result

Tensor estimation error (Lee and W. 2021)

Consider a signal plus noise model with α-smooth signal $\Theta \in \mathscr{P}_{\text{sgn}}(r)$. Let $\hat{\Theta}$ be the our nonparametric estimate, and $|\mathcal{N}^c|$ the covering number of $\mathcal{N}^c$ with Δs-bin's, i.e, $|\mathcal{N}^c| = \text{Leb}(\mathcal{N}^c)/\Delta s$. Write $t_d = d_{\max} r/|\Omega|$. With very high probability over $\mathcal{Y}_\Omega$,

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim \underbrace{(t_d)^{\alpha/(\alpha+2)}}_{\text{Error inherited from sign estimation}} + \underbrace{\frac{1 + |\mathcal{N}^c|}{H}}_{\text{Bias}} + \underbrace{H t_d}_{\text{Bariance}}.$$

In particular, setting $H \asymp ((1+|\mathcal{N}^c|)/t_d)^{1/2}$ yields the error bound

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim \max\left(t_d^{2\alpha/(\alpha+2)},\ t_d(1 + |\mathcal{N}^c|)\right)^{1/2}.$$

Back to bounded result