

Nonparametric Tensor Completion via Sign Series

Chanwoo Lee

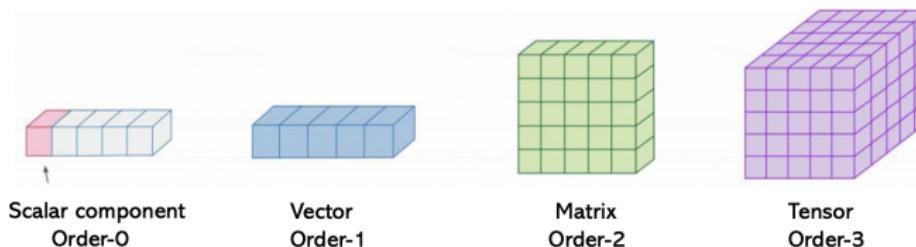
Joint work with Miaoyan Wang

Department of Statistics
University of Wisconsin - Madison

992 Seminar, Feb 3, 2021

Introduction: what is a tensor?

- ▶ Tensors are generalizations of vectors and matrices:

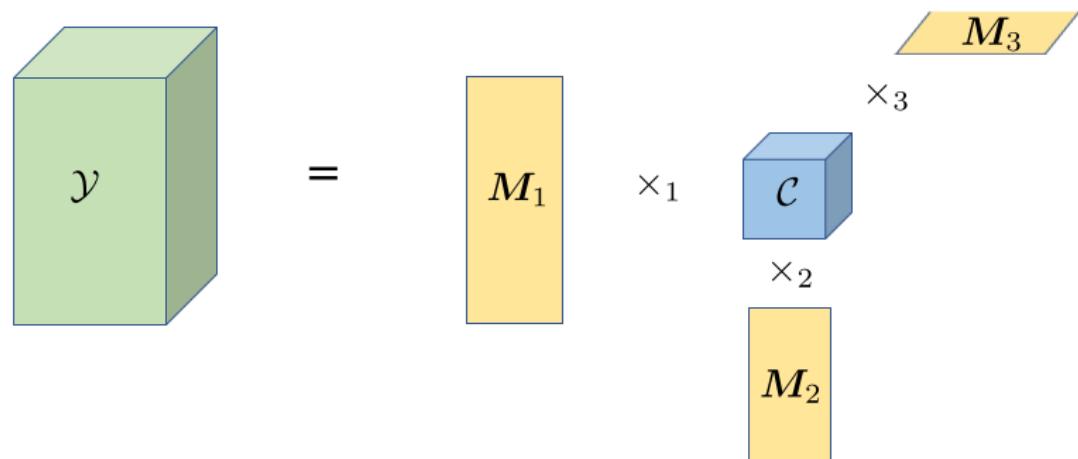


- ▶ We focus on tensors of order 3 or greater, also called **higher-order tensors**.
- ▶ Denote an order- $K(d_1, \dots, d_K)$ dimensional tensor as $\mathcal{Y} = [\![y_\omega]\!] \in \mathbb{R}^{d_1 \times \dots \times d_K}$ where $\omega \in [d_1] \times \dots \times [d_K]$.

Introduction: Tucker decomposition

- ▶ Tucker decomposition (De Lathauwer et al., 2000).
 - ▶ $\mathcal{Y} = \mathcal{C} \times_1 M_1 \times_2 M_2 \times_3 M_3$.
 - ▶ Generalization of matrix SVD to higher orders: $\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^T (= \Sigma \times_1 \mathbf{U} \times_2 \mathbf{V})$
 - ▶ Tucker rank of an order-3 tensor is defined as

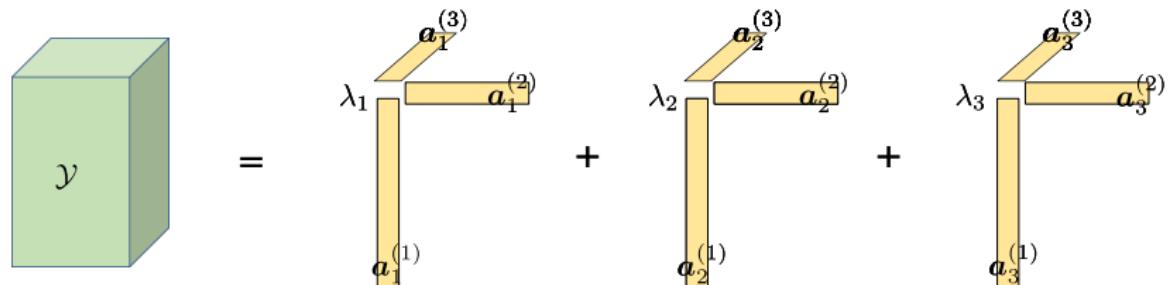
$$r(\mathcal{Y}) = (r_1, r_2, r_3).$$



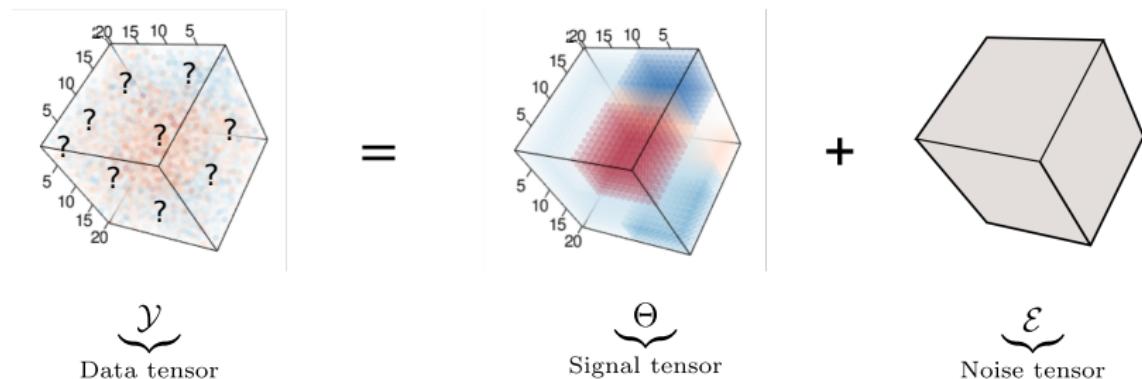
Introduction: Canonical Polyadic (CP) decomposition

- ▶ CP decomposition (Hitchcock, 1927).

- ▶ $\mathcal{Y} = \sum_{s=1}^r \lambda_s \mathbf{a}_s^{(1)} \otimes \cdots \otimes \mathbf{a}_s^{(K)}$.
- ▶ Generalization of matrix SVD to higher orders: $\mathbf{Y} = \sum_{s=1}^r \lambda_s \mathbf{u}_s \otimes \mathbf{v}_s$
- ▶ CP rank is defined as the minimal r for which the above equation holds.
- ▶ Today, we use the tensor rank as CP rank.



Main problems: the signal plus noise model



We focus on the two problems

1. **Nonparametric tensor estimation:** How to estimate the signal tensor Θ ?
2. **Complexity of tensor completion:** How many observed tensor entries do we need?

Classical approach

- ▶ Low rank models (Jain and Oh, 2014; Montanari and Sun, 2018).

$$\gamma \approx \lambda_1 \begin{matrix} a_1^{(3)} \\ a_1^{(2)} \\ a_1^{(1)} \end{matrix} + \lambda_2 \begin{matrix} a_2^{(3)} \\ a_2^{(2)} \\ a_2^{(1)} \end{matrix} + \cdots + \lambda_r \begin{matrix} a_r^{(3)} \\ a_r^{(2)} \\ a_r^{(1)} \end{matrix}$$

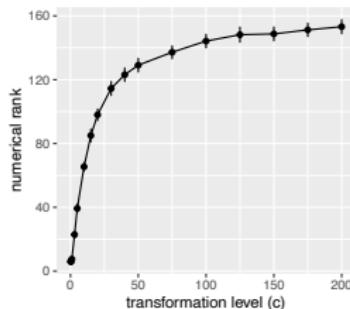
$\hat{\Theta}$

- ▶ There are two limitations of the model

1. The sensitivity to order-preserving transformation.
2. Inadequacy for special structures.

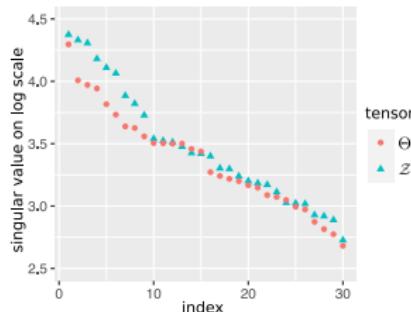
Classical approach

- ▶ The sensitivity to order-preserving transformation.



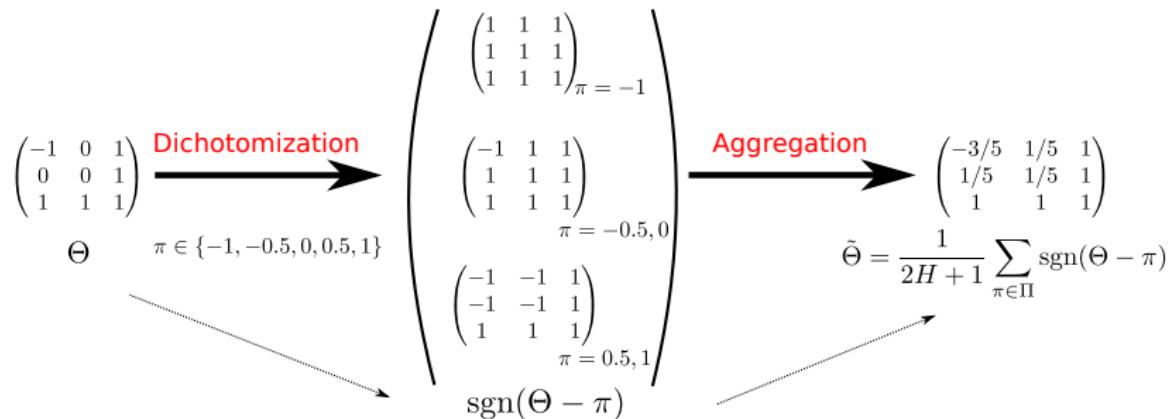
$$\Theta = \frac{1}{1 + \exp(-c(\mathcal{Z}))}, \quad \text{where}$$
$$\mathcal{Z} = \mathbf{a}^{\otimes 3} + \mathbf{b}^{\otimes 3} + \mathbf{c}^{\otimes 3}.$$

- ▶ Inadequacy for special structures.



$$\Theta = \log(1 + \mathcal{Z}), \quad \text{where}$$
$$\mathcal{Z}(i, j, k) = \frac{1}{d} \max(i, j, k).$$

Motivational toy example



where $\text{sgn}(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ -1 & \text{otherwise.} \end{cases}$

- We do not use magnitude of the signal tensor but **sign representation**.
- With a series of sign tensors, we can successfully preserve all information in the original signals.

Sign rank

- ▶ Two tensors are sign equivalent denoted as $\Theta \simeq \Theta'$ if $\text{sgn}(\Theta) = \text{sgn}(\Theta')$.
- ▶ Sign rank is defined as

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

ex) $\Theta = \begin{matrix} & \text{blue} \\ & \text{orange} \\ \text{red} & \text{yellow} \end{matrix}$, $\text{sgn}(\Theta) = \begin{matrix} & \text{blue} \\ & \text{blue} \\ \text{red} & \text{blue} \end{matrix} \implies \text{rank}(\Theta) = d, \text{srank}(\Theta) = 2$.

- ▶ In fact, for any strictly monotonic function $g: \mathbb{R} \rightarrow \mathbb{R}$ 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 $\pi \in [-1, 1]$. The collection $\{\text{sgn}(\Theta - \pi) : \pi \in [-1, 1]\}$ is called the sign tensor series.

ex1) $\Theta = \begin{pmatrix} -1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ is 2-sign representable.

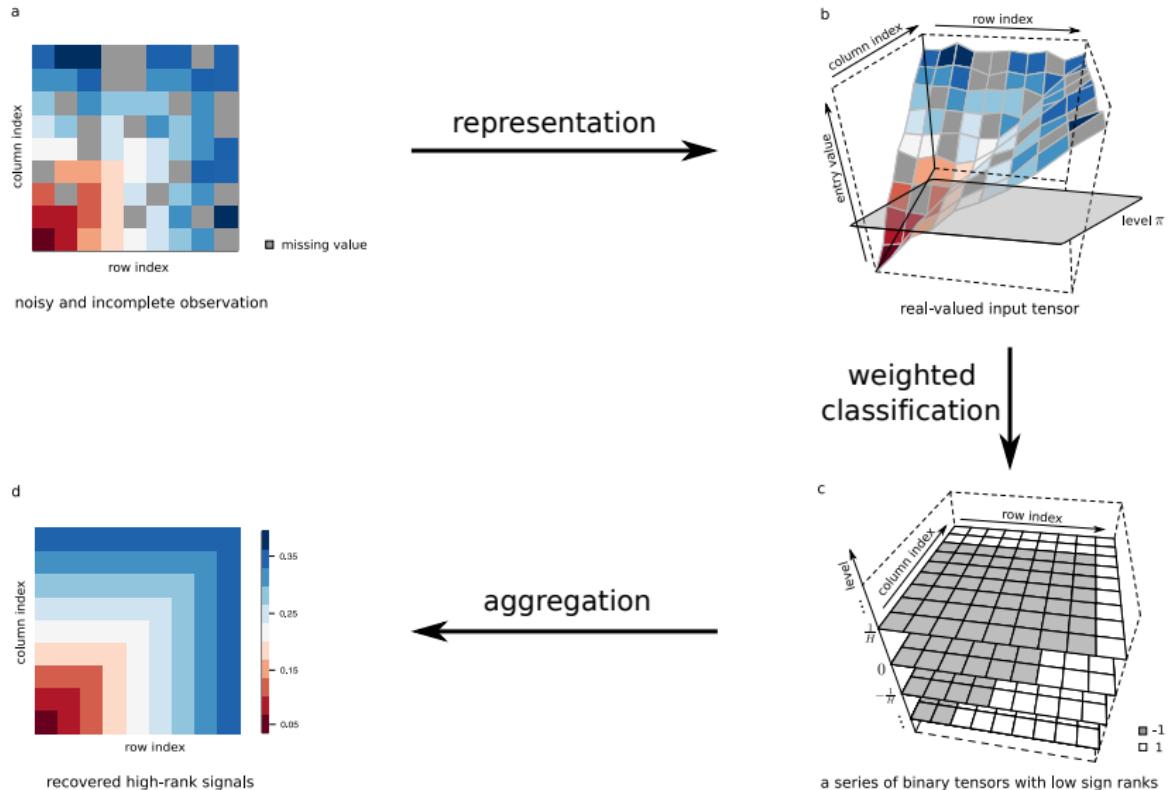
ex2) $\Theta(i_1, \dots, i_K) = \log(1 + \max(i_1, \dots, i_K))$ is 2-sign representable.

ex3) Θ such that $\text{rank}(\Theta) \leq r$, is $(r + 1)$ -sign representable.

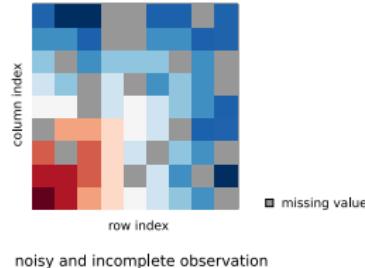
- ▶ Instead of the classical low rank assumption, we assume

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

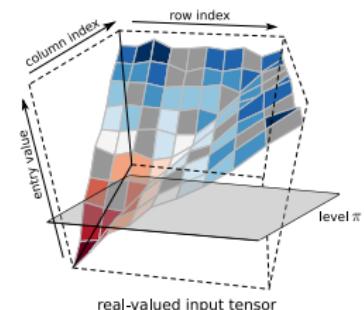
Our new approach



Our new approach: representation



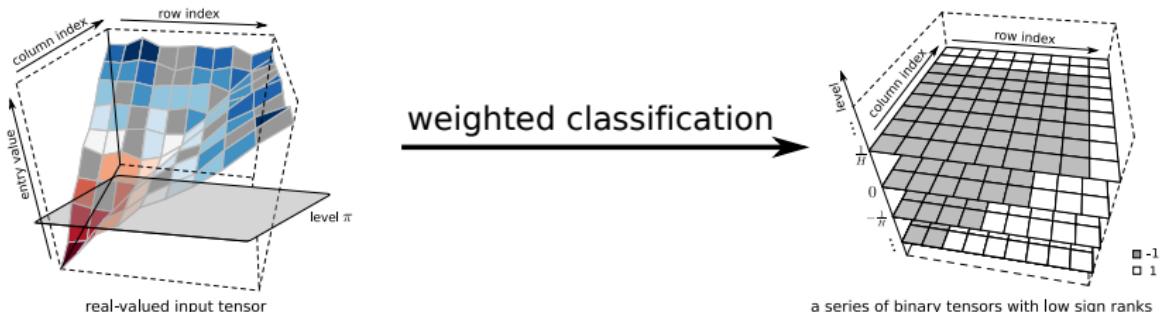
representation



- ▶ We are given the observed tensor $\mathcal{Y}_\Omega \in [-1, 1]^{d_1 \times \dots \times d_K}$ with observed index set $\Omega \in [d_1] \times \dots \times [d_K]$.
- ▶ We obtain sign tensor series

$$\{\text{sgn}(\mathcal{Y}_\Omega - \pi)\}_{\pi \in \mathcal{H}}, \quad \text{where } \mathcal{H} = \{-1, \dots, -1/H, 0, 1/H, \dots, 1\}.$$

Our new approach: weighted classification



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

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

Our new approach: weighted classification

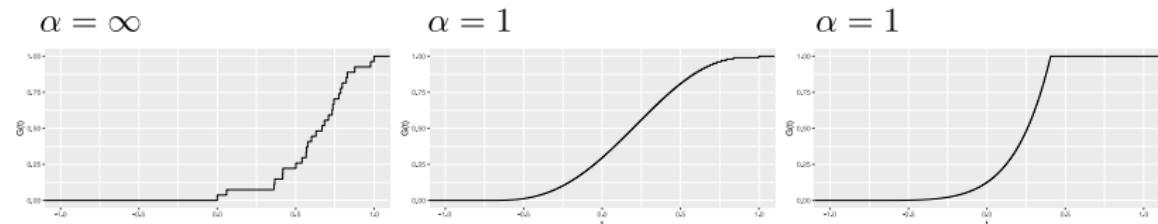
α smoothness

For fixed π , we call Θ is α smooth if there exist $\alpha = \alpha(\pi) > 0, c = c(\pi) > 0$, such that

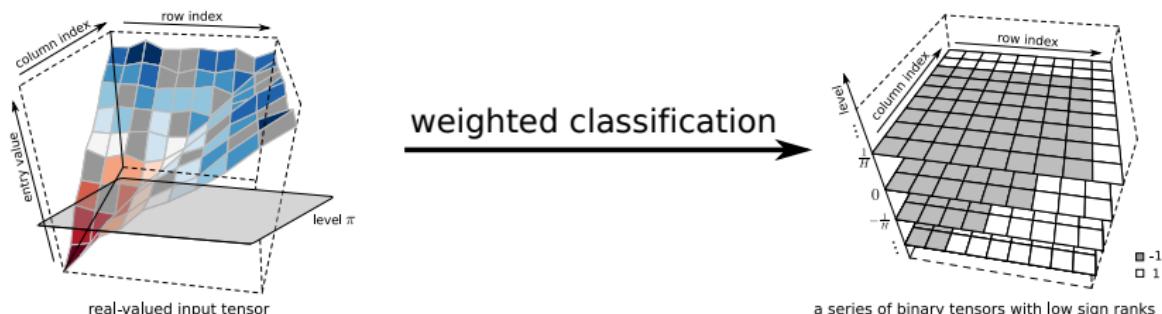
$$\sup_{0 \leq t < \rho(\pi, \mathcal{N})} \frac{\mathbb{P}_{\omega \sim \pi}[|\Theta(\omega) - \pi| \leq t]}{t} \leq c,$$

where $\rho(\pi, \mathcal{N}) = \min_{\pi' \in \mathcal{N}} |\pi - \pi'|$ and $\mathcal{N} = \{\pi : \mathbb{P}(\Theta(\omega) = \pi) \neq 0\}$. If constants α and c are independent of π , we call Θ is α -globally smooth.

ex) $G(t) := \mathbb{P}_{\omega \in \Pi}[\Theta(\omega) \leq t]$,



Our new approach: weighed classification



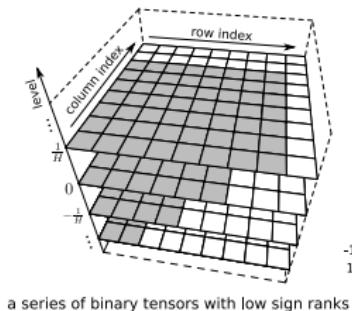
- If Θ is α smooth ($\alpha > 0$), we have **an unique optimizer** such that

$$\text{sgn}(\Theta - \pi) = \arg \min_{\mathcal{Z}: \text{rank}(\mathcal{Z}) \leq r} \mathbb{E}_{\mathcal{Y}_\Omega} L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi).$$

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

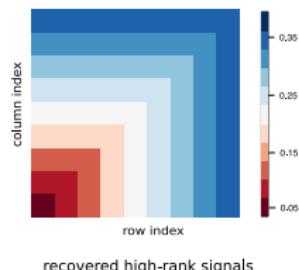
$$\hat{\mathcal{Z}}_\pi = \arg \min_{\mathcal{Z}: \text{rank}(\mathcal{Z}) \leq r} L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi).$$

Our new approach: aggregation



a series of binary tensors with low sign ranks

aggregation →



recovered high-rank signals

- ▶ From a series of optimizers $\{\hat{\mathcal{Z}}_\pi\}_{\pi \in \mathcal{H}}$ in the weighted classification, we propose the tensor estimate

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

Theoretical results: sign tensor estimation error

- ▶ For two tensors 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 fixed π (L. and Wang, 2021)

Suppose $\Theta \in \mathcal{P}_{\text{sgn}}(r)$ and $\Theta(\omega)$ is α smooth for fixed π . $d_{\max} = \max_{k \in [K]} d_k$. Then, with very high probability over \mathcal{Y}_Ω ,

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

Theoretical results: tensor estimation error

Tensor estimation error (L. and Wang 2021)

Suppose $\Theta \in \mathcal{P}_{\text{sgn}}(r)$ and $\Theta(\omega)$ is α -globally smooth. Then, with very high probability over \mathcal{Y}_Ω ,

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

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

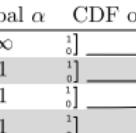
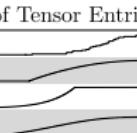
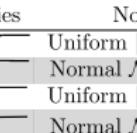
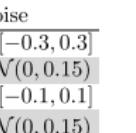
$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim \left(\frac{d_{\max} r}{|\Omega|} \right)^{\frac{\alpha}{\alpha+2} \vee \frac{1}{2}}.$$

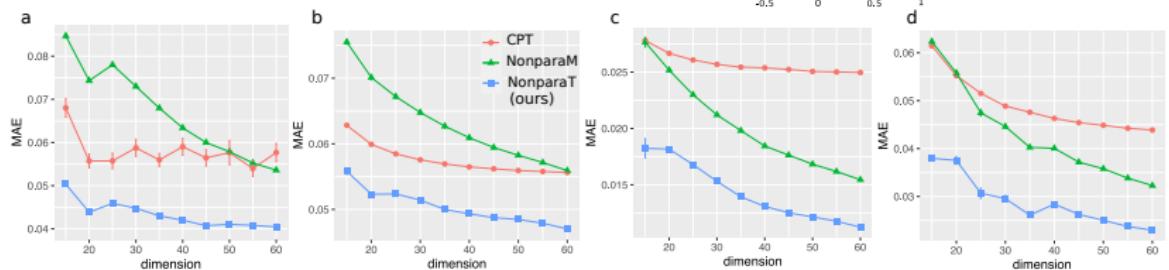
- ▶ Tensor estimation is generally no better than sign tensor estimation.
- ▶ Sample complexity:

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


18 / 26

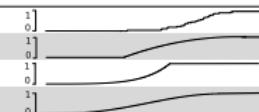
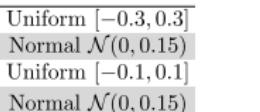
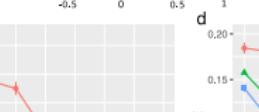
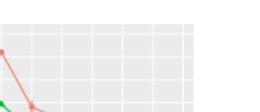
Simulations

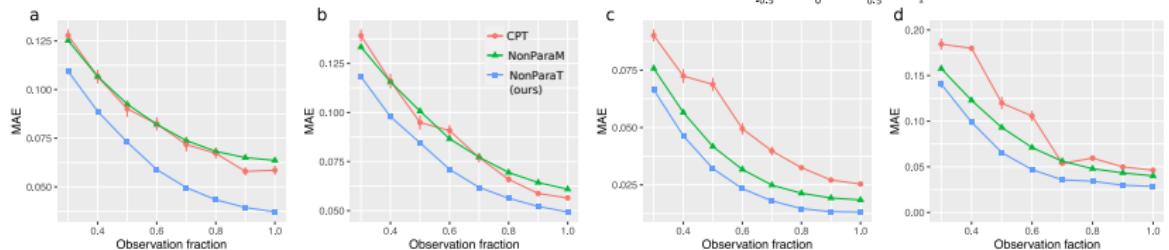
Simulation	Signal Tensor Θ	Rank	Sign Rank	Global α	CDF of Tensor Entries	Noise
1	$\mathcal{C} \times M_1 \times M_2 \times M_3$	3^3	$\leq 3^3$	∞	[]	Uniform $[-0.3, 0.3]$
2	$ \mathbf{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \mathbf{a} \otimes \mathbf{1} $	d	≤ 3	1	[]	Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + Z_{\max})$	$\geq d$	2	1	[]	Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(\mathcal{Z}_{\min}^{1/3})$	$\geq d$	2	1	[]	Normal $\mathcal{N}(0, 0.15)$



- ▶ **NonPraT**: Our nonparametric tensor method, **CPT**: low rank tensor CP decomposition, **NonPraraM**: the matrix version of our method.
- ▶ Estimation versus tensor dimension.
- ▶ Our method (NonparaT) achieves the best performance.

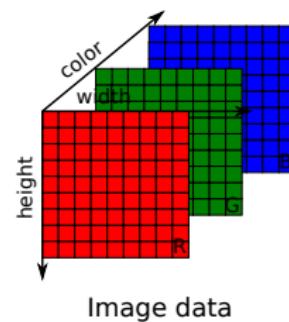
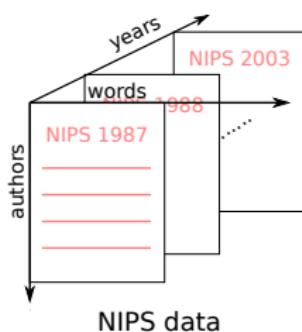
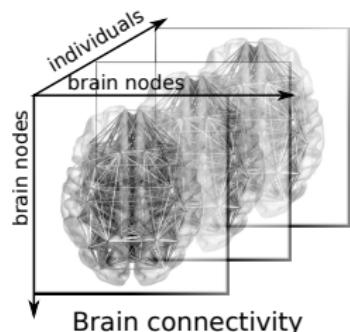
Simulations

Simulation	Signal Tensor Θ	Rank	Sign Rank	Global α	CDF of Tensor Entries	Noise
1	$\mathcal{C} \times M_1 \times M_2 \times M_3$	3^3	$\leq 3^3$	∞		Uniform $[-0.3, 0.3]$
2	$ \mathbf{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \mathbf{a} \otimes \mathbf{1} $	d	≤ 3	1		Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + Z_{\max})$	$\geq d$	2	1		Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(\mathcal{Z}_{\min}^{1/3})$	$\geq d$	2	1		Normal $\mathcal{N}(0, 0.15)$

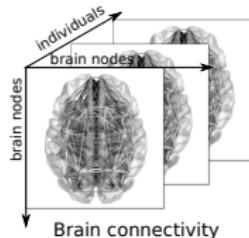


- ▶ Estimation error versus the observation fraction.
- ▶ Our method (NonparaT) achieves the best performance.

Data application

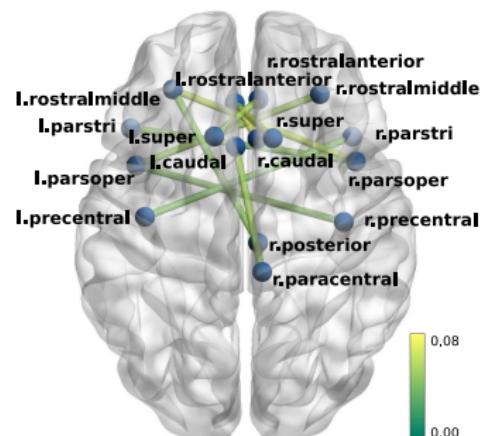


Data application: Brain connectivity

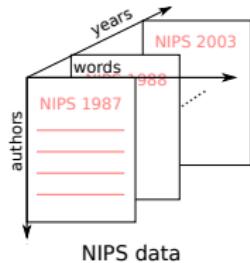


- ▶ The MRN-114 human brain connectivity data consists of 68 brain regions for 114 individuals along with their IQ scores (Wang et al., 2017).
- ▶ $\mathcal{Y} \in \{0, 1\}^{68 \times 68 \times 114}$.

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

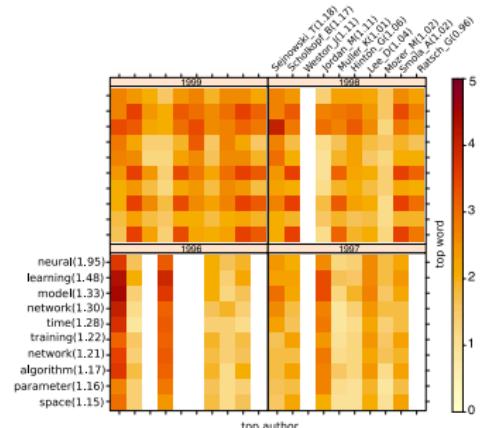


Data application: NIPS



- ▶ The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003 (Globerson et al., 2007).
- ▶ Log transformation yields the dataset $\mathcal{Y} \in \mathbb{R}^{100 \times 200 \times 17}$.

- ▶ We examine the estimated signal tensor $\hat{\Theta}$.
- ▶ Most frequent words is consistent with the active topics
- ▶ There are strong heterogeneity among word occurrences across authors and years.

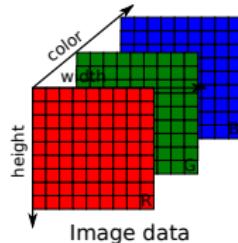


Data application: Brain connectivity + NIPS

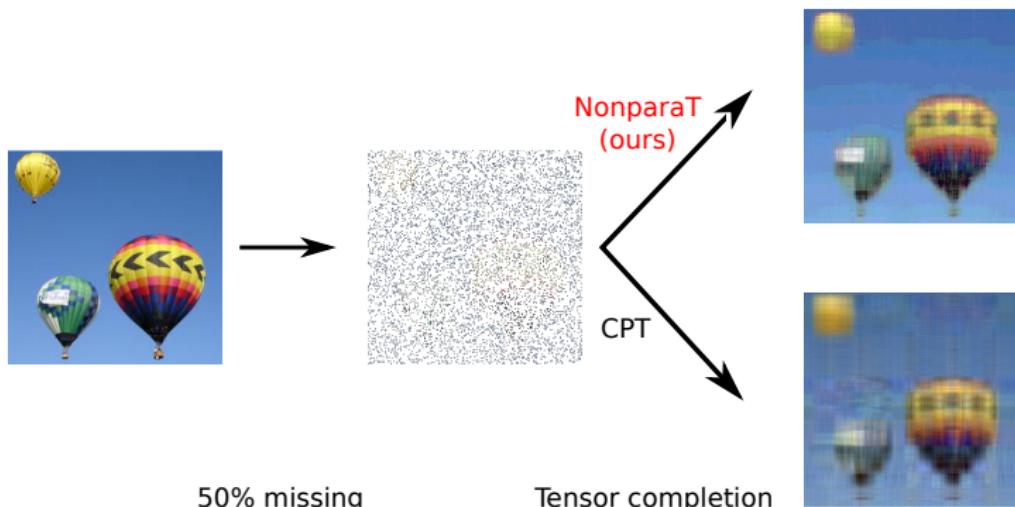
MRN-114 brain connectivity dataset					
Method	$r = 3$	$r = 6$	$r = 9$	$r = 12$	$r = 15$
NonparaT (Ours)	0.18(0.001)	0.14(0.001)	0.12(0.001)	0.12(0.001)	0.11(0.001)
Low-rank CPT	0.26(0.006)	0.23(0.006)	0.22(0.004)	0.21(0.006)	0.20(0.008)
NIPS word occurrence dataset					
Method	$r = 3$	$r = 6$	$r = 9$	$r = 12$	$r = 15$
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 on cross-validation (5 repetitions 5 folds). Standard errors are reported in parenthesis.

Data application: Image



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



Summary

- ▶ We have developed a completion method that can address both low- and high-rankness based on sign series representation.
- ▶ Estimation error rates and sample complexities are established.
- ▶ Our approach has good interpretation and prediction performance in both simulations and data applications.

Appendix

Algorithm 1 Nonparametric tensor completion

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

```
1: for  $\pi \in \mathcal{H} = \{-1, \dots, -\frac{1}{H}, 0, \frac{1}{H}, \dots, 1\}$  do
2:   Random initialization of tensor factors  $\mathbf{A}_k = [\mathbf{a}_1^{(k)}, \dots, \mathbf{a}_r^{(k)}] \in \mathbb{R}^{d_k \times r}$  for all  $k \in [K]$ .
3:   while not convergence do
4:     for  $k = 1, \dots, K$  do
5:       Update  $\mathbf{A}_k$  while holding others fixed:
6:        $\mathbf{A}_k \leftarrow \arg \min_{\mathbf{A}_k \in \mathbb{R}^{d_k \times r}} \sum_{\omega \in \Omega} |\mathcal{Y}(\omega) - \pi| F(\mathcal{Z}(\omega) \text{sgn}(\mathcal{Y}(\omega) - \pi)),$ 
7:       where  $F(\cdot)$  is the large-margin loss, and  $\mathcal{Z} = \sum_{s \in [r]} \mathbf{a}_s^{(1)} \otimes \dots \otimes \mathbf{a}_s^{(K)}$  is a rank- $r$  tensor.
8:     end for
9:   end while
10:  Return  $\mathcal{Z}_\pi \leftarrow \sum_{s \in [r]} \mathbf{a}_s^{(1)} \otimes \dots \otimes \mathbf{a}_s^{(K)}$ .
11: end for
```

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

References I

- De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.
- Globerson, A., Chechik, G., Pereira, F., and Tishby, N. (2007). Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, 8:2265–2295.
- Hitchcock, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189.
- Jain, P. and Oh, S. (2014). Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, volume 27, pages 1431–1439.

References II

- Montanari, A. and Sun, N. (2018). Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425.
- Wang, L., Durante, D., Jung, R. E., and Dunson, D. B. (2017). Bayesian network–response regression. *Bioinformatics*, 33(12):1859–1866.