

Beyond low-rankness: nonparametric tensor completion and regression

# Nonparametric signal tensor estimation via sign series

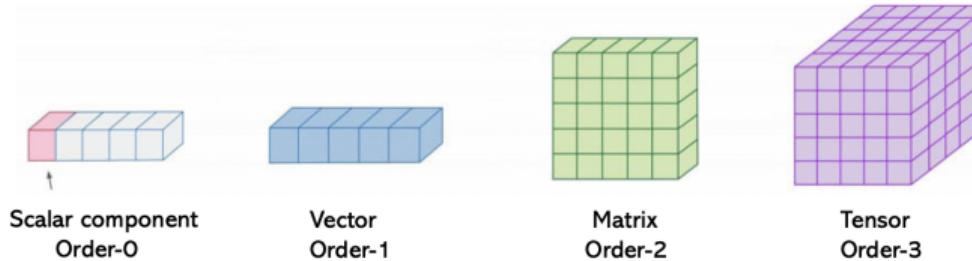
Chanwoo Lee

Department of Statistics  
University of Wisconsin - Madison

Preliminary Exam

# What is tensor?

- Tensors are generalizations of vectors and matrices:



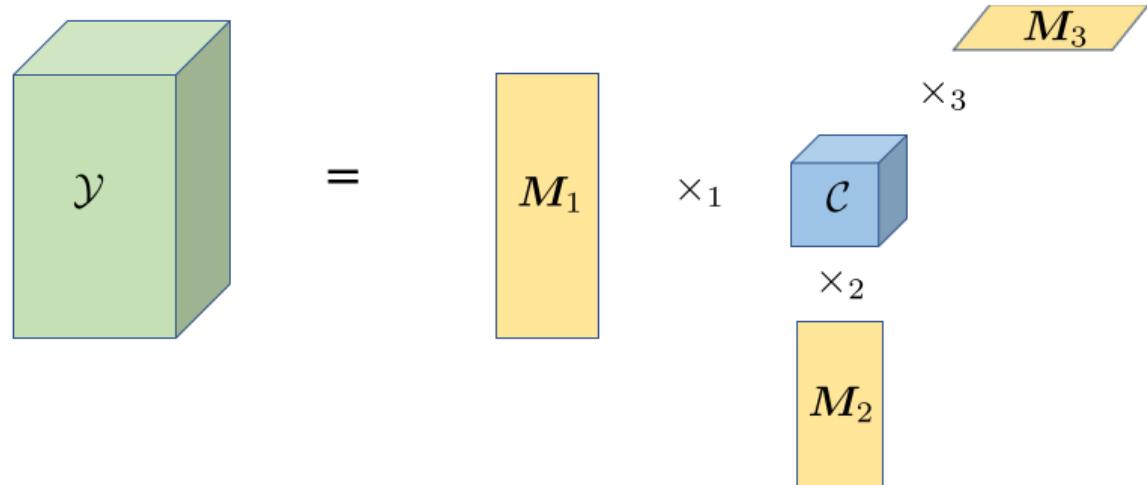
- We focus on tensors of order 3 or greater, also called **higher-order tensors**.
- Denote an order- $K(d_1, \dots, d_K)$  dimensional tensor as  $\mathcal{Y} = [\![y_\omega]\!] \in \mathbb{R}^{d_1 \times \dots \times d_K}$  where  $\omega \in [d_1] \times \dots \times [d_K]$ .

# Tensor decomposition (Tucker decomposition)

Delete

- Tucker decomposition (De Lathauwer et al., 2000).
  - $\mathcal{Y} = \mathcal{C} \times_1 \mathbf{M}_1 \times_2 \mathbf{M}_2 \times_3 \mathbf{M}_3$ .
  - 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

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

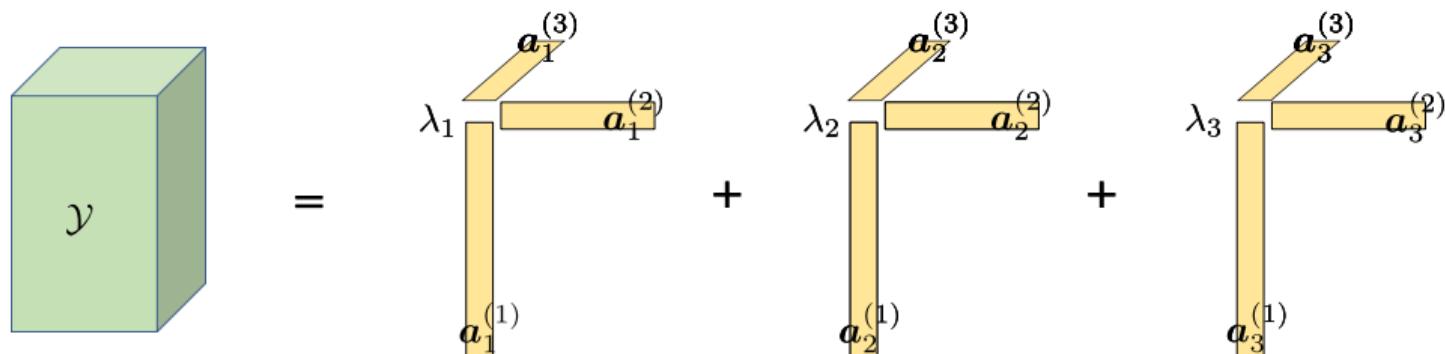


# Tensor decomposition (CP decomposition)

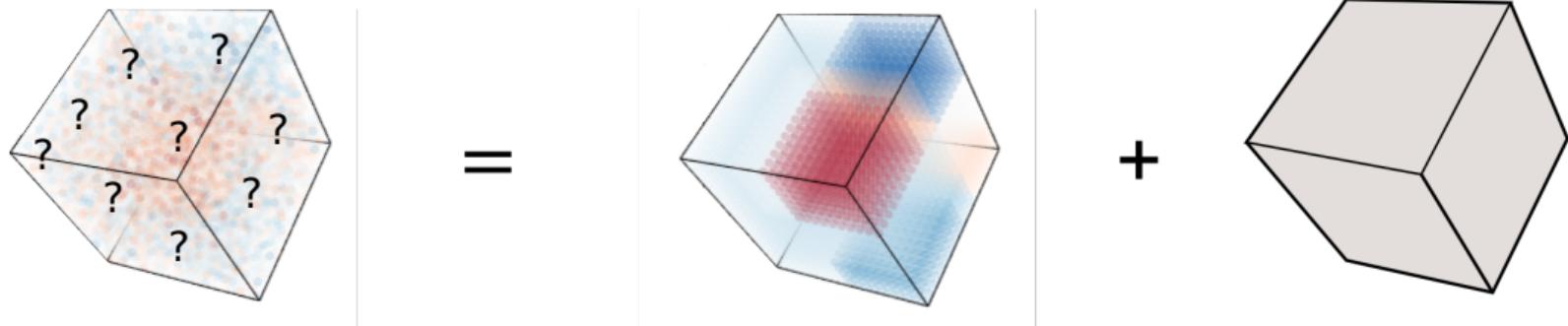
Delete

- Canonical Polyadic (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



Rewrite pages 1-5:

- My research (my p2)  
Data tensor

- What is tensor (your p2 + p5)

- Challenges (my p6. Structure + Data type)

- Talk outline (three bullets)

1. Previous work. Parametric tensor models for ordinal observations. (p5+1~2 slides).

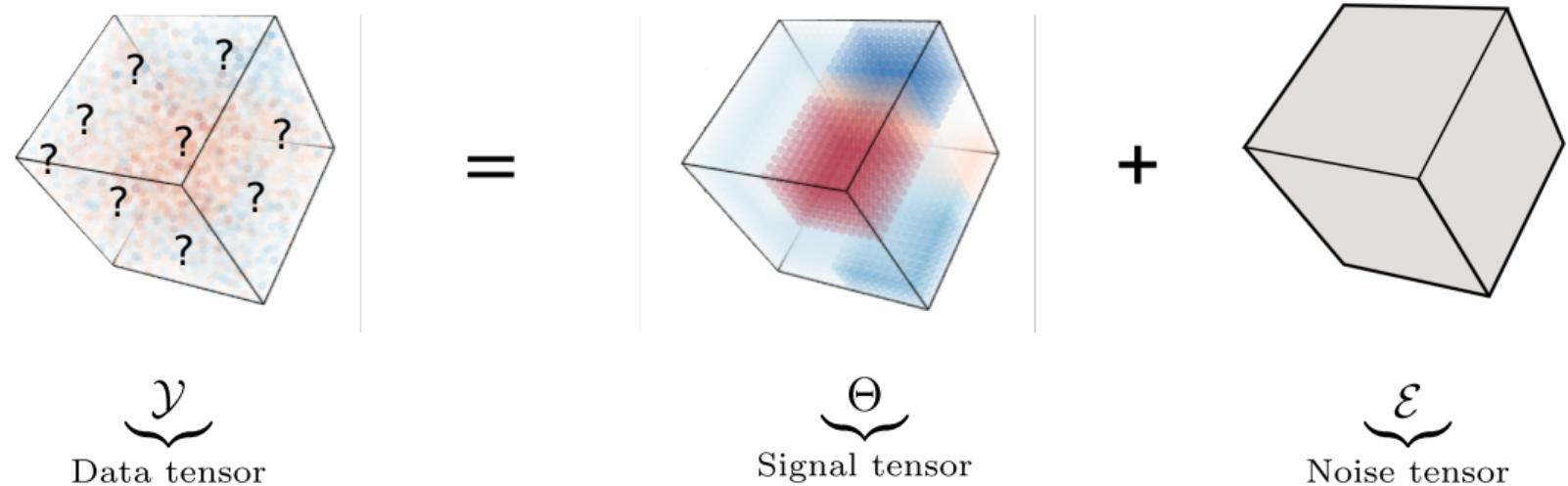
2. Current work. Nonparametric models for tensor completion and regression.

3. Future work. Other approaches for nonparametric tensor modeling.

$\Theta$   
Signal tensor

$\mathcal{E}$   
Noise tensor

# Main problems: the signal plus noise model



We focus on the two problems

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

# Tensor based learning is challenging

## High-rank matrix model

- **Matrix based model** (Cai and Zhou, 2013; Davenport et al., 2014; Ganti et al., 2015; Fan et al., 2019) update citation as well

# Tensor based learning is challenging

- **Matrix based model** (Cai and Zhou, 2013; Davenport et al., 2014; Ganti et al., 2015; Fan et al., 2019)
  - Applying matrix methods to higher-order tensor destroys structural information.
  - Tensor is different from matrix and more challenging.

# Tensor based learning is challenging

- **Matrix based model** (Cai and Zhou, 2013; Davenport et al., 2014; Ganti et al., 2015; Fan et al., 2019)
  - Applying matrix methods to higher-order tensor destroys structural information.
  - ~~Tensor is different from matrix and more challenging~~ (vage)  
tensors are more challenging because tensor rank may exceed dimension
- **Low rank tensor model** (Jain and Oh, 2014; Montanari and Sun, 2018; Cai et al., 2019)

- **Matrix based model** (Cai and Zhou, 2013; Davenport et al., 2014; Ganti et al., 2015; Fan et al., 2019)
  - Applying matrix methods to higher-order tensor destroys structural information.
  - Tensor is different from matrix and more challenging.
- **Low rank tensor model** (Jain and Oh, 2014; Montanari and Sun, 2018; Cai et al., 2019)
  - Low rank models are inadequate in many cases.

# Inadequacies of low rank models

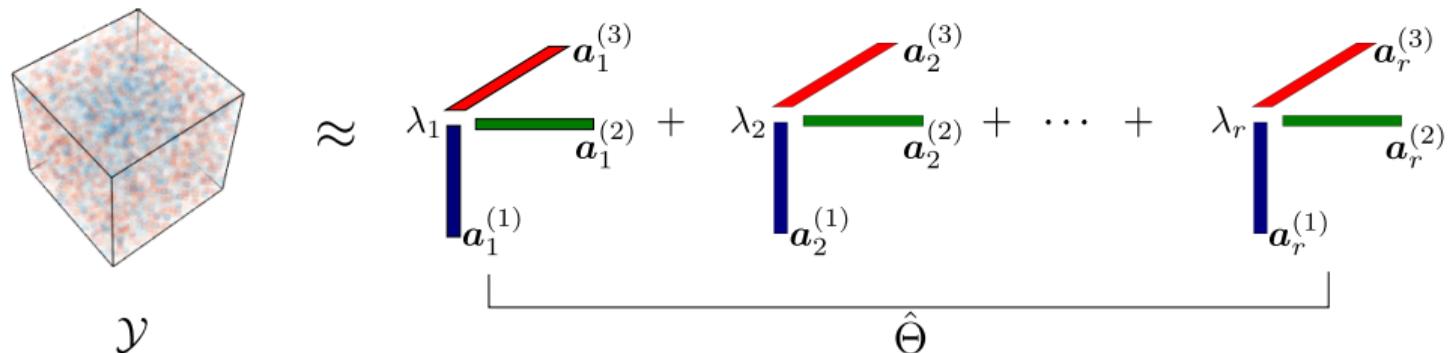
- Low rank models. (citation ....)

$$\mathcal{Y} \approx \lambda_1 a_1^{(1)} + \lambda_2 a_2^{(1)} + \cdots + \lambda_r a_r^{(1)} + \hat{\Theta}$$

# Inadequacies of low rank models

Delete.

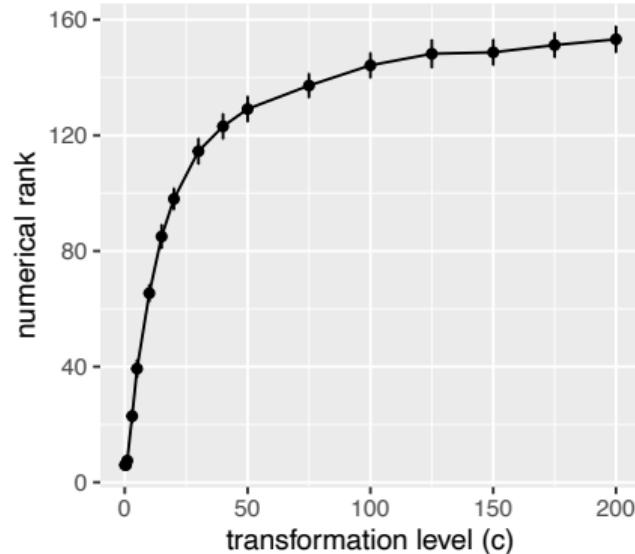
- Low rank models.



- There are two limitations of the model
  1. The sensitivity to order-preserving transformation.
  2. The inadequacy for special structures.

# Inadequacies of low rank models

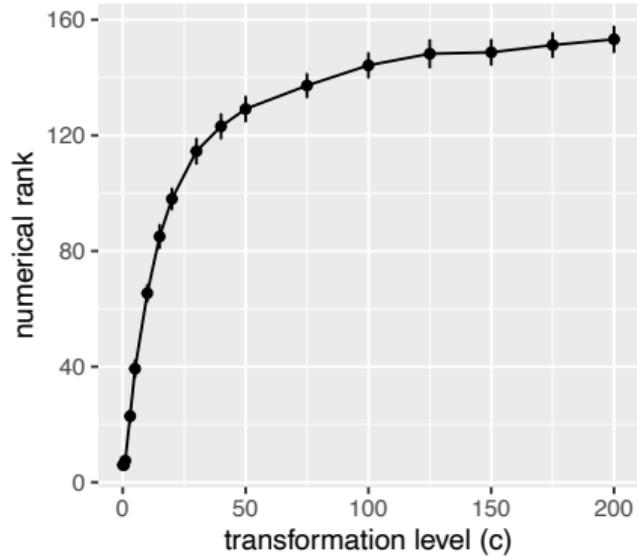
- The **sensitivity** to order-preserving transformation



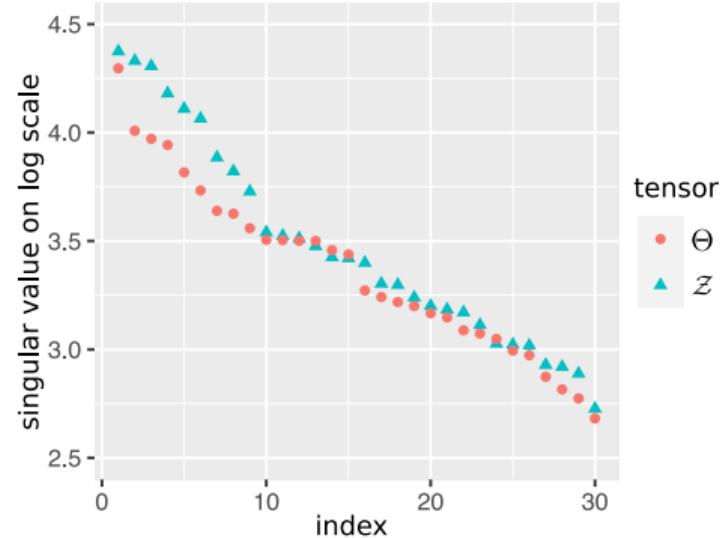
$$\Theta = \frac{1}{1 + \exp(-c(\mathcal{Z}))}, \quad \text{where}$$
$$\mathcal{Z} = \mathbf{a}^{\otimes 3} + \mathbf{b}^{\otimes 3} + \mathbf{c}^{\otimes 3}.$$

# Inadequacies of low rank models

- The **sensitivity** to order-preserving transformation
- The **inadequacy** for special structures.



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



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

## Motivating toy example in noiseless case

Delete.

$$\underbrace{\begin{pmatrix} -1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}}_{\Theta} \xrightarrow[\pi \in \{-1, -0.5, 0, 0.5, 1\}]{\text{Dichotomization}} \underbrace{\left\{ \begin{array}{lll} \pi = -1 & \pi = -0.5, 0 & \pi = 0.5, 1 \\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, & \begin{pmatrix} -1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, & \begin{pmatrix} -1 & -1 & 1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{array} \right\}}_{\text{sign}(\Theta - \pi)} \xrightarrow{\text{Aggregation}} \underbrace{\begin{pmatrix} -\frac{3}{5} & \frac{1}{5} & 1 \\ \frac{1}{5} & \frac{1}{5} & 1 \\ 1 & 1 & 1 \end{pmatrix}}_{\tilde{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \Pi} \text{sign}(\Theta - \pi)}$$

## ~~Motivating toy example in noiseless case~~

Why sign matters?

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

$\Theta \sim \frac{1}{H} \sum_{\pi \in H} \text{sign}(\Theta - \pi)$ , where  $H = \{-1, \dots, -1/H, 0, 1/H, \dots, 1\}$

$$\underbrace{\begin{pmatrix} -1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}}_{\Theta} \xrightarrow[\pi \in \{-1, -0.5, 0, 0.5, 1\}]{\text{Dichotomization}} \left\{ \begin{array}{lll} \pi = -1 & \pi = -0.5, 0 & \pi = 0.5, 1 \\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, & \begin{pmatrix} -1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, & \begin{pmatrix} -1 & -1 & 1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{array} \right\}_{\text{sign}(\Theta - \pi)} \xrightarrow{\text{Delete } \text{Aggregation}} \underbrace{\begin{pmatrix} -\frac{3}{5} & \frac{1}{5} & 1 \\ \frac{1}{5} & \frac{1}{5} & 1 \\ 1 & 1 & 1 \end{pmatrix}}_{\tilde{\Theta} = \frac{1}{2H+1} \sum_{\pi \in H} \text{sign}(\Theta - \pi)}$$

Sign (\*not signed\*) tensors are invariant to order-preserving transformation.

- Signed tensors are much simpler in the sense of rank.
- With a series of sign tensors, we successfully preserve all information in the **More flexible signal tensors allowed by using sign tensor series representation.**
- In **noise** case, we estimate  $\text{sgn}(\Theta - \pi)$  from the tensor data  $\text{sgn}(\mathcal{Y} - \pi)$ .  
**noisy**

## Sign rank

\* Key idea: we use a local notion of low-rankness to allow a richer family of signal tensors

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

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

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

# Sign rank

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

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

remove (ex)

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

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

- Most existing structure tensors belong to sign representable family:
  - Low-rank CP tensors, Tucker tensors, stochastic block models.
  - High-rank tensors from GLM, single index models,
  - Tensors with repeating patterns, e.g.  $\Theta(i_1, \dots, i_K) = \log(1 + \max(i_1, \dots, i_K))$  is 2-sign representable.

# Sign representable tensors

## Sign representable tensors

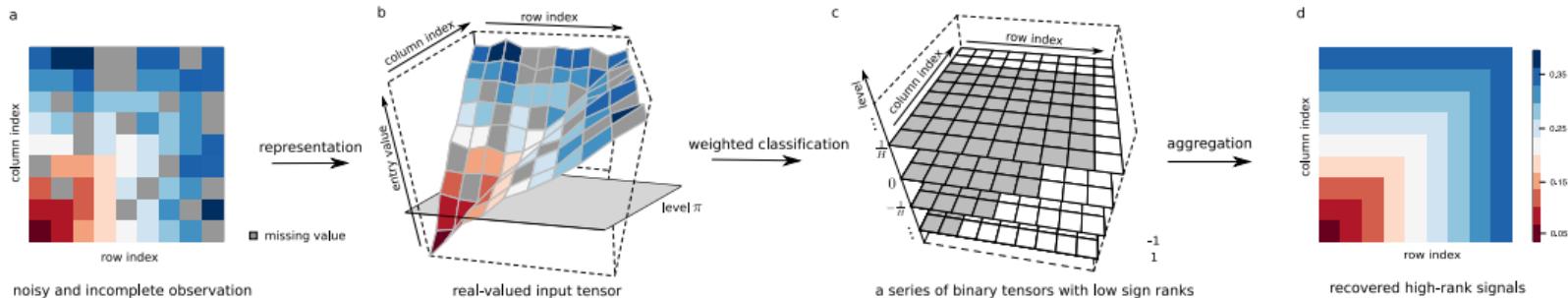
A tensor  $\Theta$  is called ***r*-sign representable** if the tensor  $(\Theta - \pi)$  has sign rank bounded by  $r$  for all  $\pi \in [-1, 1]$ .

- Most existing structure tensors belong to sign representable family:
  - Low-rank CP tensors, Tucker tensors, stochastic block models.
  - High-rank tensors from GLM, single index models,
  - Tensors with repeating patterns, e.g.  $\Theta(i_1, \dots, i_K) = \log(1 + \max(i_1, \dots, i_K))$  is 2-sign representable.
- Instead of the classical low rank assumption, we propose the **sign representable tensor family**

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

# ~~Our new approach~~ Our solution: sign signal helps!

general principle: slide title should be informative

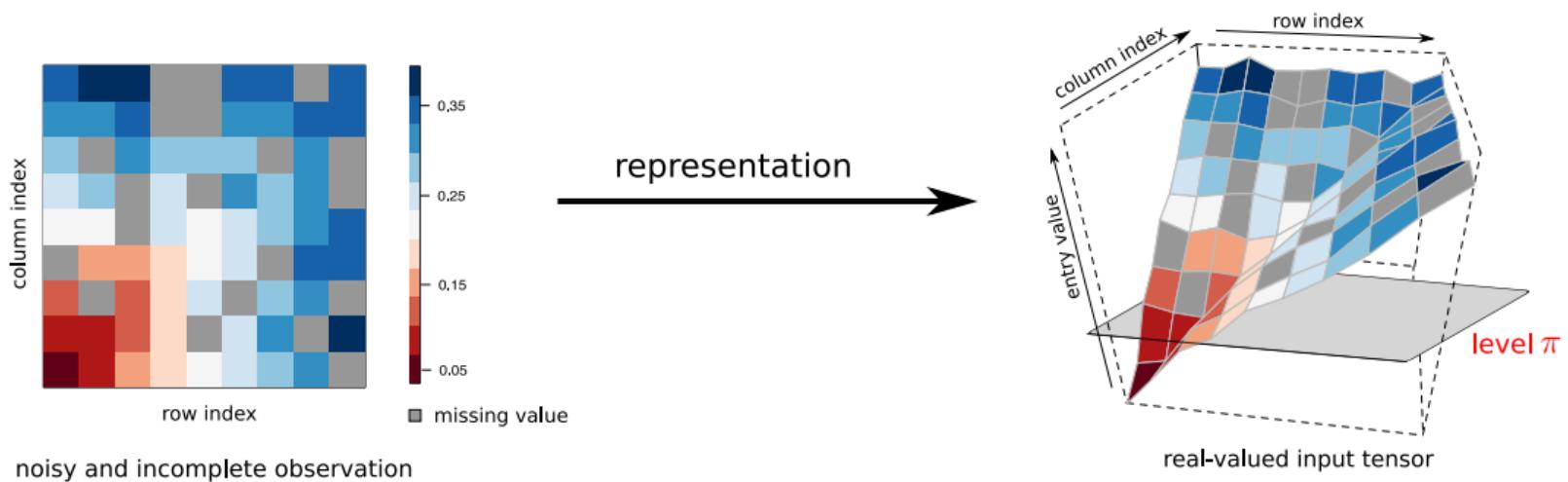


your earlier clock layout in 992 is better.  
(current figure too small to see).

two slides:

first figure shows a and d (with a question mark on top of question mark )  
second shows a-d (four pictures in a clock).

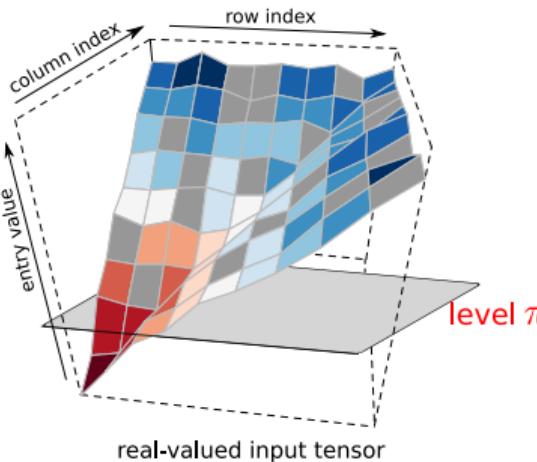
## Step 1: representation



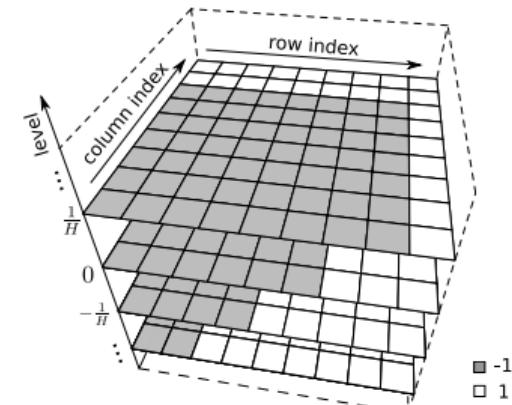
- We observe a noisy incomplete tensor  $\mathcal{Y}_\Omega \in [-1, 1]^{d_1 \times \dots \times d_K}$  with observed index set  $\Omega \subset [d_1] \times \dots \times [d_K]$ .
- We dichotomize the data into a series of sign tensors:

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

## Step 2: weighted classification



weighted classification



a series of binary tensors with low sign ranks

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

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

Do you mean weight |Y-pi|?

- Magnitude  $|\Theta(\omega) - \pi|$  plays important role in estimation.

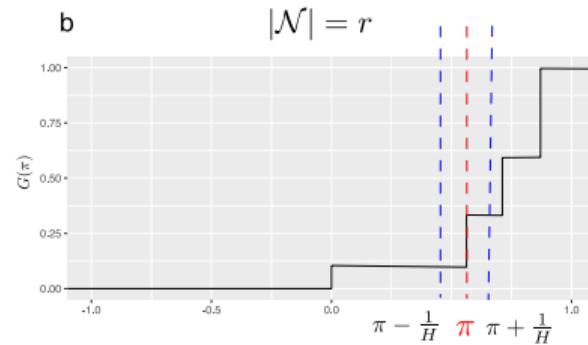
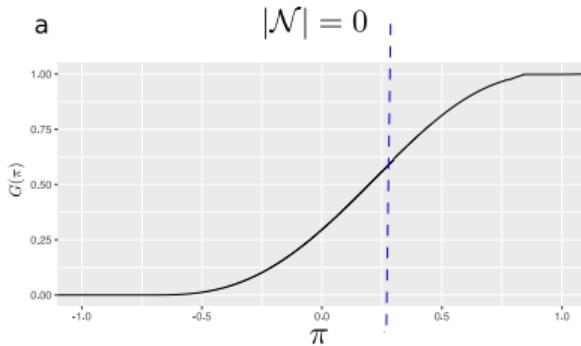
# Main assumption

S2

Combine S2+S2'. No need to explain much on N.

- We quantify difficulty of the problem using CDF  $G(\pi) = \mathbb{P}_{\omega \in \Pi}[\Theta(\omega) \leq \pi]$ .
- We define classification hard region  $\mathcal{N}$  at which point mass concentrates,

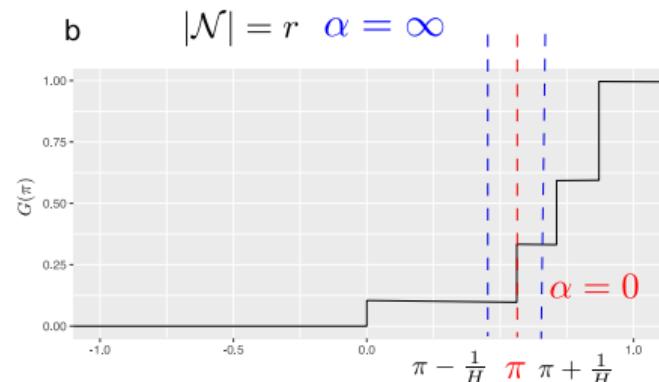
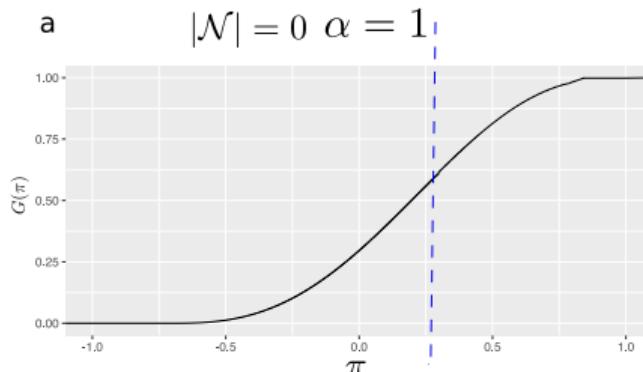
$$\mathcal{N} = \left\{ \pi : \frac{G(\pi + \Delta s) - G(\pi - \Delta s)}{\Delta s} \geq C \right\}, \text{ where } \Delta s = 1 / \prod_k^K d_k.$$



$\alpha$  smoothness

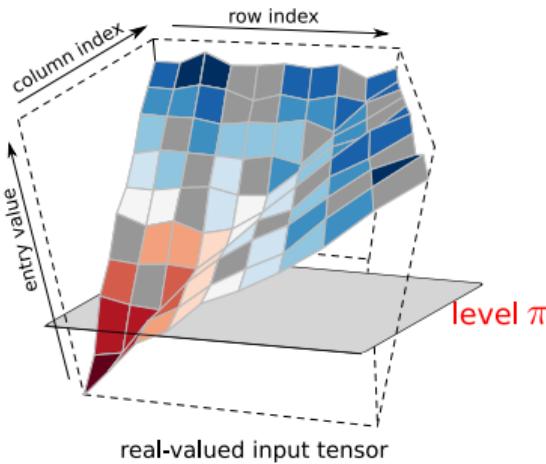
For fixed  $\pi \in \mathcal{N}^c$ , we call  $\Theta$  is  $\alpha$  smooth if there exist  $\alpha = \alpha(\pi) > 0, c = c(\pi) > 0$ , such that

- \* Partition  $[-1, 1] = N + N^c$ , where  $N^c$  consists of levels whose pseudo density is uniformly bounded, and  $N$  otherwise.
  - \*  $G$  is globally alpha-smooth in that for all  $\pi \in N^c$ ,
- ....  
 where  $\rho(\pi, N) = \min_{\pi' \in N} |\pi - \pi'| + \Delta s$ . If  $\alpha$  and  $c$  are global constants for all  $\pi$ 's, we call  $\Theta$  is  $\alpha$ -globally smooth.
- Keep only global.  
no need to say local.



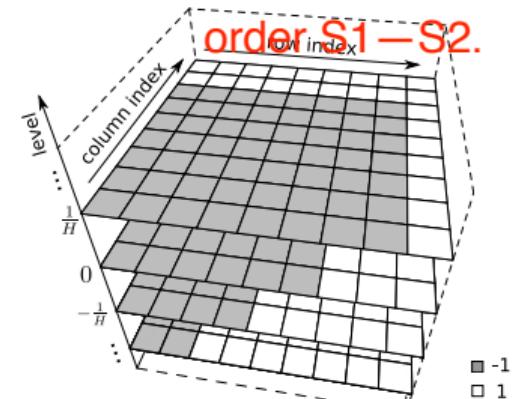
## Step 2: weighed classification

S1



weighted classification

highlight



a series of binary tensors with low sign ranks

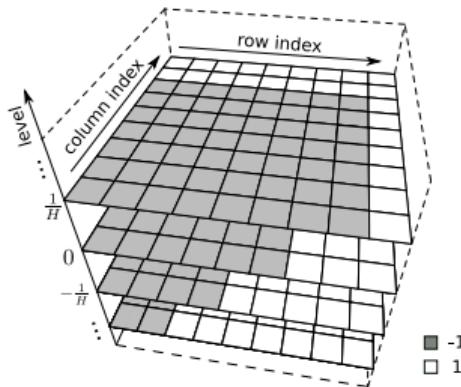
- If  $\Theta \in \mathcal{P}_{\text{sgn}}(r)$  is  $\alpha$  smooth ( $\alpha > 0$ ), we have **a unique optimizer** such that

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

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

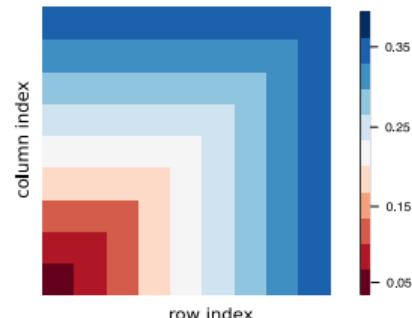
$$\hat{\mathcal{Z}}_\pi = \arg \min_{\mathcal{Z}: \text{rank}(\mathcal{Z}) \leq r} L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi).$$

## Step 3: 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 obtain the tensor estimate

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

# Sign tensor estimation error

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

## Sign tensor estimation for fixed $\pi$ (L. and Wang, 2021)

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

, and  $d_1 = \dots = d_K = d$

add footnote on \lessim (footnote: log <sub>$\alpha$</sub>  term suppressed)

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

- Sign estimation error shows a polynomial decay with the number of observed entries.  
principle: always present the simplest case in talk.

## Tensor estimation error

Tensor estimation error (L. and Wang 2021)

Suppose  $\Theta \in \mathcal{P}_{\text{sgn}}(r)$  is  $\alpha$ -globally smooth. Then, with very high probability over  $\mathcal{Y}_\Omega$ ,  
with bounded INI

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

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

add Bracket to each of the term:  
errors inherited from sign estimation  
Bias  
Variance

\* see paper for general case that allows unbounded IN^cl and sub-Gaussian noise.

Tensor estimation error (L. ~~and~~ Wang 2021) (power of  $d$ ) | Previous results

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

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

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

- For full observation case with equal dimension  $d_1 = \dots = d_K = d$  and bounded  $|\mathcal{N}| \leq C$ ,

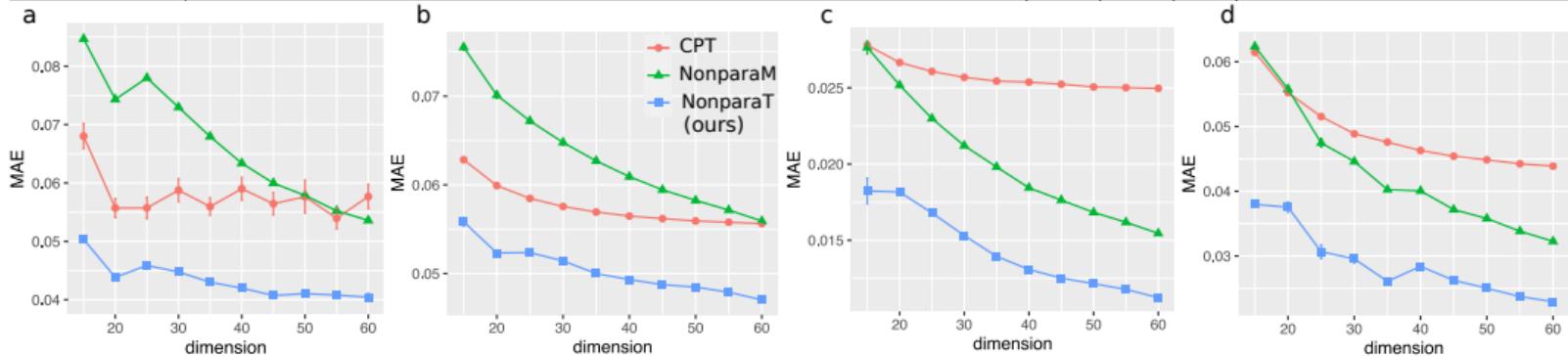
$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim \tilde{\mathcal{O}} \left( d^{-(K-1) \min\left(\frac{\alpha}{\alpha+2}, \frac{1}{2}\right)} \right).$$

- Tensor estimation is generally no better than sign tensor estimation.
- Sample complexity:  $\text{MAE}(\hat{\Theta}, \Theta) \rightarrow 0$ , as  $\frac{|\Omega|}{d_{\max} r \log^2 |\Omega|} \rightarrow \infty$ .

# Simulations for estimation error vs tensor dimension

Delete

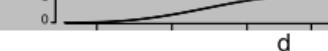
Simulation	Signal Tensor $\Theta$	Rank	Sign Rank	$\alpha$	$ \mathcal{N} $	CDF of Tensor Entries	Noise
1	$\mathcal{C} \times M_1 \times M_2 \times M_3$	$3^3$	$\leq 3^3$	$\infty$	$\leq 3^3$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Uniform $[-0.3, 0.3]$
2	$ a \otimes 1 \otimes 1 - 1 \otimes a \otimes 1 $	$d$	$\leq 3$	1	0	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + Z_{\max})$	$\geq d$	2	$\infty$	$d$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(Z_{\min}^{1/3})$	$\geq d$	2	$\infty$	$d$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Normal $\mathcal{N}(0, 0.15)$

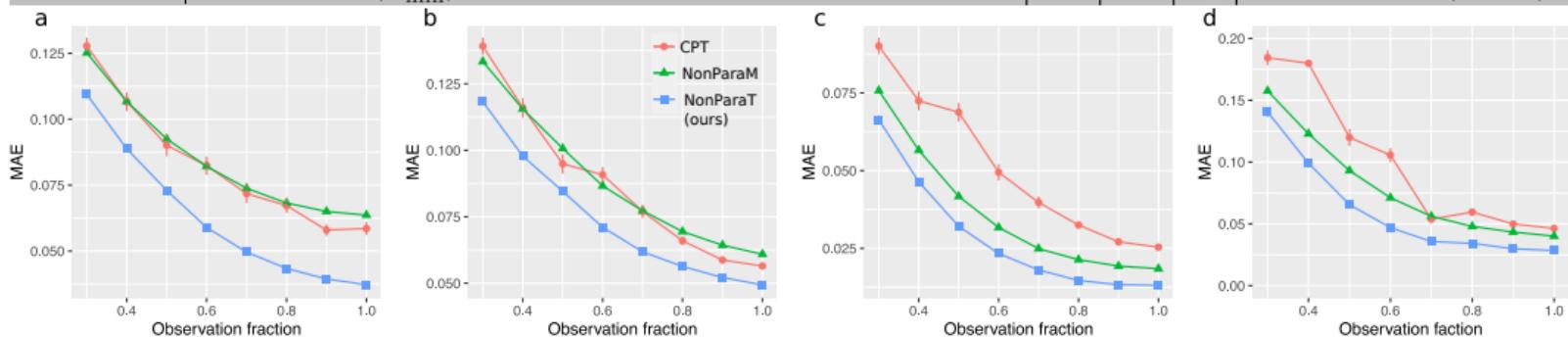


- **NonPraT**: Our nonparametric tensor method, **CPT**: low rank tensor CP decomposition, **NonPraraM**: the matrix version of our method.
- Our method (NonparaT) achieves the best performance.

# Simulations for estimation error vs the observation fraction

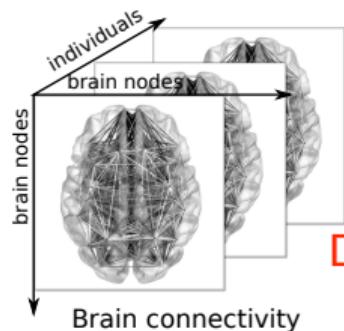
Delete

Simulation	Signal Tensor $\Theta$	Rank	Sign Rank	$\alpha$	$ \mathcal{N} $	CDF of Tensor Entries	Noise
1	$\mathcal{C} \times \mathbf{M}_1 \times \mathbf{M}_2 \times \mathbf{M}_3$	$3^3$	$\leq 3^3$	$\infty$	$\leq 3^3$		Uniform $[-0.3, 0.3]$
2	$ \mathbf{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \mathbf{a} \otimes \mathbf{1} $	$d$	$\leq 3$	1	0		Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + Z_{\max})$	$\geq d$	2	$\infty$	$d$		Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(Z_{\min}^{1/3})$	$\geq d$	2	$\infty$	$d$		Normal $\mathcal{N}(0, 0.15)$



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

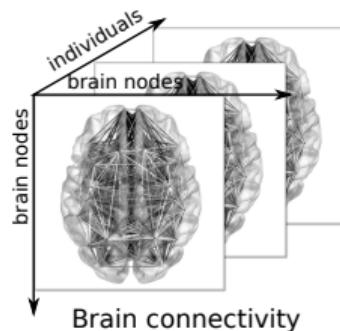
## Data application: Brain connectivity



- The MRN-114 human brain connectivity data consists of 68 brain regions for 114 individuals along with their IQ scores (Wang et al., 2017).

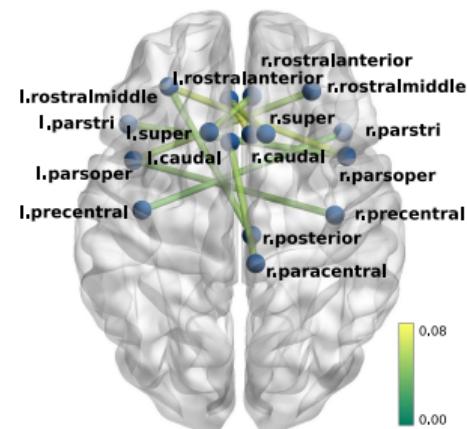
**Data tensor**  $\mathcal{Y} \in \{0, 1\}^{68 \times 68 \times 114}$ .

# Data application: Brain connectivity

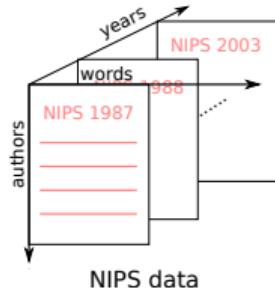


- The MRN-114 human brain connectivity data consists of 68 brain regions for 114 individuals along with their IQ scores (Wang et al., 2017).
- $\mathcal{Y} \in \{0, 1\}^{68 \times 68 \times 114}$ .

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

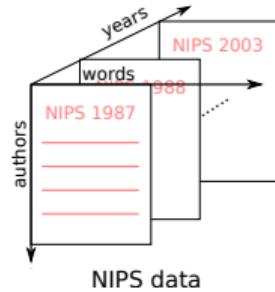


## Data application: NIPS



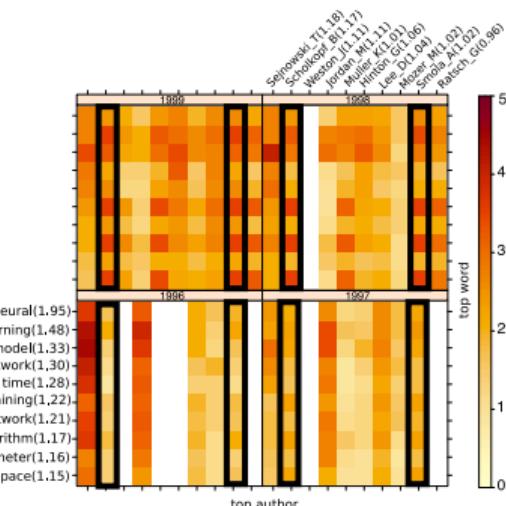
- The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003 (Globerson et al., 2007).
- Log transformation yields the dataset  $\mathcal{Y} \in \mathbb{R}^{100 \times 200 \times 17}$ .

# 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 are consistent with the active topics
- There are strong heterogeneity among word occurrences across authors and years.
- Similar word patterns (B. Schölkopf and A. Smola).



# Data application: Brain connectivity + NIPS

MRN-114 brain connectivity dataset					
Method	$r = 3$	$r = 6$	$r = 9$	$r = 12$	$r = 15$
NonparaT (Ours)	<b>0.18(0.001)</b>	<b>0.14(0.001)</b>	<b>0.12(0.001)</b>	<b>0.12(0.001)</b>	<b>0.11(0.001)</b>
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)	<b>0.18(0.002)</b>	<b>0.16(0.002)</b>	<b>0.15(0.001)</b>	<b>0.14(0.001)</b>	<b>0.13(0.001)</b>
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)		

5-folded cross-validation.

Table: MAE comparison in the brain data and NIPS data on ~~cross-validation (5 repetitions 5 folds)~~.  
~~Standard errors are reported in parenthesis.~~

- Our method substantially outperforms the low-rank CP method for every configuration under consideration.

# Nonparametric trace regression via sign series

Add transition slide:  
Outline (three bullet points)  
Current bullet (visible).  
Other bullet in gray fonts (less visible)

L26-28. Write in terms of tensor.

- $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^{d_1 \times d_2}$  denotes the matrix predictor,  $Y \in \mathbb{R}$  the scalar response. We consider the regression model,

$$Y = f(\mathbf{X}) + \epsilon,$$

where  $f: \mathcal{X} \rightarrow \mathbb{R}$  is an unknown regression function of interest, and  $\epsilon$  is mean-zero noise.

- Trace regression (Fan et al., 2019; Hamidi and Bayati, 2019) assumes that

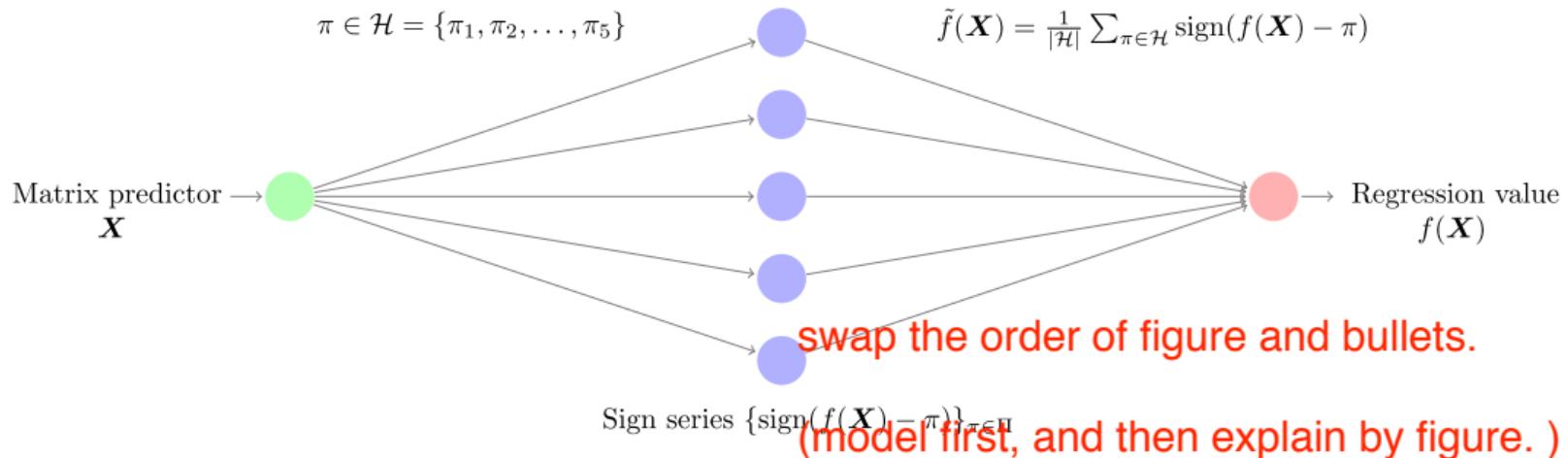
$$f(\mathbf{X}) = \langle \mathbf{B}, \mathbf{X} \rangle + b, \text{ for all } \mathbf{X} \in \mathcal{X},$$

where  $\mathbf{B}$  is usually a low rank matrix.

Functional form of  $f(\mathbf{X})$

- Low rank assumption on  $\mathbf{B}$  is not adequate in many cases.

# Nonparametric trace regression via sign series



- We apply our framework for the regression problem.

- We assume that for given  $\pi$ , **sign of regression function** is represented by **two examples:**

$$\text{sign}(f(\mathbf{X}) - \pi) = \text{sign} (\langle \mathbf{B}_\pi, \mathbf{X} \rangle + b_\pi)$$

where  $\mathbf{B}_\pi$  is a low rank matrix.

1. single index model
2. tensor completion.

# Nonparametric trace regression via sign series

- Based on **the sign rank** assumption, we estimate the series of sign functions  $\{\text{sign}(f - \pi)\}_{\pi \in \mathcal{H}}$  from the weighted classification,

$$L(\phi; \{(\mathbf{X}_i, \text{sign}(Y_i - \pi))\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^n |Y_i - \pi| \times |\text{sign}(Y_i - \pi) - \text{sign}(\phi(\mathbf{X}_i))|.$$

(add one sentence about learning reduction + meta/base algorithm + no need to reinvent the wheel)

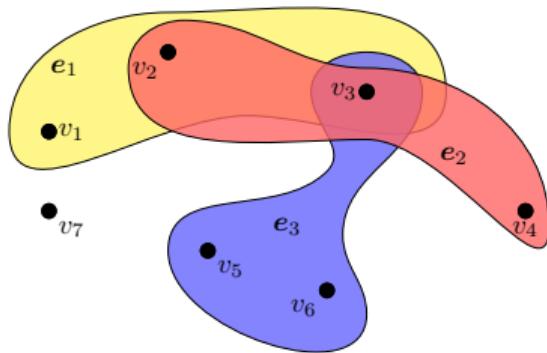
- Our framework enjoys theoretical guarantees.

- We detailed the trace regression problem in Lee et al. (2021).

Our extend the broad nonparametric paradigm to many important matrix/tensor learning problems, including regression, completion, multi-task learning, and compressed sensing

# Nonparametric probability tensor estimation from hypergraph

add transition slide



- Hypergraph considers higher-way interaction among nodes.
- An observed adjacency tensor  $\mathcal{A} \in \{0, 1\}^{d \times \dots \times d}$  corresponding to a hypergraph is generated by

$$\mathcal{A}_\omega \sim \text{Bernoulli}(\Theta_\omega), \text{ for all } \underbrace{\omega \in [d] \times \dots \times [d]}_K.$$

- The **block model** (Ahn et al., 2018; Wang and Zeng, 2019; Han et al., 2020) has

$$\Theta_\omega = \mathcal{Q}_{z(\omega_1), z(\omega_2), \dots, z(\omega_K)}, \text{ for all } \omega \in [d] \times \dots \times [d],$$

where  $\mathcal{Q} \in [0, 1]^{m \times \dots \times m}$  and  $z: [d] \rightarrow [m]$  is a hidden partition.

- We will take **nonparametric** approach considering **smooth function  $f$**  such that

$$\Theta_\omega = f(\xi_{\omega_1}, \xi_{\omega_2}, \dots, \xi_{\omega_K}), \text{ for all } \omega \in [d] \times \dots \times [d],$$

where  $\{\xi_i\}_{i=1}^d$  are i.i.d. random variables sampled from  $\text{Unif}[0, 1]$ .

Thank you!

# Appendix: Algorithm

---

**Algorithm 1** Nonparametric tensor completion

---

**Input:** Noisy and incomplete data tensor  $\mathcal{Y}_\Omega$ , 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:  $\mathbf{A}_k \leftarrow \arg \min_{\mathbf{A}_k \in \mathbb{R}^{d_k \times r}} \sum_{\omega \in \Omega} |\mathcal{Y}(\omega) - \pi| F(\mathcal{Z}(\omega) \text{sgn}(\mathcal{Y}(\omega) - \pi))$ ,  
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.
- 6:         **end for**
- 7:     **end while**
- 8:     Return  $\mathcal{Z}_\pi \leftarrow \sum_{s \in [r]} \mathbf{a}_s^{(1)} \otimes \dots \otimes \mathbf{a}_s^{(K)}$ .
- 9: **end for**

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

---

# Theoretical guarantees for a large-margin loss classification

- we consider the estimation

$$\hat{\mathcal{Z}}_\pi = \arg \min_{\substack{\text{rank}(\mathcal{Z}) \leq r \\ \omega \in \Omega}} \sum |\mathcal{Y}(\omega) - \pi| \times F(\mathcal{Z}(\omega)\text{sign}(\mathcal{Y}(\omega) - \pi)) + \lambda \|\mathcal{Z}\|_F^2,$$

where  $\lambda > 0$  is the penalty parameter and  $F$  is a large-margin loss satisfying the following assumption,

## Assumption 1

- (a) (Approximation error) For any given  $\pi \in [-1, 1]$ , there exists a sequence of tensors  $\mathcal{Z}_\pi^{(n)} \in \mathcal{P}_{\text{sgn}}(r)$ , such that  $\text{Risk}_F(\mathcal{Z}_\pi^{(n)}) - \text{Risk}_F(\Theta - \pi) \leq a_n$ , for some sequence  $a_n \rightarrow 0$  as  $n \rightarrow \infty$ . Furthermore, assume  $\|\mathcal{Z}_\pi^{(n)}\|_F \leq J$  for some constant  $J > 0$ .
- (b)  $F(z) = (1 - z)_+$  is hinge loss.

# Theoretical guarantees for a large-margin loss classification

## Estimation error for a large margin loss (L. and Wang 2021)

Denote  $t_n = \frac{d_{\max} r K \log n}{n}$ . Suppose the surrogate loss  $F$  satisfies Assumption 1 with  $a_n \lesssim t_n^{(\alpha+1)/(\alpha+2)}$ . Set  $\lambda \asymp t_n^{(\alpha+1)/(\alpha+2)} + t_n/\rho(\pi, \mathcal{N})$ . Then, with high probability, we have:

1. (Sign tensor estimation). For all  $\pi \in [-1, 1]$  except for a finite number of levels,

$$\text{MAE}(\text{sign}(\hat{\mathcal{Z}}_\pi), \text{sign}(\Theta - \pi)) \lesssim t_n^{\frac{\alpha}{2+\alpha}} + \frac{1}{\rho^2(\pi, \mathcal{N})} t_n.$$

2. (Tensor estimation).

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim (t_n \log H)^{\frac{\alpha}{2+\alpha}} + \frac{1 + |\mathcal{N}|}{H} + t_n H \log H.$$

## References I

- Ahn, K., Lee, K., and Suh, C. (2018). Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974.
- Cai, C., Li, G., Poor, H. V., and Chen, Y. (2019). Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pages 1863–1874.
- Cai, T. and Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647.
- Davenport, M. A., Plan, Y., Van Den Berg, E., and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223.
- 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.

## References II

- Fan, J., Gong, W., and Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202.
- Ganti, R. S., Balzano, L., and Willett, R. (2015). Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pages 1873–1881.
- Globerson, A., Chechik, G., Pereira, F., and Tishby, N. (2007). Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, 8:2265–2295.
- Hamidi, N. and Bayati, M. (2019). On low-rank trace regression under general sampling distribution. *arXiv preprint arXiv:1904.08576*.
- Han, R., Luo, Y., Wang, M., and Zhang, A. R. (2020). Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*.
- 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.

## References III

- Jain, P. and Oh, S. (2014). Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, volume 27, pages 1431–1439.
- Lee, C., Li, L., Zhang, H. H., and Wang, M. (2021). Nonparametric trace regression in high dimensions via sign series representation. *arXiv preprint arXiv:2105.01783*.
- 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.
- Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, pages 713–723.