Beyond low-rankness: nonparametric models for tensor completion and regression

Chanwoo Lee

Department of Statistics
University of Wisconsin - Madison

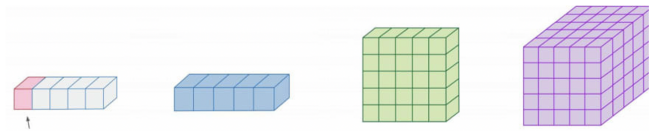
Preliminary Exam

My research

- My research interests lie at the intersection of **statistics, machine learning, and optimization**.
- Specific interests include
  - tenosor/matrix data analysis
  - high-dimensional statistics
  - non-convex optimization
  - nonparametric statistics
- My goal is to develop statistical tools for analyzing **array valued data**.

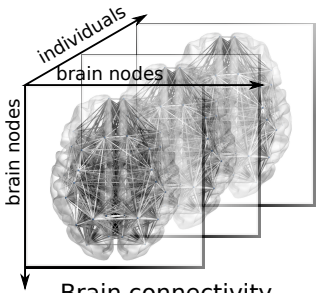
What is tensor?

- Tensors are generalizations of vectors and matrices:

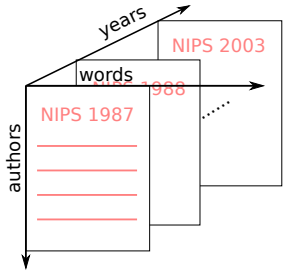


- We focus on tensors of order 3 or greater, also called higher-order tensors.
- Denote an order-K $(d_1, \cdots, d_K)$ dimensional tensor as $\mathcal{Y} = [\![y_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ where $\omega \in [d_1] \times \cdots \times [d_K]$.

Tensors in application



Brain connectivity



NIPS data

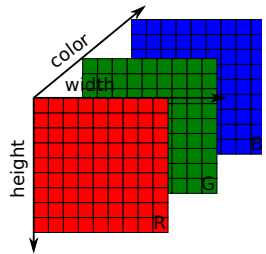


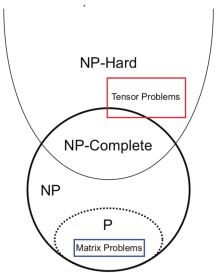
Image data

1. The human brain connectivity dataset consists of 68 brain regions for 114 individuals (Wang et al., 2017).
2. The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003 along with author information (Globerson et al., 2007).
3. An RGB image consists of pixel values across three channels

Talk outline

Prohibitive Computational Complexity

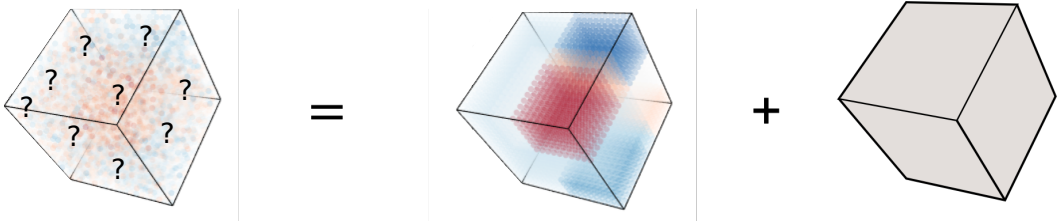
Most higher-order tensor problems are NP-hard [Hillar & Lim, 2013].



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

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

Main problems: the signal plus noise model

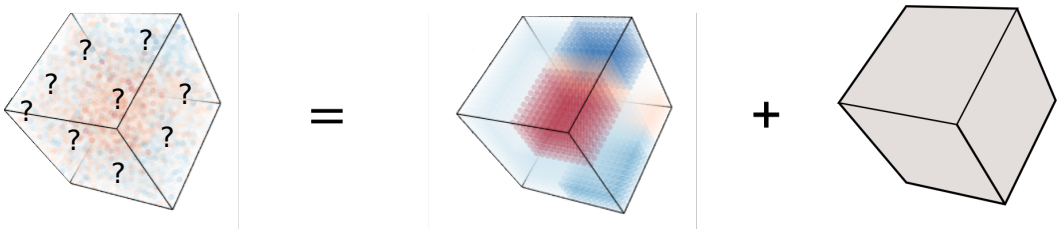


$\underbrace{\mathcal{Y}}$
Data tensor

$\underbrace{\Theta}$
Signal tensor

$\underbrace{\mathcal{E}}$
Noise tensor

Main problems: the signal plus noise model

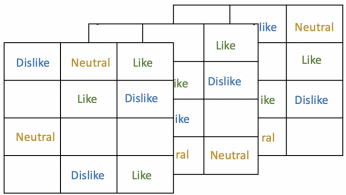


We focus on the two problems

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

Outline

A cumulative link model for ordinal-valued tensors



- Let $\mathcal{Y} = [\![y_\omega]\!] \in [L]^{d_1 \times \cdots \times d_K}$ be an ordinal tensor, where $[L] = \{1, 2, \cdots, L\}$ is the ordinal level.

- We propose a cumulative link model,

$$\mathbb{P}(y_\omega \leq \ell | \boldsymbol{b}, \Theta) = f(b_\ell - \theta_\omega), \qquad \text{for all } \ell \in [L-1],$$

  where
  - $\Theta = [\![\theta_\omega]\!]$ represents a low-rank latent parameter tensor prior to quantization.
  - $\boldsymbol{b} = (b_1, \ldots, b_{L-1})$ represents cutoff values in quantization.
- f is a known link function (quantization operation).
  e.g. $f(x) = e^x/(1+e^x)$ is a logistic link.

Low-rank assumption and estimation

- Our work generalized earlier 1-bit tensor completion.
- We propose a rank-constrained M-estimate based on log-likelihood:

$$\mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}) = \sum_{\omega\in\Omega}\sum_{\ell\in[L]} \Big\{ \mathbb{1}_{\{y_\omega=\ell\}} \log\big[f(b_\ell - \theta_\omega) - f(b_{\ell-1} - \theta_\omega)\big] \Big\}.$$

- The method achieves optimal convergence rate and nearly optimal sample complexity.
- See paper for more details (L. and M. Wang. Tensor denoising and completion based on ordinal observations. ICML, PMLR 119:5778-5788, 2020.)

Low-rank assumption and estimation

- Our work generalized earlier 1-bit tensor completion.
- We propose a rank-constrained M-estimate based on log-likelihood:

$$\mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}) = \sum_{\omega\in\Omega}\sum_{\ell\in[L]} \Big\{ \mathbb{1}_{\{y_\omega=\ell\}} \log\big[f(b_\ell - \theta_\omega) - f(b_{\ell-1} - \theta_\omega)\big] \Big\}.$$

- The method achieves optimal convergence rate and nearly optimal sample complexity.
- See paper for more details (L. and M. Wang. Tensor denoising and completion based on ordinal observations. ICML, PMLR 119:5778-5788, 2020.)

What if we have no link function information?

Outline

Tensor based learning is challenging

- **High-rank matrix model** (Ganti et al., 2015; Ongie et al., 2017; Fan and Udell, 2019)

Tensor based learning is challenging

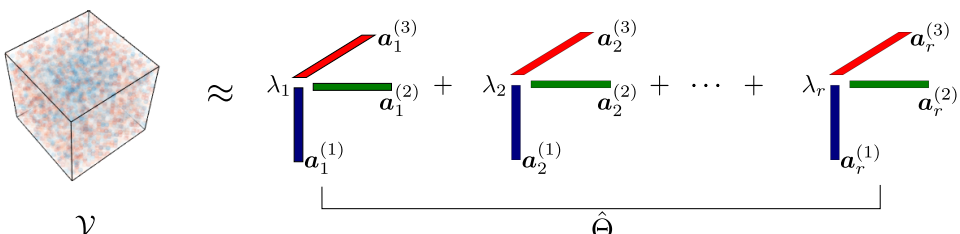
- **High-rank matrix model** (Ganti et al., 2015; Ongie et al., 2017; Fan and Udell, 2019)
  - Applying matrix methods to higher-order tensor destroys structural information.
  - Tensors are more challenging because tensor rank may exceed dimension.

Tensor based learning is challenging

- **High-rank matrix model** (Ganti et al., 2015; Ongie et al., 2017; Fan and Udell, 2019)
  - Applying matrix methods to higher-order tensor destroys structural information.
  - Tensors are more challenging because tensor rank may exceed dimension.
- **Low-rank tensor model** (Anandkumar et al., 2014; Montanari and Sun, 2018; Cai et al., 2019)
  - Low-rank models are inadequate in many cases.

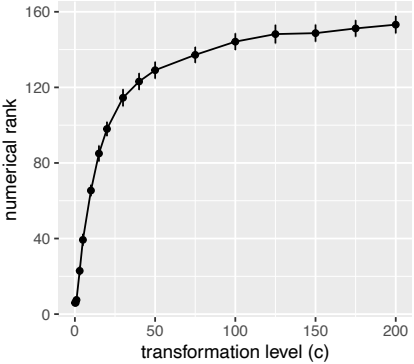
Inadequacies of low-rank models

- Low-rank models (Anandkumar et al., 2014; Montanari and Sun, 2018; Cai et al., 2019).



Inadequacies of low-rank models

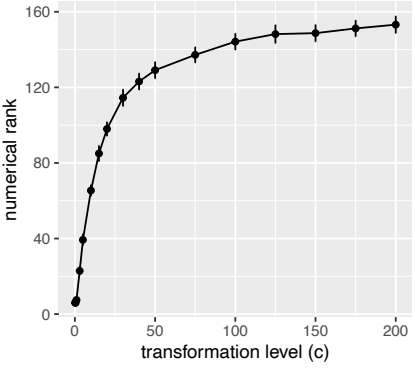
- Sensitivity to order-preserving transformation



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

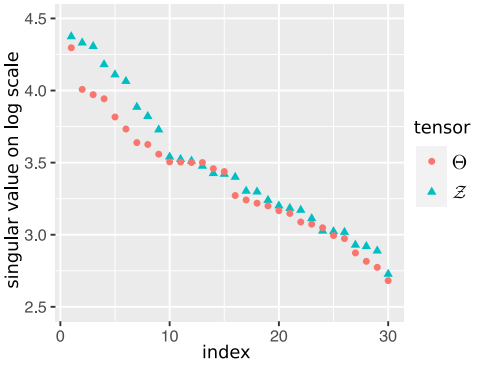
Inadequacies of low-rank models

- Sensitivity to order-preserving transformation



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

- Inadequacy for special structures.



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

Why sign matters?

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

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

- Sign tensors are invariant to order-preserving transformation.
- More flexible signal tensors are allowed by using sign tensor series representation.
- In noisy case, we estimate $\mathrm{sgn}(\Theta - \pi)$ from the tensor data $\mathrm{sgn}(\mathcal{Y} - \pi)$.

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 $\mathrm{sgn}(\Theta) = \mathrm{sgn}(\Theta')$.
- Sign rank is defined as

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

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 $\mathrm{sgn}(\Theta) = \mathrm{sgn}(\Theta')$.
- Sign rank is defined as

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

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

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 $\mathrm{sgn}(\Theta) = \mathrm{sgn}(\Theta')$.
- Sign rank is defined as

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

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

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

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

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

Sign representable tensors

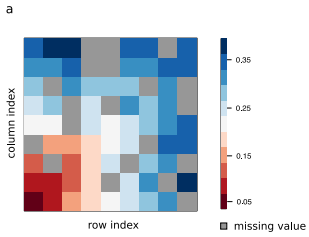
Sign representable tensors

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

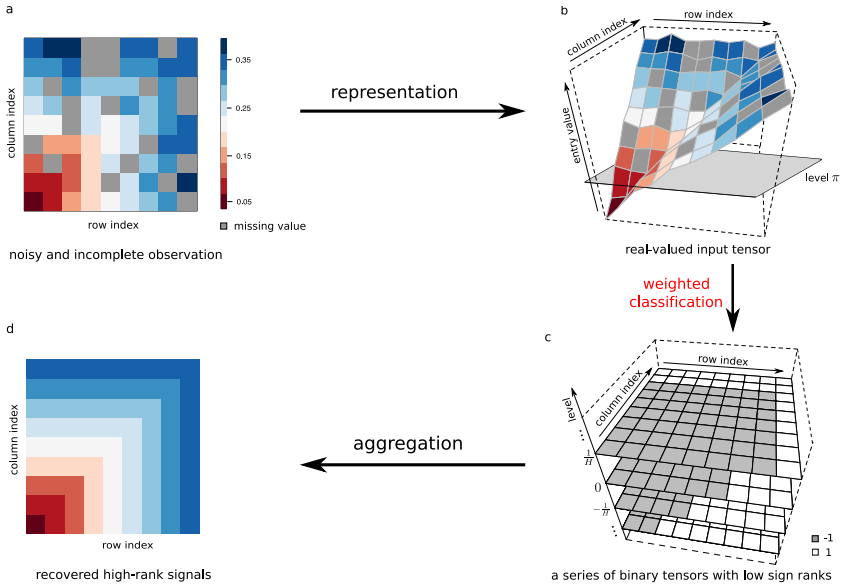
- Most existing structure tensors belong to sign representable family:
  - Low-rank CP tensors, Tucker tensors, stochastic block models.
  - High-rank tensors from GLM, single index models,
  - Tensors with repeating patterns, e.g. $\Theta(i_1, \ldots, i_K) = \log(1 + \max(i_1, \ldots, i_K))$ is 2-sign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

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

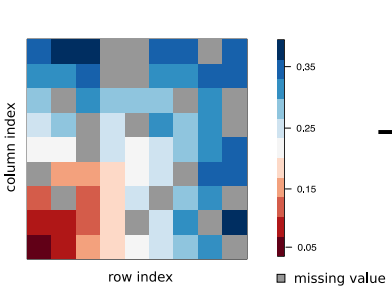
Our solution: sign signal helps!



Our solution: sign signal helps!

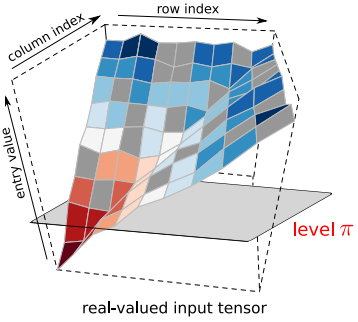


Step 1: representation



noisy and incomplete observation

representation

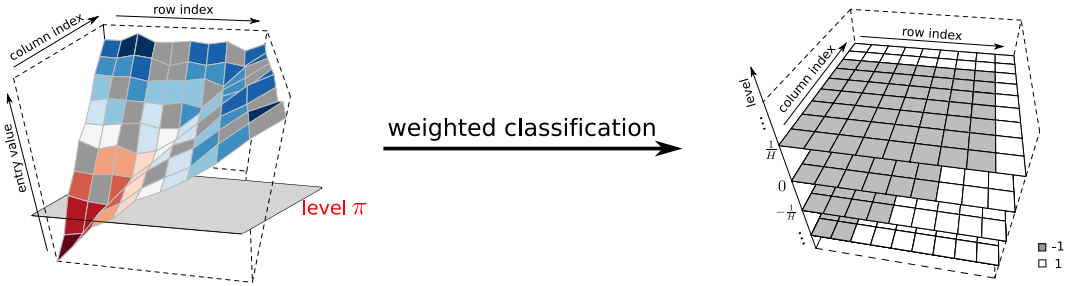


real-valued input tensor

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

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

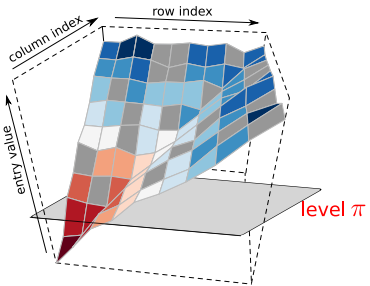
Step 2: weighted classification



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

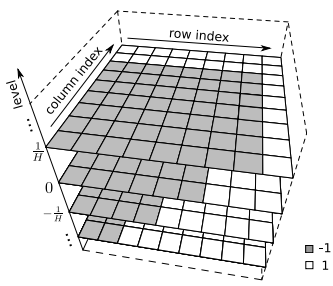
$$L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi) = \frac{1}{|\Omega|} \sum_{\pi \in \Omega} \underbrace{|\mathcal{Y}(\omega) - \pi|}_{\text{weight}} \times \underbrace{|\mathrm{sgn}(\mathcal{Z}(\omega)) - \mathrm{sgn}(\mathcal{Y}(\omega) - \pi)|}_{\text{classification loss}}$$

Step 2: weighed classification



real-valued input tensor

weighted classification



a series of binary tensors with low sign ranks

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

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

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

algorithm

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

Identification for sign tensor estimation

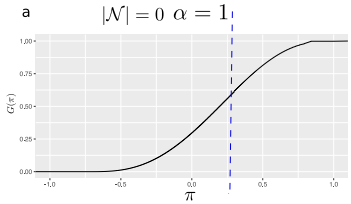
- We quantify difficulty of the problem using CDF $G(\pi) = \mathbb{P}_{\omega \in \Pi}[\Theta(\omega) \leq \pi]$.

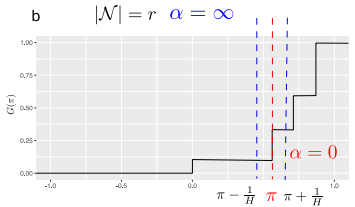
α-smoothness

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

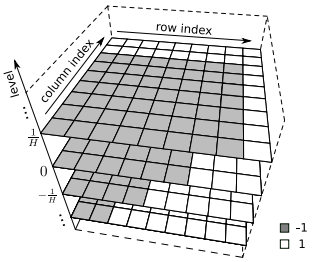
$$\sup_{\Delta s \leq t < \rho(\pi, \mathcal{N})} \frac{G(\pi + t) - G(\pi - t)}{t^\alpha} \leq c,$$

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



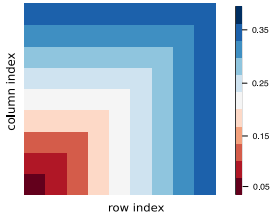


Step 3: aggregation



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

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

Estimation error

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

Estimation error (L. and Wang 2021)

Suppose $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ is α-smooth with bounded $|\mathcal{N}|$, and $d_1 = \cdots = d_K = d$.

1. (Sign tensor estimation) For all $\pi \in \mathcal{N}^c$, with high probability,

$$\mathrm{MAE}(\mathrm{sgn}\hat{\mathcal{Z}}_\pi, \mathrm{sgn}(\Theta - \pi)) \lesssim^* \left(\frac{dr}{|\Omega|}\right)^{\frac{\alpha}{\alpha+2}}.$$

2. (Tensor estimation)

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim^* \underbrace{\left(\frac{dr}{|\Omega|}\right)^{\frac{\alpha}{\alpha+2}}}_{\text{Error inherited from sign estimation}} + \underbrace{\frac{1}{H}}_{\text{Bias}} + \underbrace{\frac{Hdr}{|\Omega|}}_{\text{Variance}} \asymp^{**} \left(\frac{dr}{|\Omega|}\right)^{\min\left(\frac{\alpha}{\alpha+2}, \frac{1}{2}\right)}.$$

*log term suppressed, **$H \asymp (|\Omega|/dr)^{1/2}$

- Tensor estimation is generally no better than sign tensor estimation.
- See paper for general case that allows unbounded $|\mathcal{N}^c|$ and sub-Gaussian noise.

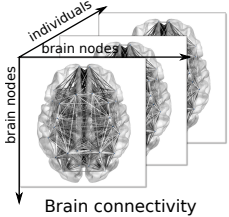
Comparison to existing results

Special case with full observation:

Model	Our rate (power of d)	Previous results
Tensor block model	$-(K-1)/2$	$\alpha = \infty$; minimax rate in Wang & 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
Generalized linear model	$-(K-1)/3$	$\alpha = 1$; close to parametric rate in L.& Wang '20
α-smooth $\mathscr{P}_{\text{sgn}}(r)$	$-(K-1)\min(\frac{\alpha}{\alpha+2} \wedge \frac{1}{2})$	faster rate as α increases

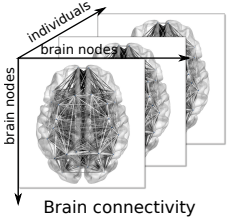
Reference: C. Lee and M. Wang. Beyond the Signs: Nonparametric tensor completion via sign series. arXiv:2102.00384, 2021.

Data application: Brain connectivity



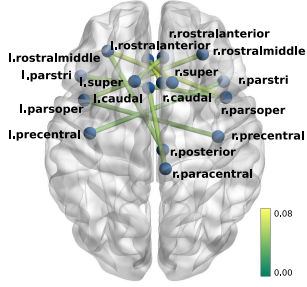
- The human brain connectivity dataset consists of 68 brain regions for 114 individuals with their IQ scores.
- Data tensor $\mathcal{Y} \in \{0,1\}^{68\times 68\times 114}$.

Data application: Brain connectivity

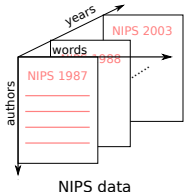


- The human brain connectivity dataset consists of 68 brain regions for 114 individuals with their IQ scores.
- Data tensor $\mathcal{Y} \in \{0,1\}^{68\times 68\times 114}$.

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

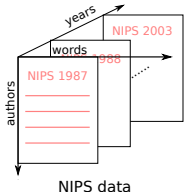


Data application: NIPS



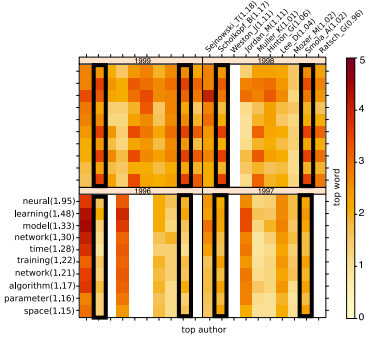
- The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003.
- Data tensor $\mathcal{Y} \in \mathbb{R}^{100 \times 200 \times 17}$.

Data application: NIPS



- 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 the active topics
- 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)	**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 5-folded cross-validation

- Our method outperforms the low-rank CP method in applications.

Outline

Nonparametric trace regression

- We extend the earlier method to a general nonparametric trace regression.
- Data: $\mathcal{X} \in \mathcal{D} \subset \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the tensor predictor, $Y \in \mathbb{R}$ the scalar response.
- Model:

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

  where $f : \mathcal{D} \to [-1, 1]$ is an unknown regression function of interest, and ϵ is mean-zero noise.
- Classical trace regression (Fan et al., 2019; Hamidi and Bayati, 2019) assumes

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

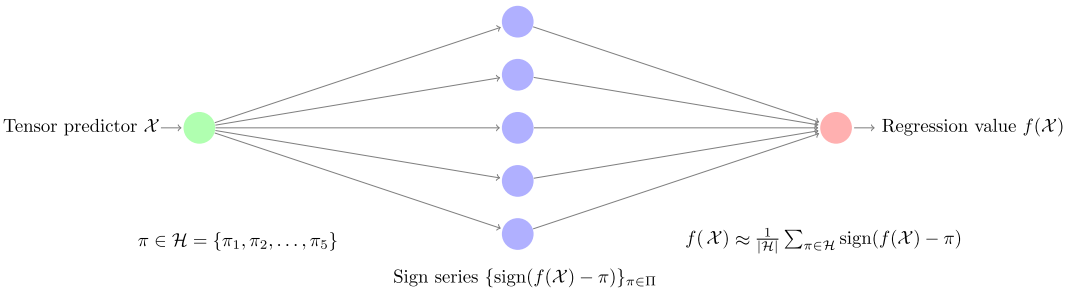
  where $\mathcal{B}$ is a low-rank tensor.
- Functional form of $f(\mathcal{X})$ is inadequate in many cases.

Rank-r sign representable function

- We apply our framework for the regression problem.
- We propose $f \in \mathcal{F}_{\text{sgn}}(r)$ belongs to rank-r sign representable functions, in that

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

  where $\text{rank}(\mathcal{B}_\pi) \leq r$ for all $\pi \in [-1, 1]$.
- We develop a learning reduction approach to solve regression using classifications.



Examples for rank-r sign representable function

Many existing function classes belong to sign representable function family.

- **Single index regression model** (Balabdaoui et al., 2019; Ganti et al., 2017):

$$f(\mathcal{X}) = g(\langle \mathcal{B}, \mathcal{X} \rangle), \text{ for unknown } g \text{ and rank-}r\ \mathcal{B}.$$

  By definition, $f \in \mathcal{F}_{\text{sgn}}(r)$.

- **Tensor completion** with $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ in previous section:
  - Predictor space $\mathcal{D} = \{e_i \otimes e_j \otimes e_k : (i, j, k) \in [d]^3\}$.
  - Tensor completion is a special case of estimating function $f : \mathcal{D} \to [-1, 1]$ such that

$$\Theta = [\![f(e_i \otimes e_j \otimes e_k)]\!] \in [-1, 1]^{d \times d \times d}.$$

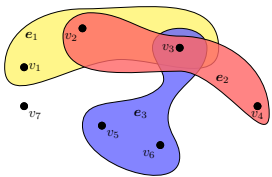
  In this setting, $f \in \mathcal{F}_{\text{sgn}}(r)$.

Summary

- We develop a new sign representation model that
  - embraces both linear and nonlinear trace effects;
  - addresses a richer class of structured tensors.
- This learning reduction from regression to classification provides a generic engine to empower existing algorithm for a wide range of structured tensor problems.
- We have extended the broad nonparametric paradigm to many learning problems:
  - tensor regression
  - tensor completion
  - multi-task learning
  - compressed sensing
- See full exposition in C. Lee and M. Wang. Beyond the Signs: Nonparametric tensor completion via sign series. arXiv:2102.00384, 2021.

Outline

Nonparametric tensor estimation from hypergraph



- Hypergraph considers higher-way interaction among nodes.
- We are currently developing a new tensor model using smooth function representation

$$\mathbb{E}(\mathcal{A}_\omega) = f(\xi_{\omega_1}, \xi_{\omega_2}, \cdots, \xi_{\omega_K}), \text{ for all } \omega \in [d] \times \cdots \times [d],$$

where $\{\xi_i\}_{i=1}^d \overset{\text{i.i.d.}}{\sim} \text{Unif}[0,1]$, and $\mathcal{A}$ is an adjacency tensor.

- Smooth functions are well approximated by piecewise constant function.
  $\Rightarrow$ Structured tensors are well approximated by stochastic block tensors,

$$\mathbb{E}(\mathcal{A}_\omega) = \mathcal{Q}_{z(\omega_1), z(\omega_2), \ldots z(\omega_K)}, \text{ for all } \omega \in [d] \times \cdots \times [d],$$

  where $\mathcal{Q} \in [0,1]^{m \times \cdots \times m}$ and $z \colon [d] \to [m]$ is a hidden partition.
- Close connection to K-uniform hypergraphons and shape-constrained regression.

Thank you!

Appendix: algorithm

back

Algorithm 1 Nonparametric tensor completion via learning reduction

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

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

2: Define π-shifted tensor $\bar{\mathcal{Y}} = \mathcal{Y} - \pi$ and corresponding sign tensor $\mathrm{sgn}(\bar{\mathcal{Y}}) = \mathrm{sgn}(\mathcal{Y} - \pi)$.

3: Perform 1-bit tensor estimation algorithm (Ghadermarzy et al., 2018; Wang and Li, 2020; Hong et al., 2020; Alquier et al., 2019) on $\bar{\mathcal{Y}}_\Omega$ and obtain

$$\hat{\mathcal{Z}}_\pi \leftarrow \underset{\text{low-rank } \mathcal{Z}}{\arg\min} \sum_{\omega \in \Omega} |\bar{\mathcal{Y}}(\omega)| F(\mathcal{Z}(\omega) \mathrm{sgn} \bar{\mathcal{Y}}(\omega)) + \lambda \|\mathcal{Z}\|_F^2,$$

where $F(\cdot)$ is the large-margin loss and λ is the penalty parameter.

4: **end for**

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

Appendix: theoretical guarantees for a large-margin loss classification

- we consider the estimation

$$\hat{\mathcal{Z}}_\pi = \underset{\text{rank}(\mathcal{Z})\le r}{\arg\min} \sum_{\omega\in\Omega} |\mathcal{Y}(\omega) - \pi| \times F(\mathcal{Z}(\omega)\text{sign}(\mathcal{Y}(\omega) - \pi) + \lambda\|\mathcal{Z}\|_F^2,$$

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 $\mathcal{Z}_\pi^{(n)} \in \mathscr{P}_{\text{sgn}}(r)$, such that $\text{Risk}_F(\mathcal{Z}_\pi^{(n)}) - \text{Risk}_F(\Theta - \pi) \le a_n$, for some sequence $a_n \to 0$ as $n \to \infty$. Furthermore, assume $\|\mathcal{Z}_\pi^{(n)}\|_F \le J$ for some constant $J > 0$.

(b) $F(z) = (1 - z)_+$ is hinge loss.

Appendix: 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, the resulting estimate from Algorithm 1 satisfies,

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).

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

Appendix: algorithmic optimality

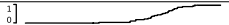
Empirical performance of the algorithm (Wang and Li, 2020)

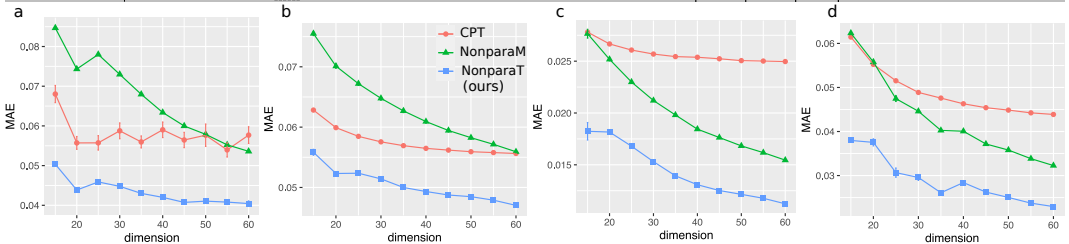
Under mild technical assumption on initialization, with high probability, there exists on iteration numnber $T_0 \geq 0$, such that,

$$Loss\left(\Theta^{(t)}, \Theta\right) \lesssim \underbrace{\rho^{t-T_0} Loss\left(\Theta^{(0)}, \Theta\right)}_{\text{algorithmic error}} + \underbrace{\sqrt{\frac{r^{K-1}\sum_k d_k}{\prod_k d_k}}}_{\text{statistical error}},$$

for all $t \geq T_0$, where $\rho \in (0,1)$ is a contraction parameter, *Loss* is logistic loss, and $\Theta^{(t)}$ is t-th iterates of Algorithm 1 .

Appendix: simulations for estimation error vs tensor dimension

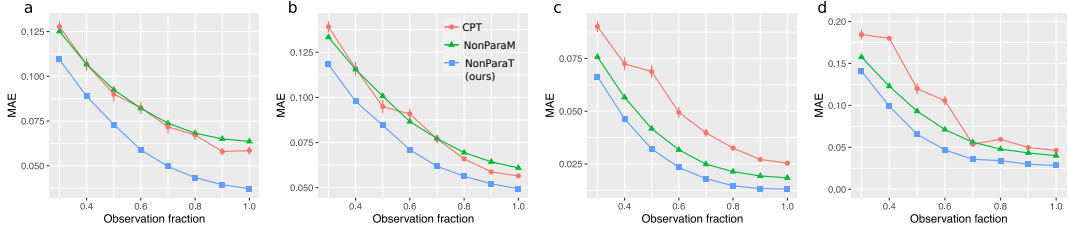
Simulation	Signal Tensor Θ	Rank	Sign Rank	α	$\lvert\mathcal{N}\rvert$	CDF of Tensor Entries	Noise
1	$\mathcal{C} \times \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \boldsymbol{M}_3$	3^3	$\leq 3^3$	∞	$\leq 3^3$		Uniform $[-0.3, 0.3]$
2	$\lvert \boldsymbol{a} \otimes \boldsymbol{1} \otimes \boldsymbol{1} - \boldsymbol{1} \otimes \boldsymbol{a} \otimes \boldsymbol{1} \rvert$	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	$2.5 - \exp(Z_{\min}^{1/3})$	$\geq d$	2	∞	d		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) achives the best performance.

Appendix: simulations for estimation error vs the observation fraction

Simulation	Signal Tensor Θ	Rank	Sign Rank	α	$\lvert\mathcal{N}\rvert$	CDF of Tensor Entries	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	$2.5 - \exp(Z_{\min}^{1/3})$	$\geq d$	2	∞	d		Normal $\mathcal{N}(0, 0.15)$



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

References I

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832.

Balabdaoui, F., Durot, C., and Jankowski, H. (2019). Least squares estimation in the monotone single index model. *Bernoulli*, 25(4B):3276–3310.

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.

Fan, J., Gong, W., and Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202.

Fan, J. and Udell, M. (2019). Online high rank matrix completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8690–8698.

References II

Ganti, R., Rao, N., Balzano, L., Willett, R., and Nowak, R. (2017). On learning high dimensional structured single index models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1898–1904.

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*.

Montanari, A. and Sun, N. (2018). Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425.

References III

Ongie, G., Willett, R., Nowak, R. D., and Balzano, L. (2017). Algebraic variety models for high-rank matrix completion. In *International Conference on Machine Learning*, pages 2691–2700.

Wang, L., Durante, D., Jung, R. E., and Dunson, D. B. (2017). Bayesian network–response regression. *Bioinformatics*, 33(12):1859–1866.

Wang, M. and Li, L. (2020). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *Journal of Machine Learning Research*, 21(154):1–38.