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

Chanwoo Lee

Department of Statistics
University of Wisconsin - Madison

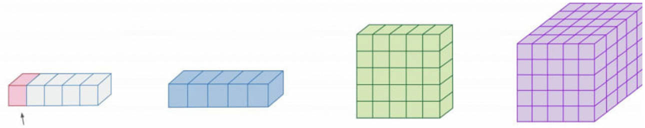
Preliminary Exam

My research

- ~~Broadly speaking,~~ my research interests lie at the intersection of **statistics, machine learning, and optimization**.
- ~~More specifically, my interests include~~ 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 **~~matrix or tensor~~ valued data**.

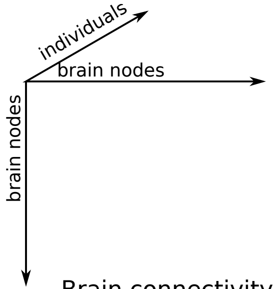
What is tensor?

- Tensors are generalizations of vectors and matrices:

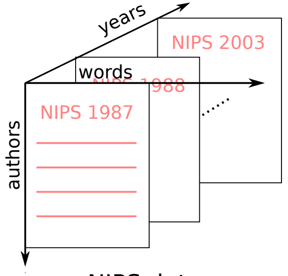


- We focus on tensors of order 3 or greater, also called higher-order tensors.
- Denote an order-$K(d_1, \cdots, d_K)$ dimensional tensor as $\mathcal{Y} = [\![y_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ where $\omega \in [d_1] \times \cdots \times [d_K]$.

Tensors in application



Brain connectivity



NIPS data

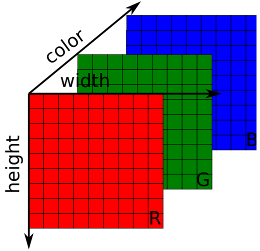


Image data

dataset (*data is plural*)

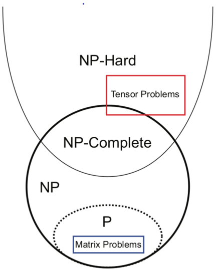
1. The ~~MRN-114~~ human brain connectivity ~~data~~ consists of 68 brain regions for 114 individuals (Wang et al., 2017).
2. The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003 along with author information (Globerson et al., 2007).
3. ~~The image data consists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

An RGB image consists of pixel values across three channels.

Talk outline

Prohibitive Computational Complexity

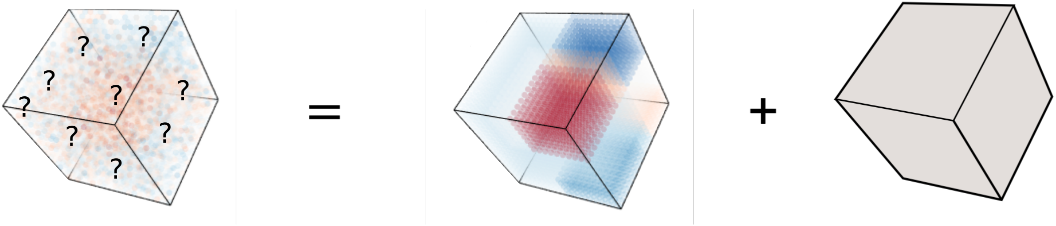
Most higher-order tensor problems are NP-hard [Hillar & Lim, 2013].



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

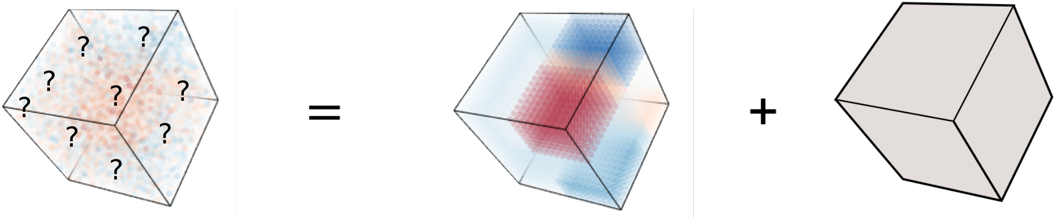
- Low-rankness
- Sparsity
- Non-negativity
- ...

Main problems: the signal plus noise model



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

Main problems: the signal plus noise model



$\underbrace{\mathcal{Y}}_{\text{Data tensor}} \qquad \underbrace{\Theta}_{\text{Signal tensor}} \qquad \underbrace{\mathcal{E}}_{\text{Noise tensor}}$

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

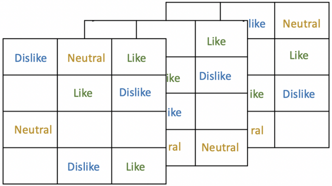
1. Previous work: parametric models for ordinal tensor completion

2. Current work: nonparametric tensor models via sign series
    2.1. Nonparametric tensor completion
    2.2. Nonparametric trace regression

3. Future work: other approaches for nonparametric tensor modeling

~~Probabilistic~~ model for ordinal-valued tensors~~: a cumulative link model~~

cumulative link



- 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. (fit into two lines?)
- We propose a cumulative link model,

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

~~ex) $f(x) = \frac{e^x}{1+e^x}$ is a logistic link.~~

where

- Theta = [\theta_\omega] represents a low-rank (highlight) latent parameter tensor prior to quantization.

- b = (b1,...,b_L-1} represents cutoff values in quantization.

- ~~If f is a cumulative function,~~

* f: a known (*highlight in red*) link function (quantization operation) is l

$$\cancel{\mathbb{P}(y_\omega = \ell) = f(b_\ell - \theta_\omega) - f(b_{\ell-1} - \theta_\omega) = \mathbb{P}(b_{\ell-1} < y^*_\omega \leq b_\ell),}$$

e.g. f(x)=.... is a logistic link

~~where $\epsilon_\omega \overset{i.i.d}{\sim} f$ and $y^*_\omega = \theta_\omega + \epsilon_\omega$.~~

Low rank assumption and estimation

Our work generlized earlier 1-bit tensor completion.

- ~~The signal tensor Θ assumes to have low-rank structure.~~
- ~~The log-likelihood associated with the observation is~~

We propose a rank-constrained M-estimate based on log-likelihood:

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

method

- ~~We propose a rank-constrained maximum likelihood estimation for Θ.~~
- The model 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, 2020.)

title in blue.

volume, page?

Low rank assumption and estimation

- The signal tensor Θ assumes to have low-rank structure.
- The log-likelihood associated with the observation is

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

- We propose a rank-constrained maximum likelihood estimation for Θ.
- The model 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, 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.

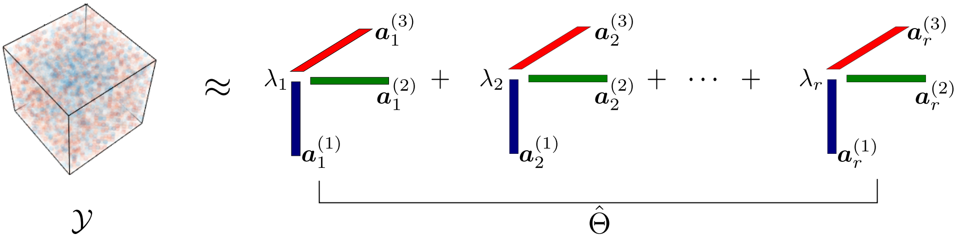
consistent

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

* low-rank models are inadequate in many cases. (I did not meat to delete this line in y
earlier version.)

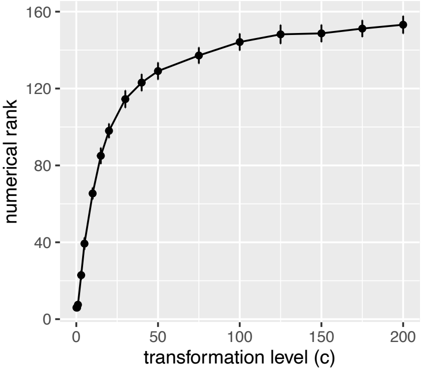
Inadequacies of low rank models

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



Inadequacies of low rank models

- ~~The~~ sensitivity to order-preserving transformation

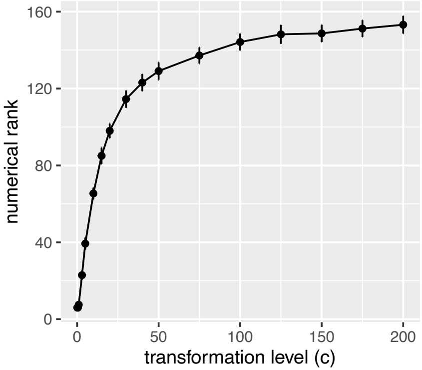


$$\Theta = \frac{1}{1+\exp(-c\,(\mathcal{Z}))}, \quad \text{where}$$

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

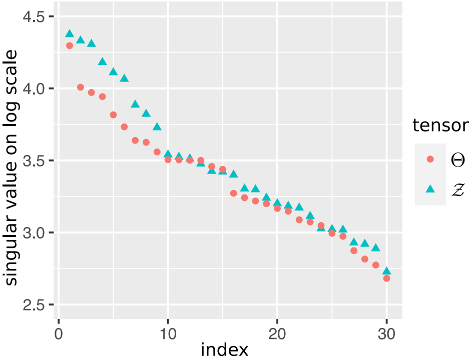
Inadequacies of low rank models

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



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

- The **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 $\Theta \in [-1,1]^{d_1\times\cdots\times d_K}$,

space?

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

- 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 $\text{sgn}(\Theta-\pi)$ from the tensor data $\text{sgn}(\mathcal{Y}-\pi)$.

Sign rank

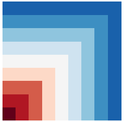
- Key idea: we use a local notion of low-rankness to allow a richer family of signal tensors.
- Two tensors are sign equivalent denoted as $\Theta \simeq \Theta'$ if $\text{sgn}(\Theta) = \text{sgn}(\Theta')$.
- Sign rank is defined as

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

Sign rank

- Key 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

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

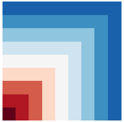
$\Theta =$  , $\quad \text{sgn}(\Theta) =$  $\Longrightarrow$

$$\begin{aligned} \text{rank}(\Theta) &= d \\ \text{srank}(\Theta) &= 2 \end{aligned}$$

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

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

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

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

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

Sign representable tensors

Sign representable tensors

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

- Most existing 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

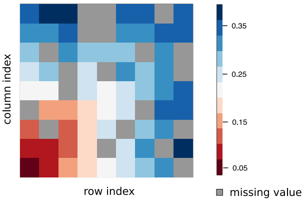
A tensor Θ is called r-sign representable if the tensor $(\Theta - \pi)$ has sign rank bounded by r for all $\pi \in [-1, 1]$.

- 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!

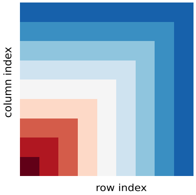
a



noisy and incomplete observation

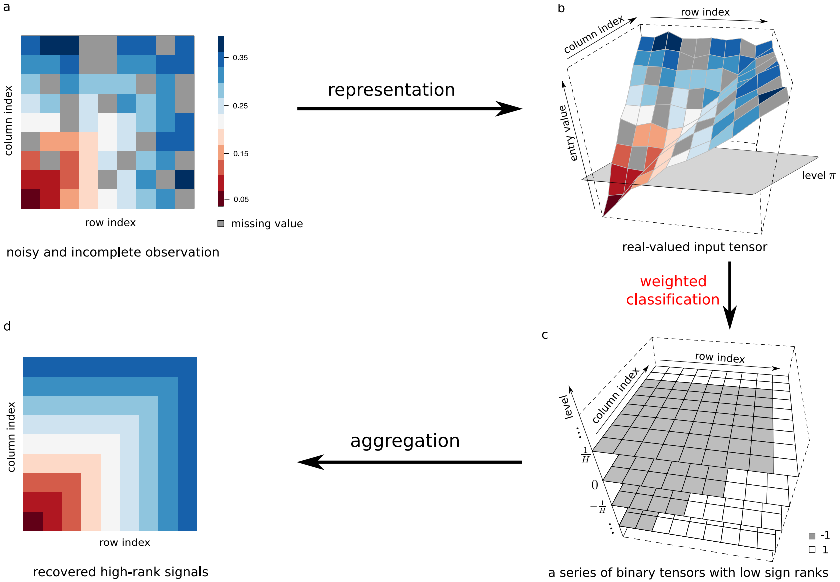
?

d

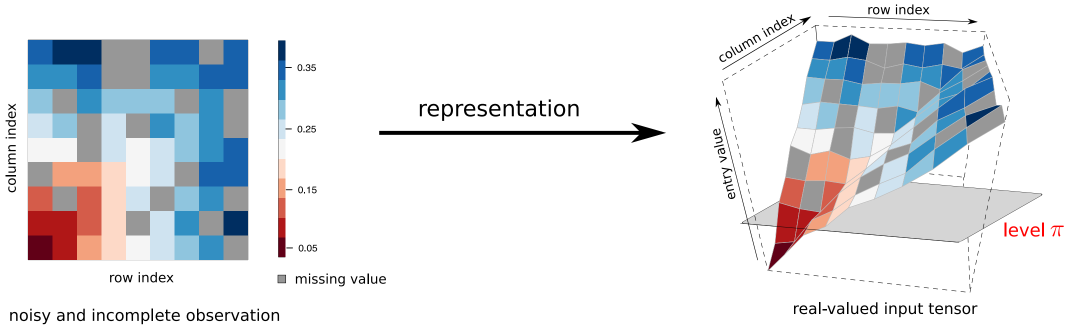


recovered high-rank signals

Our solution: sign signal helps!



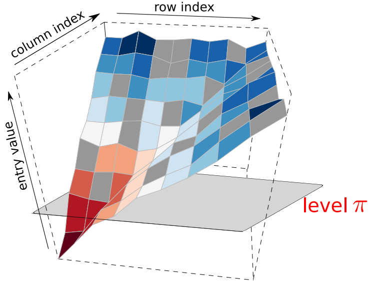
Step 1: representation



- 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:

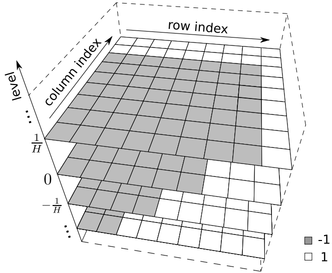
$$\{\text{sgn}(\mathcal{Y}_\Omega-\pi)\}_{\pi\in\mathcal{H}},\quad \text{where } \mathcal{H}=\left\{-1,\ldots,-\frac{1}{H},0,\frac{1}{H},\ldots,1\right\}.$$

Step 2: weighted classification



real-valued input tensor

weighted classification

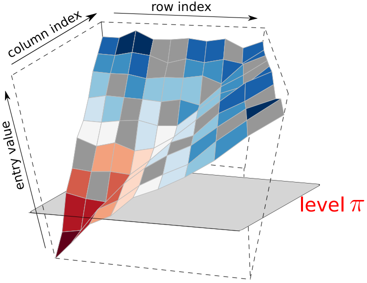


a series of binary tensors with low sign ranks

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

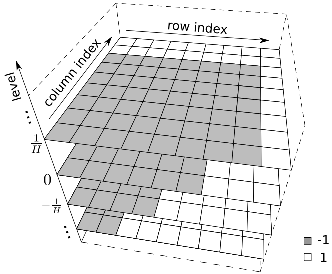
$$L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi) = \frac{1}{|\Omega|} \sum_{\pi \in \Omega} \underbrace{|\mathcal{Y}(\omega) - \pi|}_{\text{weight}} \times \underbrace{|\text{sgn}(\mathcal{Z}(\omega)) - \text{sgn}(\mathcal{Y}(\omega) - \pi)|}_{\text{classification loss}}$$

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

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

high light z.

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

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

Identification for sign tensor estimation

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

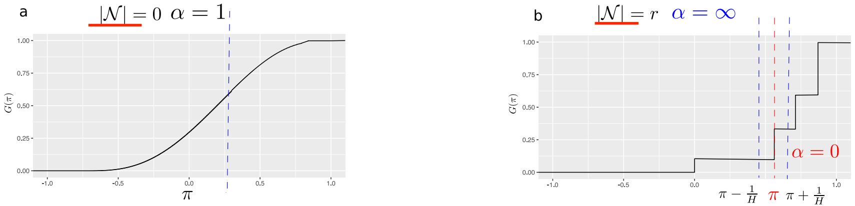
α-smoothness

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

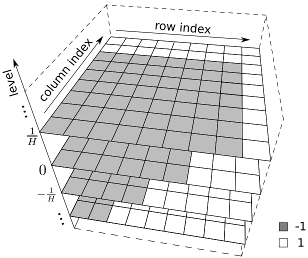
match N vs. N^c.

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

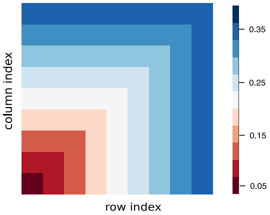
for two constants $\alpha, c > 0$, where $\rho(\pi, \mathcal{N}^c) = \min_{\pi' \in \mathcal{N}^c} |\pi - \pi'| + \Delta s$.



Step 3: aggregation



a series of binary tensors with low sign ranks



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

delete this page
(combine with page 25. In similar layout as your appendix 3.)

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

Sign tensor estimation for fixed π (L. and Wang, 2021)

Suppose $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ is α smooth for fixed π, and $d_1 = \cdots = d_K = d$. Then, with very high probability over $\mathcal{Y}_\Omega$,

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

*log term suppressed

- Sign estimation error shows a polynomial decay with the number of observed entries.

Tensor estimation error

Tensor estimation error (L. and Wang 2021)

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

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

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

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim^* \left(\frac{dr}{|\Omega|}\right)^{\min\left(\frac{\alpha}{\alpha+2}, \frac{1}{2}\right)}.$$

*log term suppressed

- 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:

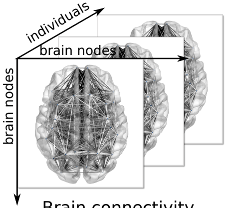
Wang

Model	Our rate (power of d)	Previous results
Tensor block model	$-(K-1)/2$	$\alpha=\infty$; minimax rate in W. & Zeng '19
Single index model	$-(K-1)/3$	$\alpha=1$; conjecture on the optimality; matrix rate $d^{-1/3}$ improves $\mathcal{O}(d^{-1/4})$ by Ganti et al. '18
Generalized linear model	$-(K-1)/3$	$\alpha=1$; close to parametric rate in W.& Li '20
α-smooth $\mathscr{P}_{\text{sgn}}(r)$	$-(K-1)\min(\frac{\alpha}{\alpha+2}\wedge\frac{1}{2})$	faster rate as α increases

change to
L. and Wang, 20
(highlight in blue)

reference: (add our paper)

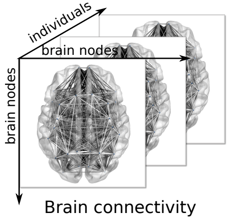
Data application: Brain connectivity



Brain connectivity

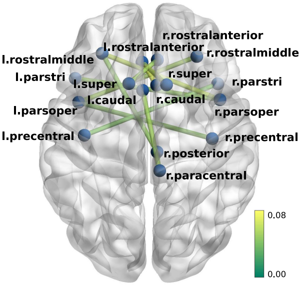
- The ~~MRN-114~~ human brain connectivity data consists of 68 brain regions for 114 individuals ~~along~~ with their IQ scores ~~(Wang et al., 2017).~~ (fit into two lines.)
- Data tensor $\mathcal{Y} \in \{0, 1\}^{68 \times 68 \times 114}$.

Data application: Brain connectivity

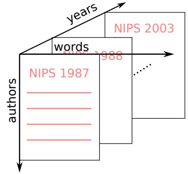


- The MRN-114 human brain connectivity data consists of 68 brain regions for 114 individuals along with their IQ scores (Wang et al., 2017).
- Data tensor $\mathcal{Y} \in \{0, 1\}^{68 \times 68 \times 114}$.

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



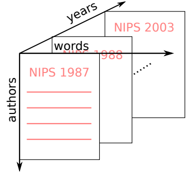
Data application: NIPS



NIPS data

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

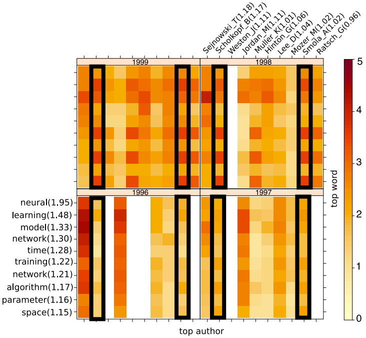
Data application: NIPS



NIPS data

- The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003 (Globerson et al., 2007).
- 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

in applications.

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

(fit into one line)

Outline

1. Previous work: parametric models for ordinal tensor completion

2. Current work: nonparametric tensor models via sign series
   2.1. Nonparametric tensor completion
   2.2. Nonparametric trace regression

3. Future work: other approaches for nonparametric tensor modeling

Nonparametric trace regression ~~via sign series~~

We extend the earlier method to a general nonparametric trace regression (*highlight in red*).

— Data:

— Model:

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

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

\tC [-1,1]

where $f: \mathbb{R}^{d_1 \times \cdots \times d_K} \to \mathbb{R}$ is an unknown regression function of interest, and ϵ is mean-zero noise.

highlight

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

(fit into one line)

Classical

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

"," highlight in red

where $\mathcal{B}$ is usually a low rank tensor.

- Functional form of $f(\mathcal{X})$ is not adequate in many cases.

inadequate

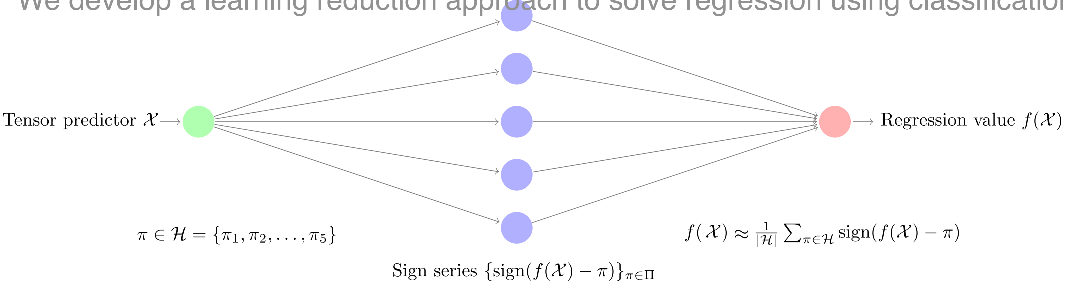
Rank-r sign representable function

- We apply our framework for the regression problem.
- We assume propose ~~that sign of regression function is represented by~~ f in tF-sign(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 call such functions rank-r sign representable function denoted by~~ ~~$f \in \mathcal{F}_{\text{sgn}}(r)$.~~

We develop a learning reduction approach to solve regression using classifications:



~~Some~~ 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) :
  ~~assumes~~ (fit into one line)

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

  ~~where g is unknown monotonic function and rank$(\mathcal{B}) \leq r$.~~
  By definition, $f \in \mathcal{F}_{\text{sgn}}(r)$.
- **Tensor completion** in previous section assumes $\Theta \in \mathscr{P}_{\text{sgn}}(r)$.
  - (bullet point)
  ~~Under the~~ predictor space $\mathcal{D} = \{e_i \otimes e_j \otimes e_k : (i, j, k) \in [d_1] \times [d_2] \times [d_3]\}$, the signal tensor Θ is represented by the bounded function $f : \mathcal{D} \to [-1, 1]$ such that
  - tensor completion is a special case of estimating regression function f: …

  $$\Theta = [\![f(e_i \otimes e_j \otimes e_k)]\!] \in [-1, 1]^{d_1 \times d_2 \times d_3}.$$

  [d]^3 (again use the simplest form)

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

We develop a new sign representation model that
— embraces both linear and nonlinear trace effects;
— addresses a richer class of structured tensors.

- ~~We estimate the series of sign functions $\{\text{sign}(f-\pi)\}_{\pi\in\mathcal{H}}$ from the weighted classification,~~

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

~~where $\phi\in\mathcal{F}_{\text{sign}}(r)$.~~

- This learning reduction from regression to classification provides a generic engine to empower existing algorithm for a wide range of structured tensor problems.

have extended

(again, try to fit into one line.)

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

4 bullet points:
— tensor regression
— tensor completion
— multi-task learning
— compressed sensing.

see full exposition in arxiv. …(title, author…)

Outline

1. Previous work: parametric models for ordinal tensor completion

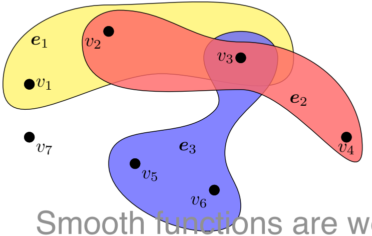
2. Current work: nonparametric tensor models via sign series
   2.1. Nonparametric tensor completion
   2.2. Nonparametric trace regression

3. Future work: other approaches for nonparametric tensor modeling

other nonparametric approaches for tensor models
(try to fit into one line)

Nonparametric ~~probability~~ tensor estimation from hypergraph

Close connection to k-uniform hypergraphons and shape-constrained regression



[3]
- Hypergraph considers higher-way interaction among nodes.
- ~~An observed adjacency tensor $\mathcal{A} \in \{0,1\}^{d\times\cdots\times d}$ corresponding to a hypergraph is generated by~~

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

Smooth functions are well approximated by piecewise constant function.

—> structured tensors are well approximated by stochastic block tensors.

[2]
- The block model (Ahn et al., 2018; Wang and Zeng, 2019; Han et al., 2020) has

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

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

[1]
- We will take nonparametric approach considering smooth function f such that

We are currently developing a new tensor model using smooth function representation

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

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

Thank you!

more back up slides.

— simulation.
— algorithmic optimality (copy thm from my 1-bit tensor paper; e.g.
``under mild technical assumption on initialization. .. error ~ algorithm + statistical")
— whatever you think might be helpful for addressing questions.

Appendix: Algorithm

change it to current version in NIPS.

Algorithm 1 Nonparametric tensor completion

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

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

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

  while not convergence **do**

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

      Update $\boldsymbol{A}_k$ while holding others fixed: $\boldsymbol{A}_k \leftarrow \arg\min_{\boldsymbol{A}_k \in \mathbb{R}^{d_k \times r}} \sum_{\omega \in \Omega} |\mathcal{Y}(\omega) - \pi| F(\mathcal{Z}(\omega)\text{sgn}(\mathcal{Y}(\omega) - \pi))$, where $F(\cdot)$ is the large-margin loss, and $\mathcal{Z} = \sum_{s \in [r]} \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}$ is a rank-r tensor.

    end for

  end while

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

end for

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

Theoretical guarantees for a large-margin loss classification

- we consider the estimation

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

  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.

Theoretical guarantees for a large-margin loss classification

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

Denote $t_n = \frac{d_{\max} rK \log n}{n}$. 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.

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.

References II

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.

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*.

References III

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*.

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

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 Zeng, Y. (2019). Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, pages 713–723.