Manuscript 1–21

Beyond the Signs: Nonparametric Tensor Completion via Sign Series

Summary: We consider the problem of tensor estimation from noisy observations with possibly missing entries. A nonparametric approach to tensor completion is developed based on a new model which we coin as sign representable tensors. The model represents the signal tensor of interest using a series of structured sign tensors. Unlike earlier methods, the sign series representation effectively addresses both low- and high-rank signals, while encompassing many existing tensor models—including CP models, Tucker models, single index models, several hypergraphon models—as special cases. We show that the sign tensor series is theoretically characterized, and computationally estimable, via classification tasks with carefully-specified weights. Excess risk bounds, estimation error rates, and sample complexities are established. We demonstrate the outperformance of our approach over previous methods on two datasets, one on human brain connectivity networks and the other on topic data mining.

Key words: Nonparametric learning, tensor completion, high dimension, classification.

This paper is under consideration for 2021 ICSA student paper award.

1. Introduction

Higher-order tensors have recently received much attention in enormous fields including social networks (Anandkumar et al., 2014), neuroscience (Wang et al., 2017), and genomics (Hore et al., 2016). Tensor methods provide effective representation of the hidden structure in multiway data. In this paper we consider the signal plus noise model,

$$\mathcal{Y} = \Theta + \mathcal{E}, \tag{1}$$

where $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is an order-K data tensor, Θ is an unknown signal tensor of interest, and $\mathcal{E}$ is a noise tensor. Our goal is to accurately estimate Θ from the incomplete, noisy observation of $\mathcal{Y}$. In particular, we focus on the following two problems:

- Q1 [Nonparametric tensor estimation]. How to flexibly estimate Θ under a wide range of structures, including both low-rankness and high-rankness?
- Q2 [Complexity of tensor completion]. How many observed tensor entries do we need to consistently estimate the signal Θ?

1.1 *Inadequacies of low-rank models*

The signal plus noise model (3) is popular in tensor literature. Existing methods estimate the signal tensor based on low-rankness of Θ (Jain and Oh, 2014; Montanari and Sun, 2018). Common low-rank models include Canonical Polyadic (CP) tensors (Hitchcock, 1927), Tucker tensors (De Lathauwer et al., 2000), and block tensors (Wang and Zeng, 2019). While these methods have shown great success in signal recovery, tensors in applications often violate the low-rankness. Here we provide two examples to illustrate the limitation of classical models.

The first example reveals the sensitivity of tensor rank to order-preserving transformations. Let $\mathcal{Z} \in \mathbb{R}^{30\times30\times30}$ be an order-3 tensor with CP rank$(\mathcal{Z}) = 3$ (formal definition is deferred to end of this section). Suppose a monotonic transformation $f(z) = (1 + \exp(-cz))^{-1}$ is applied to $\mathcal{Z}$ entrywise, and we let the signal Θ in model (1) be the tensor after transformation. Figure 1a plots the numerical rank ~~(see Appendix)~~ of Θ versus c. As we see, the rank increases

rapidly with c, rending traditional low-rank tensor methods ineffective in the presence of mild order-preserving nonlinearities. In digital processing (Ghadermarzy et al., 2018) and genomics analysis (Hore et al., 2016), the tensor of interest often undergoes unknown transformation prior to measurements. The sensitivity to transformation makes the low-rank model less desirable in practice.

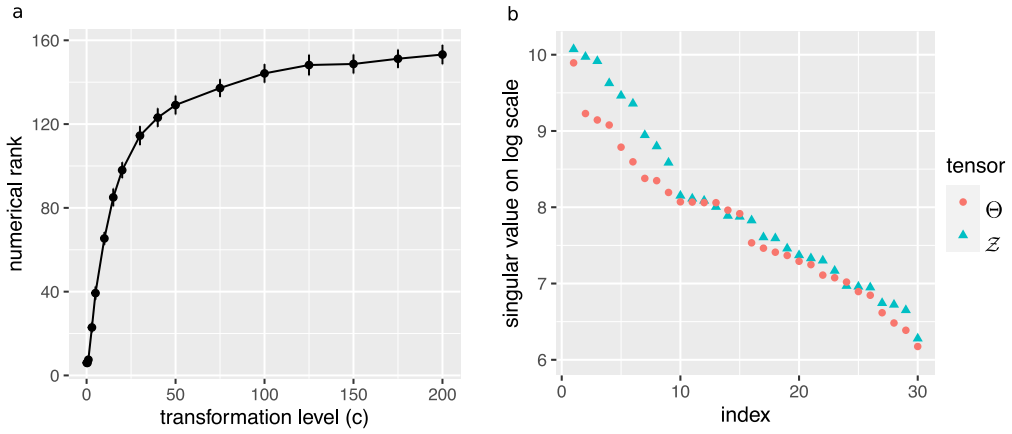


Figure 1: (a) Numerical rank of Θ versus c in the first example. (b) Top $d = 30$ tensor singular values in the second example.

The second example demonstrates the inadequacy of classical low-rankness in representing special structures. We consider the signal tensor of the form $\Theta = \log(1+\mathcal{Z})$, where $\mathcal{Z} \in \mathbb{R}^{d\times d\times d}$ is an order-3 tensor with entries $\mathcal{Z}(i,j,k) = \frac{1}{d}\max(i,j,k)$ for $i,j,k \in \{1,\ldots,d\}$. The matrix analogy of Θ was studied by Chan and Airoldi (2014) in graphon analysis. In this case neither Θ nor $\mathcal{Z}$ is low-rank; in fact, the rank is no smaller than the dimension d as illustrated in Figure 1b. Again, the low-rank models fail to address this type of tensor structure.

In the above and many other examples, the signal tensors Θ of interest have high rank. Classical low-rank models will miss these important structures. New methods that allow flexible tensor modeling have yet to be developed.

1.2 *Our contributions*

We develop a new model called sign representable tensors to address the aforementioned challenges. Figure 3 illustrates our main idea. Our approach is built on the sign series

representation of the signal tensor, and we propose to estimate the sign tensors through a series of weighted classifications. In contrast to existing methods, our method is guaranteed to recover a wide range of low- and high-rank signals. We highlight two main contributions that set our work apart from earlier literature.

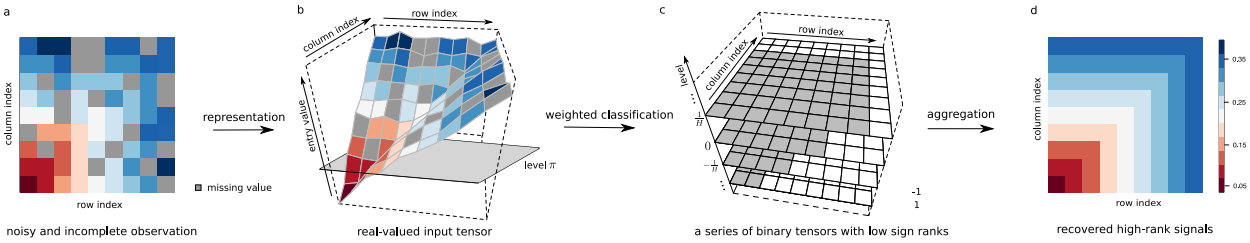


Figure 2: Illustration of our method. For visualization purpose, we plot an order-2 tensor (a.k.a. matrix); similar procedure applies to higher-order tensors. (a): a noisy and incomplete tensor input. (b) and (c): main steps of estimating sign tensor series $\mathrm{sgn}(\Theta - \pi)$ for $\pi \in \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$. (d) estimated signal $\hat{\Theta}$. The depicted signal is a full-rank matrix based on Example 5 in Section 3.

Statistically, the problem of high-rank tensor estimation is challenging. Existing estimation theory (Anandkumar et al., 2014; Montanari and Sun, 2018; Cai et al., 2019) exclusively focuses on the regime of fixed r growing d. However, such premise fails in high-rank tensors, where the rank may grow with, or even exceed, the dimension. A proper notion of nonparametric complexity is crucial. We show that the sign tensor series not only preserves all information in the original signals, but also brings the benefits of flexibility and accuracy over classical low-rank models. The results fill the gap between parametric (low-rank) and nonparametric (high-rank) tensors, thereby greatly enriching the tensor model literature.

From computational perspective, optimizations regarding tensors are in general NP-hard. Fortunately, tensors sought in applications are specially-structured, for which a number of efficient algorithms are available (Ghadermarzy et al., 2018; Wang and Li, 2020; Han et al., 2020). Our high-rank tensor estimate is provably reductable to a series of classifications, and its divide-and-conquer nature facilitates efficient computation. The ability to import and adapt existing tensor algorithms is one advantage of our method.

We also highlight the challenges associated with tensors compared to matrices. High-rank matrix estimation is recently studied under nonlinear models (Ganti et al., 2015) and subspace clustering (Ongie et al., 2017; Fan and Udell, 2019). However, the problem for high-rank tensors is more challenging, because the tensor rank often exceeds the dimension when order $K \geqslant 3$ (Anandkumar et al., 2017). We show that, applying matrix methods to higher-order tensors results in suboptimal estimates. A full exploitation of the higher-order structure is needed; this is another challenge we address in this paper.

1.3 *Notation*

We use $\mathrm{sgn}(\cdot)\colon \mathbb{R} \to \{-1, 1\}$ to denote the sign function, where $\mathrm{sgn}(y) = 1$ if $y \geqslant 0$ and -1 otherwise. We allow univariate functions, such as $\mathrm{sgn}(\cdot)$ and general $f\colon \mathbb{R} \to \mathbb{R}$, to be applied to tensors in an element-wise manner. We denote $a_n \lesssim b_n$ if $\lim_{n\to\infty} a_n/b_n \leqslant c$ for some constant $c \geqslant 0$. We use the shorthand $[n]$ to denote the n-set $\{1, \ldots, n\}$ for $n \in \mathbb{N}_+$. Let $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K $(d_1, \ldots, d_K)$-dimensional tensor, and $\Theta(\omega) \in \mathbb{R}$ denote the tensor entry indexed by $\omega \in [d_1] \times \cdots \times [d_K]$. An event E is said to occur "with very high probability" if $\mathbb{P}(E)$ tends to 1 faster than any polynomial of tensor dimension $d := \min_k d_k \to \infty$. The CP decomposition (Hitchcock, 1927) is defined by

$$\Theta = \sum_{s=1}^{r} \lambda_s \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}, \tag{2}$$

where $\lambda_1 \geqslant \cdots \geqslant \lambda_r > 0$ are tensor singular values, $\boldsymbol{a}_s^{(k)} \in \mathbb{R}^{d_k}$ are norm-1 tensor singular vectors, and $\otimes$ denotes the outer product of vectors. The minimal $r \in \mathbb{N}_+$ for which (2) holds is called the tensor rank, denoted $\mathrm{rank}(\Theta)$.

2. Model and proposal overview

Let $\mathcal{Y}$ be an order-K $(d_1, \ldots, d_K)$-dimensional data tensor generated from the model,

$$\mathcal{Y} = \Theta + \mathcal{E}, \tag{3}$$

where $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is an unknown signal tensor of interest, and $\mathcal{E}$ is a noise tensor consisting of ~~mean-zero, independently~~ zero-mean, independently but not necessarily identically distributed entries. We allow heterogenous noise, in that the marginal distribution of noise entry $\mathcal{E}(\omega)$ may depend on ω. ~~Assume that $\mathcal{Y}(\omega)$ takes value in a bounded interval $[-A, A]$; without loss of generality, we set $A = 1$ throughout the paper. We extend unbounded observation with~~ For simplicity, we assume the noise is bounded; the extension to a sub-Gaussian noise ~~in Appendix.~~ is provided in Appendix (omitted in current version). In the main paper we assume the range of $\mathcal{Y}$ is the bounded interval $[-1, 1]$ for cleaner exposition.

Our observation is an incomplete data tensor from (3), denoted $\mathcal{Y}_\Omega$, where $\Omega \subset [d_1] \times \cdots \times [d_K]$ is the index set of observed entries. We consider a general model on Ω that allows both uniform and non-uniform samplings. Specifically, let $\Pi = \{p_\omega\}$ be an arbitrarily predefined probability distribution over the full index set with $\sum_{\omega \in [d_1] \times \cdots \times [d_K]} p_\omega = 1$. Assume that the entries ω in Ω are i.i.d. draws with replacement from the full index set using distribution Π.

Before describing our main results, we provide the intuition behind our method. In the two examples in Section 1, the high-rankness in the signal Θ makes the estimation challenging. Let us examine the sign of the π-shifted signal $\mathrm{sgn}(\Theta - \pi)$ for any given $\pi \in [-1, 1]$. It turns out that these sign tensors share the same sign patterns as low-rank tensors. Indeed, the signal tensor in the first example has the same sign pattern as a rank-4 tensor, since $\mathrm{sgn}(\Theta - \pi) = \mathrm{sgn}(\mathcal{Z} - f^{-1}(\pi))$. The signal tensor in the second example has the same sign pattern as a rank-2 tensor (see Example 5 in Section 3).

The above observation suggests a general framework to estimate both low- and high-rank signal tensors. Figure 3 illustrates the main crux of our method. We dichotomize the data tensor into a series of sign tensors $\mathrm{sgn}(\mathcal{Y}_\Omega - \pi)$ for $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$. Then, we estimate the sign signals $\mathrm{sgn}(\Theta - \pi)$ by performing classification

$$\hat{\mathcal{Z}}_\pi = \underset{\text{low rank tensor } \mathcal{Z}}{\arg\min} \text{ Weighted-Loss}(\mathrm{sgn}(\mathcal{Z}), \mathrm{sgn}(\mathcal{Y}_\Omega - \pi)),$$

where Weighted-Loss$(\cdot,\cdot)$ denotes a carefully-designed classification objective function which will be described in later sections. Our final proposed tensor estimate takes the form

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

Our approach is built on the nonparametric sign representation of signal tensors. The estimate $\hat{\Theta}$ is essentially learned from dichotomized tensor series $\{\mathrm{sgn}(\mathcal{Y}_\Omega - \pi)\colon \pi \in \mathcal{H}\}$ with proper weights. We show that a careful aggregation of dichotomized data not only preserves all information in the original signals, but also brings benefits of accuracy and flexibility over classical low-rank models. Unlike traditional methods, the sign representation is guaranteed to recover both low- and high-rank signals that were previously impossible. The method enjoys statistical effectiveness and computational efficiency.

3. Statistical properties of sign representable tensors

This section develops sign representable tensor models for Θ in (3). We characterize the algebraic and statistical properties of sign tensor series, which serves ~~the~~ our theoretical foundation.

3.1 *Sign-rank and sign tensor series*

Let Θ be the tensor of interest, and $\mathrm{sgn}(\Theta)$ the corresponding sign pattern. The sign patterns induce an equivalence relationship between tensors. Two tensors are called sign equivalent, denoted $\simeq$, if they have the same sign pattern.

DEFINITION 1 (Sign-rank). The sign-rank of a tensor $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is defined by the minimal rank among all tensors that share the same sign pattern as Θ; i.e.,

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

The sign-rank is also called ~~*support rank*~~support rank (Cohn and Umans, 2013), ~~*minimal rank*~~minimal rank (Alon et al., 2016), and ~~*nondeterministic rank*~~nondeterministic rank (De Wolf, 2003). Earlier work defines sign-rank for binary-valued tensors; we extend the notion to

continuous-valued tensors. Note that the sign-rank concerns only the sign pattern but discards the magnitude information of Θ. In particular, $\mathrm{srank}(\Theta) = \mathrm{srank}(\mathrm{sgn}\Theta)$.

Like most tensor problems (Hillar and Lim, 2013), determining the sign-rank for a general tensor is NP hard (Alon et al., 2016). Fortunately, tensors arisen in applications often possess special structures that facilitate analysis. By definition, the sign-rank is upper bounded by the tensor rank. More generally, we have the following upper bounds.

PROPOSITION 1 (Upper bounds of the sign-rank). For any strictly monotonic function $g\colon \mathbb{R} \to \mathbb{R}$ with $g(0) = 0$,

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

we have $\mathrm{srank}(\Theta) \leqslant \mathrm{rank}(g(\Theta))$.

Conversely, the sign-rank can be much smaller than the tensor rank as shown earlier.

PROPOSITION 2 (Broadness). For every order $K \geqslant 2$ and dimension d, there exist tensors $\Theta \in \mathbb{R}^{d\times\cdots\times d}$ such that $\mathrm{rank}(\Theta) \geqslant d$ but $\mathrm{srank}(\Theta - \pi) \leqslant 2$ for all $\pi \in \mathbb{R}$.

~~We provide several examples in Appendix, in which the~~ The two examples in Section 1 fall into this category where tensor rank grows with dimension d but the sign-rank remains a constant. The results highlight the advantages of using sign-rank in the high-dimensional tensor analysis. Propositions 1 and 2 together demonstrate the strict broadness of low sign-rank family over the usual low-rank family.

We now introduce "sign representable tensors" for the signal model in (3).

DEFINITION 2 (Sign representable tensors). Fix a level $\pi \in [-1, 1]$. A tensor Θ is called (r, π)-sign representable, if the tensor $(\Theta - \pi)$ has sign-rank bounded by r. A tensor Θ is called r-sign (globally) representable, if Θ is (r, π)-sign representable for all $\pi \in [-1, 1]$. The collection $\{\mathrm{sgn}(\Theta - \pi)\colon \pi \in [-1, 1]\}$ is called the sign tensor series. We use

~~$\mathscr{P}_{\text{sgn}}(r) = \{\Theta\colon \text{srank}(\Theta - \pi) \leqslant r \text{ for all } \pi \in [-1,1]\}$~~ $\mathscr{P}_{\text{sgn}}(r) = \{\Theta\colon \max_{\pi\in[-1,1]} \text{srank}(\Theta - \pi) \leqslant r\}$ to denote the r-sign representable tensor family.

We show that the r-sign representable tensor family is a general model that incorporates most existing tensor models, including low-rank tensors, single index models, GLM models, and several hypergraphon models.

EXAMPLE 1 (CP/Tucker low-rank models). The CP and Tucker low-rank tensors are the two most popular tensor models (Kolda and Bader, 2009). Let Θ be a low-rank tensor with CP rank r. We see that Θ belongs to the sign representable family; i.e., $\Theta \in \mathscr{P}_{\text{sgn}}(r+1)$ (the constant 1 is due to $\text{rank}(\Theta - \pi) \leqslant r+1$). Similar results hold for Tucker low-rank tensors $\Theta \in \mathscr{P}_{\text{sgn}}(r+1)$, where $r = \prod_k r_k$ with r_k being the k-th mode Tucker rank of Θ.

EXAMPLE 2 (Tensor block models (TBMs)). Tensor block model (Wang and Zeng, 2019; Chi et al., 2020) assumes a ~~checkerbord~~ checkerboard structure among tensor entries under marginal index permutation. The signal tensor Θ takes at most r distinct values, where r is the total number of multiway blocks. Our model incorporates TBM because $\Theta \in \mathscr{P}_{\text{sgn}}(r)$.

EXAMPLE 3 (Generalized linear models (GLMs)). Let $\mathcal{Y}$ be a binary tensor from a logistic model (Wang and Li, 2020) with mean $\Theta = \text{logit}(\mathcal{Z})$, where $\mathcal{Z}$ is a latent low-rank tensor. Notice that Θ itself may be high-rank (see Section 1). By definition, Θ is a low-rank sign representable tensor. Same conclusion holds for general exponential-family models with a (known) link function (Hong et al., 2020).

EXAMPLE 4 (Single index models (SIMs)). Single index model is a flexible semiparametric model proposed in economics (Robinson, 1988) and high-dimensional statistics (Balabdaoui et al., 2019; Ganti et al., 2017). We here extend the model to higher-order tensors Θ. The SIM assumes the existence of a (unknown) monotonic function $g\colon \mathbb{R} \to \mathbb{R}$ such that $g(\Theta)$ has rank r. We see that Θ belongs to the sign representable family; i.e., $\Theta \in \mathscr{P}_{\text{sgn}}(r+1)$.

EXAMPLE 5 (Min/Max hypergraphon). Graphon is a popular nonparametric model for networks (Chan and Airoldi, 2014; Xu, 2018). Here we revisit the model introduced in Section 1 for generality. Let Θ be an order-K tensor generated from the hypergraphon $\Theta(i_1, \ldots, i_K) = \log(1+\max_k x_{i_k}^{(k)})$, where $x_{i_k}^{(k)}$ are given number in $[0, 1]$ for all $i_k \in [d_k], k \in [K]$. We conclude that $\Theta \in \mathscr{P}_{\text{sgn}}(2)$, because the sign tensor $\text{sgn}(\Theta - \pi)$ with an arbitrary $\pi \in (0,\ \log 2)$ is a block tensor with at most two blocks (see Figure 3c).

The results extend to general min/max hypergraphons. Let $g(\cdot)$ be a continuous univariate function with at most $r \geqslant 1$ distinct real roots in the equation $g(z) = \pi$; e.g., when $g(z)$ is a polynomial of degree r. Then, the tensor Θ generated from $\Theta(i_1, \ldots, i_K) = g(\max_k x_{i_k}^{(k)})$ belongs to $\mathscr{P}_{\text{sgn}}(2r)$. Same conclusion holds if the maximum is replaced by the minimum.

3.2 *Statistical characterization of sign tensors via weighted classification*

Accurate estimation of a sign representable tensor depends on the behavior of sign tensor series, $\text{sgn}(\Theta - \pi)$. In this section, we show that sign tensors are completely characterized by weighted classification. The results bridge the algebraic and statistical properties of sign representable tensors.

For a given $\pi \in [-1, 1]$, define a π-shifted data tensor $\bar{\mathcal{Y}}_\Omega$ with entries $\bar{\mathcal{Y}}(\omega) = (\mathcal{Y}(\omega) - \pi)$ for $\omega \in \Omega$. We propose a weighted classification objective function

$$L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \underbrace{|\bar{\mathcal{Y}}(\omega)|}_{\text{weight}} \times \underbrace{|\text{sgn}\mathcal{Z}(\omega) - \text{sgn}\bar{\mathcal{Y}}(\omega)|}_{\text{classification loss}}, \tag{4}$$

where $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the decision variable to be optimized, $|\bar{\mathcal{Y}}(\omega)|$ is the entry-specific weight equal to the distance from the tensor entry to the target level π. The entry-specific weights incorporate the magnitude information into classification, where entries far away from the target level are penalized more heavily in the objective. In the special case of binary tensor $\mathcal{Y} \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$ and target level $\pi = 0$, the loss (4) reduces to usual classification loss. Our proposed weighted classification function (4) is important for characterizing $\text{sgn}(\Theta - \pi)$.

Define the weighted classification risk

$$\mathrm{Risk}(\mathcal{Z}) = \mathbb{E}_{\mathcal{Y}_\Omega} L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega), \tag{5}$$

where the expectation is taken with respect to $\mathcal{Y}_\Omega$ under model (3) and the sampling distribution $\omega \sim \Pi$. Note that the form of $\mathrm{Risk}(\cdot)$ implicitly depends on π; we suppress π when no confusion arises.

PROPOSITION 3 (Global optimum of weighted risk). Suppose the data $\mathcal{Y}_\Omega$ is generated from model (3) with $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$. Then, for all $\bar{\Theta}$ that are sign equivalent to $\mathrm{sgn}(\Theta - \pi)$,

$$\mathrm{Risk}(\bar{\Theta}) = \inf\{\mathrm{Risk}(\mathcal{Z}) \colon \mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}\} = \inf\{\mathrm{Risk}(\mathcal{Z}) \colon \mathrm{rank}(\mathcal{Z}) \leqslant r\}.$$

The results show that the sign tensor $\mathrm{sgn}(\Theta - \pi)$ optimizes the weighted classification risk. This fact suggests a practical procedure to estimate $\mathrm{sgn}(\Theta - \pi)$ via empirical risk optimization of $L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega)$. In order to establish the recovery guarantee, we shall address the uniqueness (up to sign equivalence) for the optimizer of $\mathrm{Risk}(\cdot)$. The local behavior of Θ around π turns out to play a key role in the accuracy.

Some additional notation is needed ~~. We introduce a tolerance $\Delta s = 1/\prod_{i=1}^K d_i$ to account for discrete measure with finite tensor dimension. Let $\mathcal{N}$ be the set of mass points of~~ for stating the results in full generality. Let $d_t = \prod_{k=1}^K d_k$ denote the total number of tensor entries, and $\Delta s = 2/d_t$ a small tolerance. We quantify distribution of entries in tensor Θ ~~under Π, whose probability mass in a~~ using pseudo density (a.k.a. histogram with bin width Δs~~-neighborhood is heavier than the tolerance level~~). Specifically, let $G(\pi) := \mathbb{P}_{\omega \sim \Pi}[\Theta(\omega) \leqslant \pi]$ denote the cumulative distribution function (CDF) of $\Theta(\omega)$ under $\omega \sim \Pi$. We partition $[-1, 1] = \mathcal{N} \cup \mathcal{N}^c$, where $\mathcal{N}$ consists of levels whose pseudo density based on Δs-bin is asymptotically unbounded; i.e.~~,~~

$$\mathcal{N} = \{\pi \colon \mathbb{P}_{\omega \sim \Pi}(|\Theta(\omega) - \pi| \leqslant \Delta s) > C \Delta s\}, \text{ for some universal constant } C > 0.$$

$$\mathcal{N} = \left\{ \pi \in [-1,1] \colon \frac{G(\pi+\Delta s/2) - G(\pi-\Delta s/2)}{\Delta s} \geqslant C \right\}, \text{ for some universal constant } C > 0,$$

and $\mathcal{N}^c$ otherwise. Note that both ~~Π and~~ Θ and its induced CDF G implicitly depend on the tensor dimension. ~~Our assumptions are imposed to $\Pi = \Pi(d)$ and $\Theta = \Theta(d)$~~ We impose assumptions to $G = G_d$ in the high-dimensional regime uniformly ~~where $d := \min_k d_k \to \infty$~~as $d := \min_k d_k \to \infty$.

Assumption 1 (α-smoothness). Fix ~~$\pi \notin \mathcal{N}$~~$\pi \notin \mathcal{N}$. Assume there exist constants $\alpha = \alpha(\pi) > 0, c = c(\pi) > 0$, independent of tensor dimension, such that,

$$\sup_{\Delta s \leqslant t < \rho(\pi,\mathcal{N})} \frac{\mathbb{P}_{\omega\sim\Pi}[|\Theta(\omega) - \pi| \leqslant t]}{t^\alpha} \sup_{\Delta s \leqslant t < \rho(\pi,\mathcal{N})} \frac{G(\pi + t/2) - G(\pi - t/2)}{t^\alpha} \leqslant c, \qquad (6)$$

where $\rho(\pi, \mathcal{N}) := \min_{\pi' \in \mathcal{N}} |\pi - \pi'|$ denotes the distance from π to the nearest point in $\mathcal{N}$. The largest possible $\alpha = \alpha(\pi)$ in (6) is called the smoothness index at level π. We make the ~~convention that $\alpha = \infty$ if the set $\{\omega \colon |\Theta(\omega) - \pi| \leqslant t\}$ has zeromeasure~~conversion that $\alpha = \infty$ if the numerator in (6) is zero, implying almost no entries of ~~which~~ $\Theta(\omega)$ ~~is~~ around the level π. We call a tensor Θ is α-globally smooth, if (6) holds with ~~a global constant $c > 0$ for all $\pi \in [-1, 1]$ except for a finite number of levels~~global constants $\alpha > 0, c > 0$ for all $\pi \notin \mathcal{N}$.

The smoothness index α quantifies the intrinsic hardness of recovering $\operatorname{sgn}(\Theta - \pi)$ from Risk$(\cdot)$. The value of α depends on both the sampling distribution $\omega \sim \Pi$ and the behavior of $\Theta(\omega)$. The recovery is easier at levels where points are less concentrated around π with a large value of $\alpha > 1$, or equivalently, when ~~the cumulative distribution function (CDF) $G(\pi) := \mathbb{P}_{\omega\sim\Pi}[\Theta(\omega) \leqslant \pi]$~~ $G(\pi)$ remains almost flat around π. A small value of $\alpha < 1$ indicates the nonexistent (infinite) density at level π, or equivalently, when the $G(\pi)$ jumps by greater than the ~~threshold level~~ tolerance Δs at π. ~~Table 1 illustrates the~~ Fig 3 illustrates three examples of $G(\pi)$~~for various models of Θ~~.

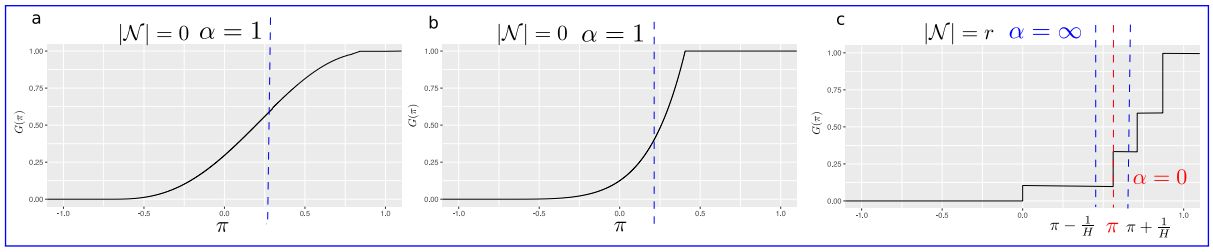


Figure 3: Three examples of CDF, $G(\pi) = \mathbb{P}_{\boldsymbol{X}}(f(\boldsymbol{X}) \leqslant \pi)$, and the associated smoothness index α. (a) and (b) $G(\pi)$ with $\alpha = 1$ because the function $G(\pi)$ has no large jumps in the range of π. (c) $G(\pi)$ with $\alpha = \infty$ at most π (in blue) except for a few jump points (in red). The dashed lines correspond to local (α, π)-smoothness.

We now reach the main theorem in this section. For two tensors Θ_1, Θ_2, define the mean absolute error (MAE) as $\mathrm{MAE}(\Theta_1, \Theta_2) \stackrel{\text{def}}{=} \mathbb{E}_{\omega \sim \Pi} |\Theta_1(\omega) - \Theta_2(\omega)|$.

THEOREM 1 (Identifiability). *Under Assumption 1, for all tensors* $\bar{\Theta} \simeq \mathrm{sgn}(\Theta - \pi)$ *and tensors* $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$,

$$\mathrm{MAE}(\mathrm{sgn}\mathcal{Z}, \mathrm{sgn}\bar{\Theta}) \leqslant C(\pi) \left[\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})\right]^{\alpha/(\alpha+1)},$$

where $C(\pi) > 0$ *is independent of* $\mathcal{Z}$.

The result establishes the recovery stability of sign tensors $\mathrm{sgn}(\Theta - \pi)$ using optimization with population risk (5). The bound immediately shows the uniqueness of the optimizer for $\mathrm{Risk}(\cdot)$ up to a zero-measure set under Π. We find that a higher value of α implies more stable recovery, as intuition would suggest. Similar results hold for optimization with sample risk (4) (see Section 4).

We conclude this section by applying Assumption 1 to the examples described in Section 3.1. For simplicity, suppose Π is the uniform sampling. The tensor block model is ∞-globally smooth. This is because the set $\mathcal{N}$ consists of finite Δs-~~neighbors whose centers are~~ -bin's covering the distinct block means in Θ. Furthermore, we have $\alpha = \infty$ for all $\pi \notin \mathcal{N}$, since the numerator in (6) is zero. Similarly, the min/max hypergaphon model such that $\Theta(i_1, \ldots, i_K) = \log(1 + \max_{\ell=1,\ldots K} i_\ell/d)$ is ∞-globally smooth because $\alpha = \infty$ for all π except

~~points~~ those in $\mathcal{N}$ ~~consisting of at most~~ , where $\mathcal{N}$ collects d ~~number of~~ many Δs-~~neighbors centering at~~ -bin's covering $\log(1+i/d)$ for $i = 1, \ldots, d$.

4. Nonparametric tensor completion via sign series

In previous sections we have established the sign series representation and its relationship to classification. Now, we present our algorithm proposed in Section 2 in details. We provide the estimation error bound and address the empirical implementation of the algorithm.

4.1 *Estimation error and sample complexity*

Given a noisy incomplete tensor observation $\mathcal{Y}_\Omega$ from model (3), we cast the problem of estimating Θ into a series of weighted classifications. Specifically we propose the tensor estimate using the sign representation,

$$\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \operatorname{sgn} \hat{\mathcal{Z}}_\pi, \tag{7}$$

where, for each $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$, the tensor $\hat{\mathcal{Z}}_\pi \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the π-weighted classifier estimated ~~at levels $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$,~~ by

$$\hat{\mathcal{Z}}_\pi = \underset{\mathcal{Z}:\, \operatorname{rank} \mathcal{Z} \leqslant r}{\arg\min}\ L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi). \tag{8}$$

Here $L(\cdot, \cdot)$ denotes the weighted classification objective defined in (4), where we have plugged $\bar{\mathcal{Y}}_\Omega = (\mathcal{Y}_\Omega - \pi)$ in the expression, and the rank constraint follows from Proposition 3. For the theory, we assume the true r is known; in practice, r could be chosen in a data adaptive fashion via cross-validation or elbow method (Hastie et al., 2009).

The next theorem establishes the statistical convergence for the sign tensor estimate (8), which is an important ingredient for the final signal tensor estimate $\hat{\Theta}$ in (7).

THEOREM 2 (Sign tensor estimation). *Suppose $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ and $\Theta(\omega)$ is α-globally smooth under $\omega \sim \Pi$. Let $\hat{\mathcal{Z}}_\pi$ be the estimate in (8), $d_{\max} = \max_{k \in [K]} d_k$, and $d_{\max} r \lesssim |\Omega|$. Then, for all $\pi \in [-1, 1] \setminus \mathcal{N}$, with very high probability over $\mathcal{Y}_\Omega$,*

$$\mathrm{MAE}(\operatorname{sgn} \hat{\mathcal{Z}}_\pi, \operatorname{sgn}(\Theta - \pi)) \lesssim \left(\frac{d_{\max} r}{|\Omega|} \right)^{\frac{\alpha}{\alpha+2}} + \frac{1}{\rho^2(\pi, \mathcal{N})} \frac{d_{\max} r}{|\Omega|}. \tag{9}$$

Theorem 2 provides the error bound for the sign tensor estimation. Compared to the population results in Theorem 1, we explicitly reveal the dependence of accuracy on the sample complexity and the level π. The result demonstrates the polynomial decay of sign errors with $|\Omega|$. Our sign estimate achieves consistent recovery using as few as $\tilde{O}(d_{\max}r)$ noisy entries.

Recall that $\mathcal{N}$ collects the levels for which the sign tensor is possibly nonrecoverable. Let $|\mathcal{N}|$ be the ~~number of centers in~~ covering number of $\mathcal{N}$ ~~defined as $|\mathcal{N}| = \lceil \mu(\mathcal{N})/2\Delta s \rceil$,~~ with Δs-bin's, i.e, $|\mathcal{N}| = \lceil \mu(\mathcal{N})/\Delta s \rceil$, where μ is the Lebesgue measure. ~~$|\mathcal{N}|$ implies the possible levels in $\mathcal{H}$ that cannot recover the sign signal.~~ Combining the sign representability of the signal tensor and the sign estimation accuracy, we obtain the main results on our nonparametric tensor estimation method.

THEOREM 3 (Tensor estimation error). *Consider the same conditions of Theorem 2. Let $\hat{\Theta}$ be the estimate in* (7) *and $t_d = d_{\max}r/|\Omega|$. With very high probability over $\mathcal{Y}_\Omega$,*

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim \left(d_{\max} r |\Omega| t_d \right)^{\frac{\alpha}{\alpha+2}\alpha/(\alpha+2)} + \frac{1+|\mathcal{N}|\vee 1}{H} + H d_{\max} r |\Omega| H t_d. \tag{10}$$

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

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim \max\left(d_{\max} r |\Omega|^{\frac{\alpha}{\alpha+2}\wedge\frac{1}{2}} t_d^{2\alpha/(\alpha+2)},\ t_d(1+|\mathcal{N}|d_{\max}r|\Omega|) \right)^{\frac{1}{2}1/2}. \tag{11}$$

Theorem 3 demonstrates the convergence rate of our tensor estimation. The bound (10) reveals three sources of errors: the estimation error for sign tensors, the bias from sign series representations, and the variance thereof. The resolution parameter H controls the bias-variance tradeoff. We remark that the signal estimation error (11) is generally no better than the corresponding sign error (9). This is to be expected, since magnitude estimation is a harder problem than sign estimation.

In the special case of full observation with equal dimension $d_k = d, k \in [K]$, our signal estimate achieves convergence

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim \max\left(\left(r d^{K-1} r d^{-(K-1)} \right)^{\frac{\alpha}{\alpha+2}\vee\frac{1}{2}2\alpha/(\alpha+2)},\ r d^{-(K-1)}(1+|\mathcal{N}|r d^{K-1}) \right)^{\frac{1}{2}1/2}.$$

Compared to earlier methods, our estimation accuracy applies to both low- and high-rank

signal tensors. The rate depends on the sign complexity $\Theta \in \mathscr{P}_{\text{sgn}}(r)$, and this r is often much smaller than the usual tensor rank (see Section 3.1). Our result also reveals that the convergence becomes favorable as the order of data tensor increases.

We apply our method to the main examples in Section 3.1, and compare the results with existing literature. The numerical comparison is provided in Section 5.

EXAMPLE 2 (TBMs). Consider a tensor block model with r multiway blocks. Our result implies a rate $\mathcal{O}(d^{-(K-1)/2})$ by taking $\alpha = \infty$ and $|\mathcal{N}| \leqslant r^K$. This rate agrees with the previous root-mean-square error (RMSE) for block tensor estimation (Wang and Zeng, 2019).

EXAMPLE 3 (~~GLM~~GLMs). Consider a GLM tensor $\Theta = g(\mathcal{Z})$, where g is a known link function and $\mathcal{Z}$ is a latent low-rank tensor. Suppose the marginal density of $\Theta(\omega)$ is uniformly bounded as $d \to \infty$. Applying our results with $\alpha = 1$ and $|\mathcal{N}| < C$ for some constant $C > 0$ yields $\mathcal{O}(d^{-(K-1)/3})$. This rate is slightly slower than the parametric RMSE rate (Zhang and Xia, 2018; Wang and Li, 2020). One possible reason is that our estimate remains valid for unknown g and general high-rank tensors. The nonparametric rate is the price one has to pay for not knowing the form $\Theta = g(\mathcal{Z})$ as a priori.

EXAMPLE 4 (SIMs). The earlier example has shown the nonparametric rate $\mathcal{O}(d^{-(K-1)/3})$ when applying our method to single index tensor model. In the matrix case with $K = 2$, our result yields a nonparametric rate $\mathcal{O}(d^{-1/3})$, which is faster compared to the RMSE rate $\mathcal{O}(d^{-1/4})$ obtained by Ganti et al. (2015).

EXAMPLE 5 (Min/Max hypergraphon). We consider a more general model than that in Section 1. Consider a r-sign representable tensor $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ with at most d distinct entries with repetition pattern not necessarily $\mathcal{Z}_{\max}$. Applying our results with $\alpha = \infty$ and $|\mathcal{N}| = d$ yields the rate $\mathcal{O}(d^{-(K-2)/2})$. Intuitively, the rate roughly reflects the total degree of freedom

d^2, where the factor d corresponds to the number of distinct entries, and the other factor $d \approx \log(d^K)^d$ corresponds to combinatorics complexity for repetition patterns.

The following sample complexity for nonparamtric tensor completion is a direct consequence of Theorem 3.

COROLLARY 1 (Sample complexity for nonparametric completion). *Under the same conditions of Theorem 3 with $\alpha \neq 0$ and bounded $|\mathcal{N}|$, with high probability over $\mathcal{Y}_\Omega$,*

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

Our result improves earlier work (Yuan and Zhang, 2016; Ghadermarzy et al., 2019; Lee and Wang, 2020) by allowing both low- and high-rank signals. Interestingly, the sample requirements depend only on the sign complexity r but not the nonparametric complexity α. Note that $\tilde{\mathcal{O}}(d_{\max} r)$ roughly matches the degree of freedom of sign tensors, suggesting the optimality of our sample requirements.

4.2 *Numerical implementation*

This section addresses the practical implementation of our estimation (7) illustrated in Figure 3. Our sign representation of the signal estimate $\hat{\Theta}$ is an average of $2H+1$ sign tensors, which can be solved in a divide-and-conquer fashion. Briefly, we estimate the sign tensors $\mathcal{Z}_\pi$ (detailed in the next paragraph) for the series $\pi \in \mathcal{H}$ through parallel implementation, and then we aggregate the results to yield the output. The estimate enjoys low computational cost similar to a single sign tensor estimation.

For the sign tensor estimation (8), the problem reduces to binary tensor decomposition with a weighted classification loss. A number of algorithms have been developed for this problem (Ghadermarzy et al., 2018; Wang and Li, 2020; Hong et al., 2020). We adopt similar ideas by tailoring the algorithms to our contexts. Following the common practice in classification, we replace the binary loss $\ell(z, y) = |\mathrm{sgn} z - \mathrm{sgn} y|$ with a surrogate loss

$F(m)$ using a continuous function of margin $m := z\text{sgn}(y)$. Examples of large-margin loss are hinge loss $F(m) = (1-m)_+$, logistic loss $F(m) = \log(1+e^{-m})$, and nonconvex ψ-loss $F(m) = 2\min(1, (1-m)_+)$ with $m_+ = \max(m, 0)$. Similar estimation properties hold under Fisher consistency of surrogate loss (Bartlett et al., 2006) and technical lemmas.

Algorithm 1 Nonparametric tensor completion

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

1: **for** $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$ **do**
2: 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 $k \in [K]$.
3: **while** not convergence **do**
4: **for** $k = 1, \ldots, K$ **do**
5: 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.
6: **end for**
7: **end while**
8: Return $\mathcal{Z}_\pi \leftarrow \sum_{s \in [r]} \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}$.
9: **end for**

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

Figure 4: The full procedure of ~~Algorithm 1~~algorithm for nonparametric tensor completion.

The rank constraints in the optimization (8) have been extensively studied in literature. Recent developments involve convex norm relaxation (Ghadermarzy et al., 2018) and nonconvex optimization (Wang and Li, 2020; Han et al., 2020). Unlike matrices, computing the tensor convex norm is NP hard, so we choose (non-convex) alternating optimization due to its numerical efficiency. Briefly, we use the rank decomposition (2) of $\mathcal{Z} = \mathcal{Z}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K)$ to optimize the unknown factor matrices $\boldsymbol{A}_k = [\boldsymbol{a}_1^{(k)}, \ldots, \boldsymbol{a}_r^{(k)}] \in \mathbb{R}^{d_k \times r}$, where we choose to collect tensor singular values into $\boldsymbol{A}_K$. We numerically solve (7) by optimizing one factor $\boldsymbol{A}_k$ at a time while holding others fixed. Each suboptimization reduces to a convex problem with a low-dimensional decision variable. Following common practice in tensor optimization (Anandkumar et al., 2014; Hong et al., 2020), we implement multiple initializations to locate a final estimate with the lowest objective value. Figure 4 describes the full procedure.

5. Simulations

In this section, we compare our nonparametric tensor method (**NonParaT**) with two alternative approaches: low-rank tensor CP decomposition (**CPT**), and the matrix version of our method applied to tensor unfolding (**NonParaM**). We assess the performance under

both complete and incomplete observations. The signal tensors are generated based on four models listed in Table 1. The simulation covers a wide range of complexity, including block tensors, transformed low rank tensors, min/max hypergraphon with logarithm and exponential functions. We consider order-3 tensors of equal dimension $d_1 = d_2 = d_3 = d$, and set $d \in \{15, 20, \ldots, 55, 60\}$, $r = 2$, $H = 10 + (d - 15)/5$ in Algorithm 1. For **NonParaM**, we apply Algorithm 1 to each of the three unfolded matrices and report the average error. All summary statistics are averaged across 30 replicates.

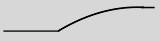
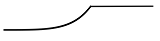
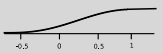
Simulation	Signal Tensor Θ	Rank	Sign Rank	Global α	CDF	Noise
1	$\mathcal{C} \times \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \boldsymbol{M}_3$	3^3	$\leq 3^3$	∞		Uniform $[-0.3, 0.3]$
2	$\lvert \boldsymbol{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \boldsymbol{a} \otimes \mathbf{1}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

Table 1: Simulation models used for comparison. We use $\boldsymbol{M}_k \in \{0,1\}^{d\times 3}$ to denote membership matrices, $\mathcal{C} \in \mathbb{R}^{3\times 3\times 3}$ the block means, $\boldsymbol{a} = \frac{1}{d}(1, 2, \ldots, d)^T \in \mathbb{R}^d$, $\mathcal{Z}_{\max}$ and $\mathcal{Z}_{\min}$ are order-3 tensors with entries ~~$\frac{1}{d}\max(i,j,k)$~~ $\frac{1}{d}\max(i,j,k)/d$ and $\frac{1}{d}\min(i,j,k)$, respectively.

Figure 5 compares the estimation error under full observation. The MAE decreases with tensor dimension for all three methods. We find that our method **NonParaT** achieves the best performance in all scenarios, whereas the second best method is **CPT** for models 1-2, and **NonParaM** for models 3-4. One possible reason is that models 1-2 have controlled multilinear tensor rank, which makes tensor methods **NonParaT** and **CPT** more accurate than matrix methods. For models 3-4, the rank exceeds the tensor dimension. Therefore, the two nonparametric methods exhibit the greater advantage for signal recovery.

Figure 6 shows the completion error against observation fraction. We fix $d = 40$ and gradually increase the observation fraction $\frac{|\Omega|}{d^3}$ from 0.3 to 1. We find that **NonParaT** achieves the lowest error among all methods. Our simulation covers a reasonable range of complexities; for example, model 1 has 3^3 jumps in the CDF of signal Θ, and models 2 and 4 have unbounded noise. Nevertheless, our method shows good performance in spite of model

misspecification. This robustness is appealing in practice because the structure of underlying signal tensor is often unknown.

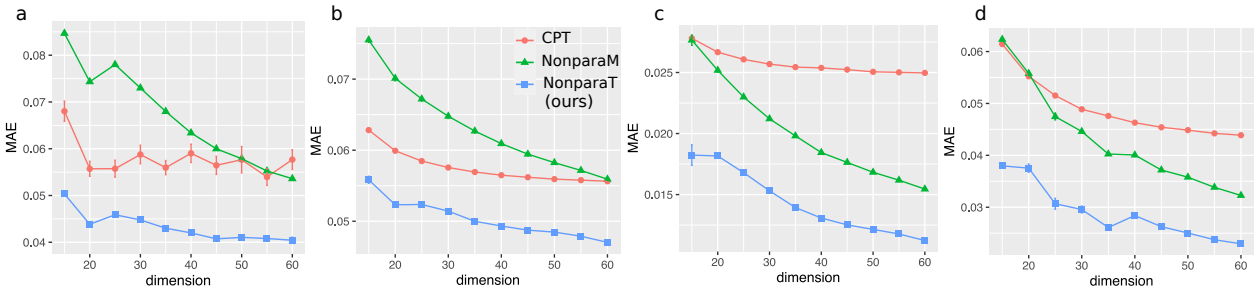


Figure 5: Estimation error versus tensor dimension. Panels (a)-(d) correspond to simulation models 1-4 in Table 1.

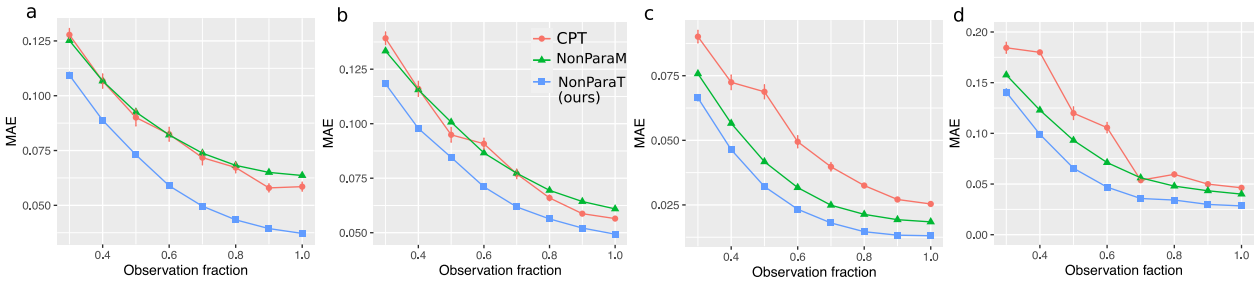


Figure 6: Completion error versus observation fraction. Panels (a)-(d) correspond to simulation models 1-4 in Table 1.

6. Data applications

We apply our method to two tensor datasets, the MRN-114 human brain connectivity data (Wang et al., 2017), and NIPS data ~~Globerson et al. (2007)~~(Globerson et al., 2007) . The brain dataset records the structural connectivity among 68 brain regions for 114 individuals along with their Intelligence Quotient (IQ) scores. We organize the connectivity data into an order-3 tensor, where entries encode the presence or absence of fiber connections between brain regions across individuals. The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003. We focus on the top 100 authors, 200 most frequent words, and normalize each word count by log transformation with pseudo-count 1. The resulting dataset is an order-3 tensor with entry representing the log counts of words by authors across years.

Table 2 compares the prediction accuracy of different methods. Reported MAEs are averaged

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 2: MAE comparison in the brain data and NIPS data analysis. Reported MAEs are averaged over five runs of cross-validation, with 20% entries for testing and 80% for training, with standard errors in parentheses. Bold numbers indicate the minimal MAE among three methods. For low-rank CPT, we use R function `rTensor` with default hyperparameters, and for our method, we set $H = 20$.

over five runs of cross-validation, with 20% entries for testing and 80% for training. Our method substantially outperforms the low-rank CP method for every configuration under consideration. Further increment of rank appears to have little effect on the performance, and we find that increased missingness gives more advantages to our method~~(see details in Appendix)~~. The comparison highlights the advantage of our method in achieving accuracy while maintaining low complexity.

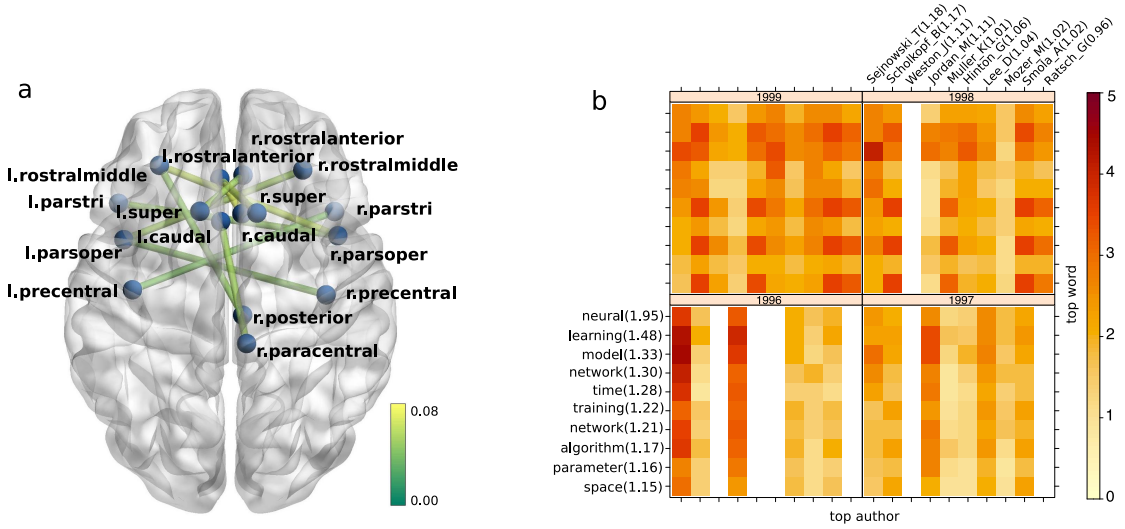


Figure 7: Estimated signal tensors in the data analysis. (a) top edges associated with IQ scores in the brain connectivity data. The color indicates the estimated IQ effect size. (b) top authors and words for years 1996-1999 in the NIPS data. Authors and words are ranked by marginal averages based on $\hat{\Theta}$, where the marginal average is denoted in the parentheses.

We next examine the estimated signal tensor $\hat{\Theta}$ from our method. Figure 7a shows the results from brain connectivity dataset. We plot the top 10 brain edges based on regression analysis of denoised connection strengths against normalized IQ scores. We find that top connections are mostly inter-hemisphere edges, consistent with recent research on brain connectivity with intelligence ~~Li et al. (2009); Wang et al. (2017)~~(Li et al., 2009; Wang et al., 2017). Figure 7b illustrates the results from NIPS data, where we plot the entries in $\hat{\Theta}$ corresponding to top authors and most-frequent words (after excluding generic words such as *figure*, *results*, etc). The identified pattern is consistent with the active topics in the NIPS publication. Among the top words are *neural* (marginal mean = 1.95), *learning* (1.48), and *network* (1.21), whereas top authors are *T. Sejnowski* (1.18), *B. Scholkopf* (1.17), *M. Jordan* (1.11), and *G. Hinton* (1.06). We also find strong heterogeneity among word occurrences across authors and years. For example, *training* and *algorithm* are popular words for *B. Scholkopf* and *A. Smola* in 1998-1999, whereas *model* occurs more often in *M. Jordan* and in 1996. The detected pattern and achieved accuracy demonstrate the applicability of our method.

7. Conclusion

We have developed a tensor completion method that addresses both low- and high-rankness based on sign series representation. Our work provides a nonparametric framework for tensor estimation, and we obtain results previously impossible. We hope the work opens up new inquiry that allows more researchers to contribute to this field.

References

Alon, N., S. Moran, and A. Yehudayoff (2016). Sign rank versus VC dimension. In *Conference on Learning Theory*, pp. 47–80.

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

Anandkumar, A., R. Ge, and M. Janzamin (2017). Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research 18*(1), 752–791.

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

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association 101*(473), 138–156.

Cai, C., G. Li, H. V. Poor, and Y. Chen (2019). Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pp. 1863–1874.

Chan, S. and E. Airoldi (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pp. 208–216.

Chi, E. C., B. J. Gaines, W. W. Sun, H. Zhou, and J. Yang (2020). Provable convex co-clustering of tensors. *Journal of Machine Learning Research 21*(214), 1–58.

Cohn, H. and C. Umans (2013). Fast matrix multiplication using coherent configurations. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1074–1087.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications 21*(4), 1253–1278.

De Wolf, R. (2003). Nondeterministic quantum query and communication complexities. *SIAM Journal on Computing 32*(3), 681–699.

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

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

Ganti, R. S., L. Balzano, and R. Willett (2015). Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pp. 1873–1881.

Ghadermarzy, N., Y. Plan, and O. Yilmaz (2018). Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing 67*(1), 29–40.

Ghadermarzy, N., Y. Plan, and Ö. Yilmaz (2019). Near-optimal sample complexity for convex tensor completion. *Information and Inference: A Journal of the IMA 8*(3), 577–619.

Globerson, A., G. Chechik, F. Pereira, and N. Tishby (2007). Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research 8*, 2265–2295.

Han, R., R. Willett, and A. Zhang (2020). An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hillar, C. J. and L.-H. Lim (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM) 60*(6), 45.

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.

Hong, D., T. G. Kolda, and J. A. Duersch (2020). Generalized canonical polyadic tensor decomposition. *SIAM Review 62*(1), 133–163.

Hore, V., A. Viñuela, A. Buil, J. Knight, M. I. McCarthy, K. Small, and J. Marchini (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics 48*(9), 1094.

Jain, P. and S. Oh (2014). Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, Volume 27, pp. 1431–1439.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Review 51*(3), 455–500.

Lee, C. and M. Wang (2020). Tensor denoising and completion based on ordinal observations. In *International Conference on Machine Learning*, pp. 5778–5788.

Li, Y., Y. Liu, J. Li, W. Qin, K. Li, C. Yu, and T. Jiang (2009). Brain anatomical network and intelligence. *PLoS Comput Biol 5*(5), e1000395.

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

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

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society 56*(4), 931–954.

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

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

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

Xu, J. (2018). Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, pp. 5433–5442.

Yuan, M. and C.-H. Zhang (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics 16*(4), 1031–1068.

Zhang, A. and D. Xia (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory 64*(11), 7311 – 7338.