Beyond the Signs: Nonparametric Tensor Completion via Sign Series

Anonymous Authors[1]

Abstract

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 a real-valued signal tensor using a series of sign tensors with low sign ranks. Unlike earlier methods, the sign series representation allows high-rankness in the original signal, and the model also encompasses many existing low-rank tensor methods—including CP models, Tucker models, single index models, certain hypergraphon models—as special cases. We show that the sign tensor series are theoretically characterized, and computationally estimable, by classification tasks with carefully-specified weights. The excess risk rate, estimation error bound, and sample complexity are established. The results uncover the joint contribution of statistical bias-variance errors and discretization errors. Numerical results demonstrate the robustness of our proposal over previous tensor methods.

1. Introduction

2. Preliminaries

We use $\mathrm{sgn}(\cdot)\colon \mathbb{R} \to \{-1, 1\}$ to denote the sign function, where $\mathrm{sgn}(y) = 1$ if $y \geq 0$ and -1 otherwise. We allow univariate functions, such as $\mathrm{sgn}(\cdot)$ or general $f\colon \mathbb{R} \to \mathbb{R}$, to be applied to tensors in an element-wise manner. We use the shorthand $[n]$ to denote the n-set $\{1, \ldots, n\}$ for $n \in \mathbb{N}_+$. We use $\otimes$ to denote the outer product of vectors, $\|\boldsymbol{x}\|_2$ to denote the vector 2-norm, and $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}\colon \|\boldsymbol{x}\|_2 = 1\}$ to denote the $(d-1)$-dimensional unit sphere.

We use $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ to denote an order-K $(d_1, \ldots, d_k)$-dimensional tensor, and use $\mathcal{Y}(\omega) \in \mathbb{R}$ to denote the tensor entry indexed by $\omega \in [d_1] \times \cdots \times [d_K]$. The Frobenius norm of $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_F = \sqrt{\sum_\omega \mathcal{Y}^2(\omega)}$. Unlike matrices, various notions of decomposition have been developed for tensors of order $K \geq 3$. The Canonical Polyadic (CP) tensor decomposition (Hitchcock, 1927) for a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is defined as

$$\Theta = \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)}, \tag{1}$$

where $\lambda_1 \geq \cdots \geq \lambda_R > 0$ are called tensor singular values, and $\boldsymbol{a}_r^{(k)} \in \mathbf{S}^{d_k - 1}$ are called tensor singular vectors, for all $r \in [R]$, $k \in [K]$. The minimal R for which the decomposition (1) holds is called the tensor rank, denoted as $\mathrm{rank}(\Theta)$.

3. Motivation and method overview

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

$$\mathcal{Y} = \Theta + \mathcal{E}, \tag{2}$$

where $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the unknown signal tensor of interest, and $\mathcal{E}$ is a noise tensor consisting of mean-zero, independent 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; this incorporates, for example, a Bernoulli tensor whose entries $\mathcal{Y}(\omega)$ have mean $\Theta(\omega)$ and variance $\mathrm{Var}(\mathcal{E}(\omega)) = \Theta(\omega)(1 - \Theta(\omega))$. In general, assume the entries of the tensor $\mathcal{Y}$ take values in a bounded interval $[-L, L]$. Without loss of generality, we assume $L = 1$ throughout the paper.

Our observation is an incomplete data tensor from (2), denoted $\mathcal{Y}_\Omega$, where $\Omega \subset [d_1] \times \cdots \times [d_K]$ denotes the index set of observed entries. We consider an observation model of Ω that allows both uniform and non-uniform sampling. Specifically, let $\Pi = \{\pi_\omega\}$ be an arbitrarily predefined probability distribution over the full index set such that $\sum_{\omega \in [d_1 \times \cdots \times d_K]} \pi_\omega = 1$. We assume the entries ω in Ω are i.i.d. draws with replacement from the full index set using distribution Π. The sampling rule will be denoted as $\omega \sim \Pi$.

Our goal is to accurately estimate Θ from the incomplete, noisy observation $\mathcal{Y}_\Omega$. In particular, we focus on the following two problems:

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

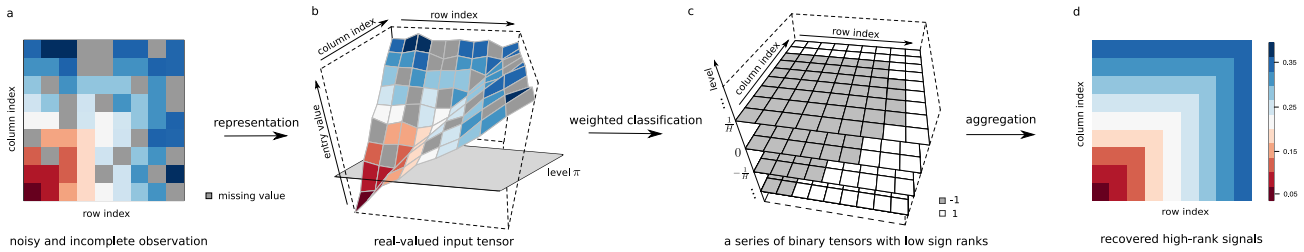

Figure 1. Illustration of our proposed method. For visualization purpose, we plot an order-2 tensor (a.k.a. matrix) in the figure; similar procedure applies to higher-order tensors. (a): input tensor $\mathcal{Y}_\Omega$ with noisy and incomplete entries. (b) and (c): the algorithm uses weighted classification to estimate sign tensors $\mathrm{sgn}(\Theta - \pi)$ for a sequence of levels $\pi \in \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$. (d) output tensor $\hat{\Theta}$ with denoised and imputed entries. The depicted example is based on Example 5, where the true signal matrix has full rank.

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

3.1. Inadequacies of common low-rank models

The signal plus noise model (2) is common in tensor literature. Most existing methods perform estimation based on the low-rankness of Θ (Anandkumar et al., 2014; Montanari and Sun, 2018; Kadmon and Ganguli, 2018; Cai et al., 2019). While these methods have shown great success in low-rank recovery, little is explored when the underlying signal tensor is of high rank. Here we provide two examples to illustrate the limitation of classical low-rank models.

The first example reveals the sensitivity of tensor rank to order-preserving transformations. Let $\mathcal{Z} \in \mathbb{R}^{d\times d\times d}$ be an order-3 tensor with $\mathrm{rank}(\mathcal{Z}) = 3$. Suppose a monotonic transformation $f(z) = (1 + \exp(-cz))^{-1}$ is applied to $\mathcal{Z}$ entrywise, and we observe data from model (2) with the signal tensor $\Theta = f(\mathcal{Z})$. Figure 2a plots the numerical rank of Θ versus c when the tensor dimension $d = 30$. As we can see, the rank increases rapidly with c, rending traditional low-rank tensor methods ineffective in the presence of order-preserving nonlinearities. In applications such as digital processing and genomics analysis, the tensor of interest often undergoes some unknown transformation prior to measurements. The sensitivity to transformation therefore makes the low-rank model less desirable in practice.

The second example shows that the classical low-rankness may exclude interesting tensor structures. Here we consider the signal tensor of the form $\Theta = \log(1 + \mathcal{Z})$, where $\mathcal{Z}$ is an order-3 tensor with entries $\mathcal{Z}(i, j, k) = \frac{1}{d}\max(i, j, k)$ for $(i, j, k) \in [d]^3$. In this example neither Θ nor $\mathcal{Z}$ is low rank; in fact, we are able to show that both tensor ranks are lower bounded by dimension d (see proofs in Appendix and Figure 1b). The matrix analogy of this structured Θ was first studied in Chan and Airoldi (2014) in the context of

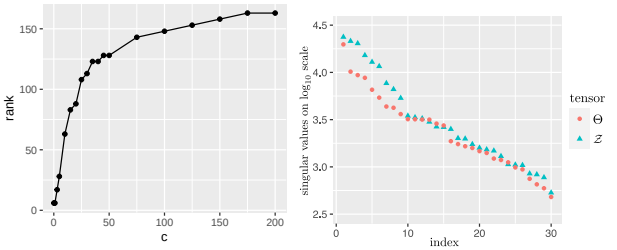

Figure 2. (a) Numerical rank of $\Theta = f(\mathcal{Z})$ versus c in Example 1. Here $\mathcal{Z}$ is generated using a rank-3 tensor whose (unnormalized) eigenvectors consist of i.i.d. entries from $N(0, 1)$. The numerical rank is computed as the minimal R such that $\min_{\mathrm{rank}(\mathcal{A})\le R} \frac{\|\Theta-\mathcal{A}\|_F}{\|\Theta\|_F} \le 0.1$. (b). Top $d = 30$ numerical singular values of tensors Θ and $\mathcal{Z}$ in Example 2. The reported values in both figures are obtained by running CP decomposition ten times using random initializations.

graphon analysis. However, classical tensor models fail to incoporate these types of structures.

In the above and many other examples, the signal tensor of interest is structured but of high rank.

3.2. Overview of our proposal

Before describing the main results, we outline the main crux of our idea. In the earlier two examples, the tensor Θ is of high rank. However, the runcated $(\Theta - \pi)$ has the same sign pattern as low-rank tensors. This observation suggests a a more general framework that allows flexibly modeling..... Indeed as we see in Section **??**, the model not only corporates common low-rank models such as CP/Tucker and single index model, but also allows substantial class of structured tensors.

Figure 1 illustrates the main approach of our method. We convert the data tensor into a tensor series $\bar{\mathcal{Y}}_\pi = (\mathcal{Y} - \pi)$ for $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$ and perform

classification to estimate the sign signal of each,

$$\hat{\mathcal{Z}}_\pi = \underset{\mathcal{Z} \in \text{classifier model}}{\arg\min} \text{Risk}(\mathcal{Z}, \text{sgn}\bar{\mathcal{Y}}_\pi).$$

Then, the nonparametric estimate takes the form of

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

Here $\text{Risk}_\pi(\cdot, \cdot)$ is a weighted extension of usual classification risk, and the classifier models refer to a tensor model which will be described in next section.

The main approach is to estimate $\hat{\Theta}$ from the binary tensor series $\{\text{sgn}(\mathcal{Y} - \pi) \colon \pi \in \mathcal{H}\}$. 1) exactly recover the signal tensor under a wide range of currently used tensor models; 2) brings the benefit of high-rank tensor estimation and completion; 3) leverage efficient classification algorithms. We choose the classifier model that is broader that incorporates low rank tensor model, single index tensor model, and nonlinear tensor model, etc, in the literature. It may be counterintuitive that dichotomization may lose information in the data. It turns out the ensemble model based on dichotomizations not only preserves all information in the data, but also brings the nonparametric benefit of model-free and flexibility over classical parameter models.

4. Sign representable tensors

In this section, we introduce the sign representable tensors and the properties of sign tensor series.

4.1. Sign rank of tensors

Let Θ be a real-valued tensor, and $\text{sgn}(\Theta)$ be the induced sign tensor. The sign pattern induces an equivalence relationship between tensors. Two tensors are called sign equivalent, denoted $\simeq$, if they share the same sign pattern.

Definition 1 (Sign rank)**.** The sign rank of a tensor Θ is the minimal rank among all tensors that share the same sign pattern as Θ,

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

The sign rank is also called *support rank* (Cohn and Umans, 2013), *minimal rank*, *nondeterministic rank* (De Wolf, 2003), in the filed of combinatorics and computational complexity. Some literature defines sign rank for binary tensors/matrices only; we extend the notion to real-valued tensors. Note that the sign rank concerns only the sign pattern of Θ but discards the magnitudes information. In particular, $\text{srank}(\Theta) = \text{srank}(\text{sgn}\Theta)$.

Determining the sign rank is NP hard, just as the determining tensor rank is NP hard. Fortunately, many tensors arising in practice have low sign rank. We show here the family of low rank sign tensors is strictly larger than usual low rank tensors. By definition, the sign rank is upper bounded by the usual rank. More generally,

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

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

Conversely, we show that the sign rank can be dramatically smaller than tensor rank.

Proposition 1 (Strictness)**.** *For every order K and tensor dimension d, there exists a tensor whose usual rank is d but the sign rank is 2.*

The Proposition 1 is shown by construction, and we provide the proof in Appendix. Corollary 1 and Proposition 1 demonstrates that low sign rank tensors constitute a strictly broader family than the usual low rank tensors. Some examples include identity tensors, banded tensor, ... etc. ...

We now introduce the family of tensor model (2) of our interest, which we coin as "sign representable tensors".

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. The binary tensor $\text{sgn}(\Theta - \pi)$ is called the π-level sign tensor. If a tensor Θ is (r, π)-sign representable for all $\pi \in [-1, 1]$ except for a finite number of points, then we say that Θ is r-globally sign representable, denoted $\Theta \in \mathscr{P}_{\text{sgn}}(r) = \{\Theta \colon \text{srank}(\Theta - \pi) \le r \text{ for all } \pi \in [-1, 1]\}$.

We show that the r-globally sign representable tensor is very broad family that incorporates many usual common tensor models, such as single index model, GLM models, and certain hypergraphon models.

emphasize the original tensor may be high rank....Give a toy example.

Example 1 (CP/Tucker low-rank tensor model)**.** Let Θ be a tensor with CP rank bounded by r. Then the usual signal + noise model is a special cases of low sign rank tensor assemble model.

Example 2 (Generalized linear model (GLM) for tensors)**.** Let $\mathcal{Y}$ be a binary tensor. Then the GLM low rank model is a special case of low sign rank tensor assemble model.

Example 3 (Single index model)**.** Suppose there exists a (unknown) monotonic function $g \colon \mathbb{R} \to \mathbb{R}$ such that $g(\Theta)$ has rank bounded by r. Then, Θ belongs to the a low sign rank assemble tensor; i.e., $\Theta \in \mathscr{P}_{\text{sgn}}(r + 1)$.

Example 4 (Tensor block model)**.** Finite possible π. Suppose Θ has at most r multiway blocks. Then $\Theta \in \mathscr{P}_{\text{sgn}}(r)$.

Example 5 (Min/Max hypergraphon)**.** Suppose the matrix $\Theta = [\![\Theta(i,j,k)]\!]$ are generated from the graphon model,

$$\Theta(i,j,k) = \exp(1 + 0.5\max(x_i, y_j, z_k)),$$

where $x_i, y_i, z_k \sim$ Uniform$[0,1]$ i.i.d. for all $(i,j,k) \in [d_1] \times [d_2] \times [d_3]$. Then the affine matrix $(\Theta - \pi)$ has sign rank bounded by 2; i.e., $\Theta \in \mathscr{P}_{\text{sgn}}(2)$ for every tensor dimension. In contrast, the regular matrix rank of Θ has full rank almost surely for every matrix dimension d. This example highlights the benefit of our assemble model.

More generally, if $\Theta(i,j,k) = g(\max(x_i, y_i, z_i))$ where $g(\cdot)$ is a continuous function with at most r real roots in the equation $g(z) = \pi$ (for example, this holds for all $\pi \in [-1,1]$ when g is an rth polynomial.) Then $\Theta \in \mathscr{P}_{\text{sgn}}(r+1)$. Same properties hold if the maximum is replaced by minimum.

4.2. Characterization of sign tensors via weighted classification

Accurate estimation of sign representable tensors crucially depends on the behavior of sign tensors, $\text{sgn}(\Theta - \pi)$. In this section, we show that weighted classification completely characterizes the sign tensors. The results bridge the algebraic and statistical properties of sign representable tensors, and provide a building block for our nonparametric algorithm (see Figure 1).

For a given $\pi \in [-1,1]$, define a new data tensor $\bar{\mathcal{Y}} = (\mathcal{Y} - \pi)$. We introduce a weighted classification loss function

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

where $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the decision variable to be estimated. In the case of binary tensor, taking $\pi = 1/2$ gives unweighted usual classification risk.

Denote the weighted classification risk

$$\text{Risk}(\mathcal{Z}) = \mathbb{E}_{\mathcal{Y},\Omega} L(\mathcal{Z}, \bar{\mathcal{Y}}), \tag{3}$$

where the expectation is taken with respect to both $\mathcal{Y}$ under the sign assemble model, and Ω under the sampling distribution. The form of Risk$(\cdot)$ implicitly depends on π.

Proposition 2 (Global optimum of weighted risk)**.** *Suppose the data $\mathcal{Y}$ is generated from model 2 with $\Theta \in \mathscr{P}_{\text{sgn}}(r)$. Then, for all $\bar{\Theta}$ that are sign equivalent to* $\text{sgn}(\Theta - \pi)$,

$$\begin{aligned}\text{Risk}(\bar{\Theta}) &= \min\{\text{Risk}(\mathcal{Z})\colon \mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}\},\\ &= \min\{\text{Risk}(\mathcal{Z})\colon \text{rank}\mathcal{Z} \le r\}.\end{aligned}$$

The support sign tensor $\text{sgn}(\Theta - \pi)$ serves the bridge between the target parameter Θ and statistical properties of Risk$(\cdot)$. Highlight the importance of weight!

The above property shows that the $\text{sgn}(\Theta - \pi)$ is one of the optimum of Risk$(\cdot)$. However, the converse may not true. We establish the uniqueness; that is, is the tensor $\text{sgn}(\Theta - \pi)$ the unique (up to sign equivalence) optimizer of Risk$(\cdot)$? The following assumption quantifies the identifiability of $\bar{\Theta}$ from Risk$(\cdot)$.

We introduce some additional notation. Recall that $\omega \in \Pi$ denotes the sampling distribution over index set. Our theory concerns the high-dimensional ... Both Π and Θ implicitly depend on the tensor dimension. We are interested in the high dimensional region $d := \min_k d_k \to \infty$ for the sequence $\Pi = \Pi_d$ and $\Theta = \Theta_d$. Unless otherwise stated, all relevant assumptions should be interpreted as uniform conditions that hold for all d. Let $\mathcal{N} = \{\pi\colon \mathbb{P}_{\omega\sim\Pi}(\Theta(\omega) = \pi) \neq 0\}$ denote the set of mass points of $\Theta(\omega)$. Assume there exists a constant $c_1 > 0$, independent of tensor dimension, such that $|\mathcal{N}| \le c_1$. Furthermore,

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

$$\sup_{0 \le t < \rho(\pi,\mathcal{N})} \frac{\mathbb{P}_{\omega\sim\Pi}(\omega\colon |\Theta(\omega) - \pi| \le t)}{t^\alpha} \le c, \tag{4}$$

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 $\alpha = \alpha(\pi)$ is called the smoothness index at level π. We make the convention that $\alpha = \infty$ if the set $\{\omega\colon |\Theta(\omega) - \pi| \le t\}$ is measure-zero set, implying the density of $\Theta(\omega)$ has a jump around the level π.

If Θ is α-smooth for all $\pi \in [-1,1]$ except for a finite number of points, then we call Θ is α-globally smooth. The smoothness index is determined by the marginal density of $\Theta(\omega)$, or equivalently, by the underlying Θ and the sampling distribution $\omega \sim \Pi$.

To gain some intuition about (4), we consider the case of uniform sampling model where each tensor entry has equal probability to be observed. Larger value of α corresponds to an easier problem (faster statistical rate). The case $\alpha < 1$ may indicate a local extremum or inflection point of the density of $\Theta(\omega)$ at the level π. Typical cases are $\alpha = 1$, for example, when the density of $\Theta(\omega)$ is locally bilipschitz.

For tensor block models, $|\mathcal{N}|$ equals the number of blocks with distinct means which is finite. Furthermore, we can take $\alpha = \infty$ for all $\pi \notin \mathcal{N}$ (because the numerator in (4) is zero). For max/min hypergraphon model with r-th polynomial function, we have $\alpha = 1$ for all π except at most r many points (local extremums). For single index model $f(\Theta)$....

We consider the mean absolute error (MAE)

$$\text{MAE}(\Theta_1, \Theta_2) \overset{\text{def}}{=} \mathbb{E}_{\omega\sim\Pi}\|\Theta_1 - \Theta_2\|_1.$$

We now reach the main theorem in this section.

Theorem 4.1 (Identifiability)**.** *Under Assumption 1, for tensors* $\bar{\Theta} \simeq \mathrm{sgn}(\Theta - \pi)$ *and all tensors* $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$,

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

where $C(\pi) > 0$ *is a constant.*

The result immediately suggests uniqueness of the optimizer of $\mathrm{Risk}(\cdot)$ up to a zero-measure set under Π. Furthermore, it establish the stability of recovering $\mathrm{sgn}\Theta$ from the optimization (3).

5. Nonparametric tensor completion via sign series

Now, we are ready to state the assemble algorithm. We use the sample formulation ... , and propose the estimate

$$\hat{\mathcal{Z}}_\pi = \underset{\mathcal{Z}:\ \mathrm{rank}(\mathcal{Z}) \leq r,\ \|\mathcal{Z}\|_F \leq 1}{\arg\min} L(\mathcal{Z}, \bar{\mathcal{Y}}_\pi), \tag{5}$$

for a series of levels $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$, and the aggregated tensor estimate

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

Theorem 5.1 (Estimation of sign series)**.** *Suppose* $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ *and Assumption 1 holds at level* $\pi \notin \mathcal{N}$ *with smoothness index* $\alpha \in [0, 1]$*. Then, for every such* π*, with very high probability over* $\mathcal{Y}_\Omega$*,*

$$\mathrm{MAE}(\mathrm{sgn}\hat{\mathcal{Z}}_\pi, \mathrm{sgn}\bar{\Theta}_\pi) \lesssim \left(\frac{dr}{|\Omega|}\right)^{\alpha/(\alpha+2)} + \frac{1}{\rho^2(\pi, \mathcal{N})} \frac{dr}{|\Omega|}.$$

Theorem 5.1 shows that the estimate from (5) recovers the desired sign tensor $\mathrm{sgn}\bar{\Theta}$. The result suggests the sample complexity for sign tensor recovery is of the order dr. We will show that later, this complexity remains for recovering the entire tensor.

Theorem 5.2 (Tensor estimation error)**.** *With very high probability over* $\mathcal{Y}_\Omega$*,*

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \leq \left(\frac{dr}{|\Omega|}\right)^{\alpha/(\alpha+2)} + \frac{1}{H} + H\frac{dr}{|\Omega|}.$$

Setting $H \asymp \left(\frac{|\Omega|}{rd}\right)^{1/2}$ *gives*

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \leq \left(\frac{dr}{|\Omega|}\right)^{\frac{\alpha}{\alpha+2} \vee \frac{1}{2}}.$$

Corollary 2 (Sample complexity for completion)**.** With very high probability

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

Completion is consistent provided that $|\Omega|/dr \to \infty$. Furthermore, in the full observation case, $|\Omega| = \prod_k d_k$, then the

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \leq \left(\frac{r}{d^{-(K-1)}}\right)^{\frac{\alpha}{\alpha+2} \vee \frac{1}{2}}.$$

Example 6 (Tensor block model)**.** Consider tensor block model. $\alpha = \infty$. Earlier results show the estimation error

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \leq \sqrt{\frac{r \log d}{|\Omega|}}.$$

this is the best rate. reached the minimax rate??

Example 7 (Single index tensor model)**.** $\mathcal{O}(1/3)$???

6. Algorithm

The optimization (5) is a non-convex problem because of the nonconvexity in L and the rank constraints. We present an efficient ADMM algorithm with a good empirical performance.

$$\min_{\mathcal{Z}} L(\mathcal{Z}) := \frac{1}{|\Omega|} \sum_{\omega \in \Omega} |\bar{y}_\omega|(1 - z_\omega \bar{y}_\omega)_+$$
$$\text{subject to } \mathrm{rank}(\mathcal{Z}) \leq r \text{ and } \|\mathcal{Z}\|_F \leq 1.$$

$$L(\mathcal{Z}, \mathcal{W}, \Lambda, \rho, \lambda) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} |\bar{y}_\omega|(1 - z_\omega \bar{y}_\omega)_+ + \lambda\|\mathcal{Z}\|_F^2 + \rho\|\mathcal{Z} - \mathcal{W}\|_F^2 + \langle \mathcal{Z} - \mathcal{W}, \Lambda \rangle$$

Two changes: 1. make convex relaxation; 2. alternatively... Gradually increasing ρ and λ.

Proof. We verify two conditions.

1. Approximation error. For $\mathcal{Z}$ with $\text{rank}(\mathcal{Z}) \leq r$, we have $\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) = 0$ for all d.

2. Variance-to-mean relationship

$$\text{Var}_{\mathcal{Y},\Omega}[L(\mathcal{Z}, \bar{\mathcal{Y}}_\pi) - L(\bar{\Theta}, \mathcal{Y}_\pi)] \leq [\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})]^{\alpha/(1+\alpha)} + \frac{1}{\rho(\pi, \mathcal{N})}[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})].$$

Apply Lemma 1 to the above condition, we obtain

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \leq t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})} t_n, \quad \text{where } t_n = \frac{Krd}{n}.$$

□

Lemma 1. *Because the classification rate is scale-free; Risk*$(\mathcal{Z}) =$ *Risk*$(c\mathcal{Z})$ *for every* $c > 0$*. Therefore, without loss of generality, we solve the estimate subject to* $\|\mathcal{Z}\|_F \leq 1$*,*

$$\hat{\mathcal{Z}} = \underset{\mathcal{Z}\colon \text{rank}(\mathcal{Z}) \leq r, \|\mathcal{Z}\|_F \leq 1}{\arg\min} L(\mathcal{Z}, \bar{\mathcal{Y}}_\pi).$$

Write $|\Omega| = n$*. We have*

$$\mathbb{P}[\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \geq t_n] \leq \frac{7}{2} \exp(-Cnt_n).$$

The rate of convergence $t_n > 0$ *is determined by the solution to the following inequality,*

$$\frac{1}{t_n} \int_{t_n}^{\sqrt{t_n^\alpha + \rho^{-1} t_n}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon,\ \mathcal{F},\ \|\cdot\|_2)} d\varepsilon \leq n^{1/2},$$

where $\mathcal{F} = \{\mathcal{Z}\colon \text{rank}(\mathcal{Z}) \leq r,\ \|\mathcal{Z}\|_F^2 \leq 1\}$ *and* $\rho = \rho(\pi, \mathcal{N})$*. By Lemma 2, we obtain*

$$t_n \asymp \left(\frac{Kdr}{n}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho^2(\pi, \mathcal{N})} \frac{Kdr}{n}.$$

Finally, we obtain

$$\mathbb{P}[\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \geq t_n] \leq \frac{7}{2} \exp(-Cd^{\frac{\alpha+1}{\alpha+2}} n^{\frac{1}{\alpha+2}}) \leq \frac{7}{2} \exp(-C\sqrt{d}),$$

where $C = C(k, r) > 0$ *is a constant independent of* d *and* n*.*

Lemma 2 (Bracketing number for bounded low rank tensor)**.**

$$\sqrt{\mathbb{E}_{\omega \sim \Pi} |\mathcal{Z}_1(\omega) - \mathcal{Z}_2(\omega)|^2} \leq \|\mathcal{Z}_1 - \mathcal{Z}_2\|_\infty \leq \|\mathcal{Z}_1 - \mathcal{Z}_2\|_F.$$

Therefore

$$\mathcal{H}_{[\,]}(2\varepsilon, \mathcal{F}, \|\cdot\|_2) \leq \mathcal{H}(\varepsilon, \mathcal{F}, \|\cdot\|_F) \leq C(1 + Kdr) \log \frac{d}{\varepsilon},$$

where the covering number for low rank tensor is based on (Mu et al., 2014; Ibrahim et al., 2020).

References

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

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.

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. SIAM.

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

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.

Ibrahim, S., X. Fu, and X. Li (2020). On recoverability of randomly compressed tensors with low cp rank. *IEEE Signal Processing Letters 27*, 1125–1129.

Kadmon, J. and S. Ganguli (2018). Statistical mechanics of low-rank tensor decomposition. *Advances in Neural Information Processing Systems 31*, 8201–8212.

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

Mu, C., B. Huang, J. Wright, and D. Goldfarb (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pp. 73–81.