Nonparametric tensor estimation with unknown permutations

Chanwoo Lee
Department of Statistics
University of Wisconsin-Madison
chanwoo.lee@wisc.edu

Miaoyan Wang
Department of Statistics
University of Wisconsin-Madison
miaoyan.wang@wisc.edu

Abstract

We consider the problem of structured tensor denoising in the presence of unknown permutations. Such data problems arise commonly in recommendation system, neuroimaging, community detection, and multiway comparison applications. Here, we develop a general family of smooth tensors up to arbitrarily index permutations; the model incorporates the popular block models and graphon models. We show that a constrained least-squares estimate in the block-wise polynomial family achieves the minimax error bound. A phase transition phenomenon is revealed with respect to the smoothness threshold needed for optimal recovery. In particular, we find that a polynomial of degree of $m(m-1)/2$ is sufficient for accurate recovery of order-m tensors, whereas higher degree exhibits no further benefits. Furthermore, we provide an efficient polynomial-time Borda count algorithm that provably achieves optimal rate under monotonicity assumptions. The efficacy of our procedure is demonstrated through both simulations and Chicago crime date analysis.

1 Introduction

Higher-order tensor datasets are rising ubiquitously in modern data science applications, for instance, recommendation systems [3], social networks [19], genomics [23], and neuroimaging [29]. Tensor structure provides effective representation of data that classical vector- and matrix-based methods fail to capture. One example is music recommendation system that records ratings of songs from users on different contexts [3]. This three-way tensor of user×song×context allows us to investigate interaction of users and songs under a context-specific manner. Another example is network analysis that studies the connection pattern among nodes. Pairwise interactions are often insufficient to capture the complex relationships, whereas multi-way interactions improve understanding the networks in molecular system [18] and computer vision [1]. In both examples, higher-order tensors represent multi-way interactions in an efficient way.

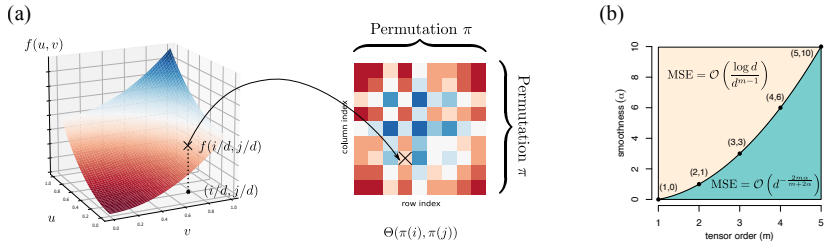

Figure 1: (a): Illustration of order-m d-dimensional permuted smooth tensor models with $m = 2$. (b): Regimes of mean square error (MSE) depending on the smoothness α and tensor order m. Bold dots show the critical smoothness α^* and tensor order m.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Tensor estimation problem cannot be solved without imposing structure. We study a class of structured tensors, *permuted smooth tensors* of the following form:

$$\mathcal{Y} = \Theta \circ \pi + \text{noise}, \quad \text{where} \quad \Theta_{i_1,\ldots,i_K} = f(i_1,\ldots,i_m), \tag{1}$$

where $\pi\colon [d] \to [d]$ is an unknown latent permutation, Θ is an unknown order-m d-dimensional signal tensor, and f is an unknown multivariate function with smoothness index $\alpha > 0$; see Figure 1(a) for an illustration. Our primary goal is to estimate a permuted smooth signal tensor from a noisy observation.

Related work and our contributions. The estimation problem of (1) falls into the general category of structured learning with *latent permutation*, which has recently observed a surge of interest. Models involving latent permutations include graphon [4, 9, 15], stochastic transitivity models [5, 21], statistical seriation [7, 12], graph matching [6, 17], and crowd labeling [22]. Most of theses methods are developed for matrices. The tensor counterparts, however, are far less well understood. Table 1 summarizes the related works on tensor learning with latent permutations.

	Pananjady et al [20]	Balasubramanian [2]	Li et al [16]	**Ours***
model structure	monotonic	Lipschitz	Lipschitz	α-smoothness
minimax lower bound	$\surd$	$\times$	$\times$	$\surd$
error rate for order-3 tensors	d^{-1}	$d^{-6/5}$	d^{-1}	d^{-2}
polynomial algorithm	$\surd$	$\times$	$\surd$	$\surd$

Table 1: Comparison of our results with previous works. *We list here only the result for infinitely smooth order-3 tensors. Our results allow general tensors of arbitrary order m and smoothness α; See Theorems 1 and 3.

The primary goal of our work is to provide statistical and computational estimation accuracy for the permuted smooth tensor model (1). We summarize our major contributions below.

(a) We develop a general permuted α-smooth tensor model for an arbitrary smoothness index $\alpha > 0$. In contrast to earlier work [2, 16] that focuses only on $\alpha = 1$, we fully establish the statistically optimal error rate and its dependence on tensor order, dimension, and smoothness index.

(b) We discover an intriguing phase transition phenomenon with respect to the smoothness threshold needed for optimal tensor recovery in model (1). The critical threshold α^* (defined in Theorem 1 and 3) characterizes two distinct error dependence behaviors on the smooth index α. We proved that the error decreases with α in the range $\alpha < \alpha^*$, whereas the error is a constant of α in the range $\alpha > \alpha^*$. Figure 1(b) plots the critical smoothness α^* as a function of tensor order m. These results are distinct from the matrix counterparts [8, 15, 9], thereby highlighting the fundamental challenges with tensors.

(c) We provide an efficient polynomial-time Borda count algorithm that provably achieves optimal rate under monotonicity assumptions. Simulation and data studies demonstrate the competitive performance of our algorithm.

Notation. We use $[d] = \{1,\ldots,d\}$ for d-set with $d \in \mathbb{N}_+$. For a set S, $|S|$ denotes its cardinality and $\mathbb{1}_S$ denotes the indicator function. For positive two sequences $\{a_n\}, \{b_n\}$, we denote $a_n \lesssim b_n$ if $\lim_{n\to\infty} a_n/b_n \le c$ for some constant $c > 0$, and $a_n \asymp b_n$ if $c_1 \le \lim_{n\to\infty} a_n/b_n \le c_2$ for some constants $c_1, c_2 > 0$. Given number $a \in \mathbb{R}$, the floor function $\lfloor a \rfloor$ is the largest integer no greater than a, and the ceiling function $\lceil a \rceil$ is the smallest integer no less than a. We use $\mathcal{O}(\cdot)$ to denote the big-O notation, $\tilde{\mathcal{O}}(\cdot)$ the variant hiding logarithmic factors. An event A is said to occur *with high probability* if $\mathbb{P}(A)$ tends to 1 as the tensor dimension $d \to \infty$. Let $\Theta \in \mathbb{R}^{d\times\cdots\times d}$ be an order-m d-dimensional tensor, and $\pi\colon [d] \to [d]$ be an index permutation. We use $\Theta_{i_1,\ldots,i_m}$ to denote the tensor entry indexed by $(i_1,\ldots,i_m)$, and use $\Theta \circ \pi$ to denote the permuted tensor such that $(\Theta \circ \pi)_{i_1,\ldots,i_m} = \Theta_{\pi(i_1),\ldots,\pi(i_m)}$ for all $(i_1,\ldots,i_m) \in [d]^m$. We use $S(d) = \{\pi\colon [d] \to [d]\}$ to denote all possible permutations on $[d]$.

2 Smooth tensor model with unknown permutation

Suppose we observe an order-m d-dimensional symmetric data tensor from the following model,

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

where $\pi\colon [d] \to [d]$ is an unknown latent permutation, $\Theta \in \mathbb{R}^{d\times\cdots\times d}$ is an unknown symmetric signal tensor under certain smoothness (to be specified in next paragraph), and $\mathcal{E}$ is a symmetric noise tensor consisting of zero-mean, independent sub-Gaussian entries with variance bounded by σ^2. For simplicity of presentation, we focus on symmetric tensors in the main paper; our models and

techniques easily generalize to non-symmetric tensors. Here, we do not assume identical distributions among entries in $\mathcal{E}$. In particular, we allow the error variance to depend on mean. Therefore, our model (1) allows a wide range of data types including Gaussian and Bernoulli tensors.

We now describe the smooth model on the signal Θ. Assume that there exists a multivariate function $f\colon [0,1]^m \to \mathbb{R}$ underlying the signal tensor, such that

$$\Theta_{i_1,\ldots,i_m} = f\left(\frac{i_1}{d},\ldots,\frac{i_m}{d}\right). \tag{3}$$

Assume the generating function f is in the α-Hölder smooth family.

Definition 1 (α-Hölder smooth)**.** A function $f\colon [0,1]^m \to \mathbb{R}$ is α-Hölder smooth, denoted as $f \in \mathcal{H}(\alpha)$, if there exists a polynomial $P_{\lfloor\alpha\rfloor}(\boldsymbol{x}-\boldsymbol{x}_0)$ of degree $\lfloor\alpha\rfloor$, such that

$$|f(\boldsymbol{x}) - P_{\lfloor\alpha\rfloor}(\boldsymbol{x}-\boldsymbol{x}_0)| \le C\|\boldsymbol{x}-\boldsymbol{x}_0\|_\infty^\alpha, \tag{4}$$

for all $\boldsymbol{x}, \boldsymbol{x}_0 \in [0,1]^m$ and a universal constant $C > 0$.

Hölder smooth function class is one of the most popular function classes considered in the nonparametric regression literature [15, 9]. In addition to the function class $\mathcal{H}(\alpha)$, we also define the smooth tensor class based on discretization (3),

$$\mathcal{P}(\alpha) = \left\{\Theta \in \mathbb{R}^{d\times\cdots\times d}\colon \Theta(\omega) = f\left(\frac{\omega}{d}\right) \text{ for all } \omega = (i_1,\ldots,i_m) \in [d]^m \text{ and } f \in \mathcal{H}(\alpha)\right\}.$$

Combining (2) and (3) yields our proposed *permuted smooth tensor model*. The generating process is visualized in Figure 1(a) for the case $m = 2$ (matrices).

We give two concrete examples to show the applicability of our permuted smooth tensor model.

Example 1 (Four-player game tensor)**.** Consider a four-player board game. Suppose there are in total d players, among which all combinations of four have played against each other. The game results are naturally summarized as an order-4 (asymmetric) tensor, with entries encoding the winner of four-player games. Our model is then given by

$$\mathbb{E}(Y_{i_1,\ldots,i_4}) = \mathbb{P}(\text{user } i_1 \text{ wins over } (i_2,i_3,i_4)) = f\left(\frac{\pi(i_1)}{d},\cdots,\frac{\pi(i_4)}{d}\right).$$

In this setting, we can interpret the permutation π as the unknown ranking among d players, and the function f the unknown four-players interaction. Operationally, players with similar ranking would have similar performance encoded by the smoothness of f.

Example 2 (Co-authorship networks)**.** Consider co-authorship networks. Suppose there are in total d authors. We say there exists a hyperedge between nodes $(i_1,\ldots,i_m)$ if the authors $i_1,\ldots,i_m$ have co-authored at least one paper. The resulting hypergraph is represented as an order-m (symmetric) adjacency tensor. Our model is then expressed as

$$\mathbb{E}(Y_{i_1,\ldots,i_m}) = \mathbb{P}(\text{authors } i_1,\ldots,i_m \text{ co-authored}) = f\left(\frac{\pi(i_1)}{d},\cdots,\frac{\pi(i_m)}{d}\right).$$

In this setting, we can interpret the permutation π as the affinity measures of authors, and the function f represents the m-way interaction among authors.

3 Block-wise tensor approximation

Our general strategy for estimating the signal tensor is based on the block-wise tensor approximation. We first introduce the tensor block model [24, 11]. Then, we extend this model to the block-wise polynomial approximation.

3.1 Tensor block model

The tensor block model describes a checkerbroad pattern in the signal tensor. Specifically, suppose that there are k clusters in the tensor dimension d, and the clusters are represented by a clustering function $z\colon [d] \to [k]$. Then, the tensor block model assumes that signal tensor $\Theta \in \mathbb{R}^{d\times\cdots\times d}$ takes values from a mean tensor $\mathcal{S} \in \mathbb{R}^{k\times\cdots\times k}$ according to the clustering function z:

$$\Theta_{i_1,\ldots,i_m} = \mathcal{S}_{z(i_1),\ldots,z(i_m)}, \quad \text{for all } (i_1,\ldots,i_m) \in [d]^m. \tag{5}$$

A tensor Θ satisfying (5) is called a block-k tensor. The tensor block model has shown great success in discovering hidden group structure in many applications including hypergraph clustering [14], collaborative filtering [28] and signal dehe ection in 3D/4D imaging [27]. Despite its popularity and great applicability, the tensor block models cannot describe delicate structure of the signal tensor when the tensor dimension d is very large. This parametric model aims to explain data with a finite number of blocks; this approach is useful when the sample outsizes the parameters. Our nonparametric models (3), on the other hand, use infinite number of parameters to allow growing model complexity as sample increases. Therefore, we shift the goal of tensor block model from discovering hidden group structure to approximating the generative process of the function f in (3). Thus, the number of blocks k should be interpreted as a resolution parameter (i.e., a bandwidth) of the approximation similar to the notion of number of bins in histogram and polynomial regression.

3.2 Block-wise polynomial approximation

The tensor block model (5) can be viewed as a discrete version of piece-wise *constant* function with $\alpha = 0$ in (2). This connection motivates us to use block-wise *polynomial* tensors to approximate α-Hölder functions. Now we extend (5) to block-wise polynomial models. For a given block number k, we use $z\colon [d] \to [k]$ to denote the canonical clustering function that partitions $[d]$ into k clusters,

$$z(i) = \lceil ki/d \rceil, \quad \text{for all } i \in [d].$$

The collection of inverse images $\{z^{-1}(j)\colon j \in [k]\}$ consists of disjoint and equal-sized subsets in $[d]$, and we have $\cup_{j\in[k]} z^{-1}(j) = [d]$ by the construction. We denote $\mathcal{E}_k$ as the m-way partition as a collection of k^m disjoint, equal-sized blocks in $[d]^m$, such that

$$\mathcal{E}_k = \{z^{-1}(j_1) \times \cdots \times z^{-1}(j_m)\colon (j_1, \ldots, j_m) \in [k]^m\}.$$

We propose to approximate the signal tensor Θ in (3) by degree-ℓ polynomial tensor within each $\mathcal{E}_k$-block. Specifically, we use $\mathscr{B}(k, \ell)$ to denote the class of block-k, degree-ℓ polynomial tensors,

$$\mathscr{B}(k,\ell) = \left\{ \mathcal{B} \in (\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega) = \sum_{\Delta \in \mathcal{E}_k} \mathrm{Poly}_{\ell,\Delta}(\omega) \mathbb{1}\{\omega \in \Delta\} \text{ for all } \omega \in [d]^m \right\},$$

where $\mathrm{Poly}_{\ell,\Delta}(\cdot)$ denotes a degree-ℓ polynomial function in $\mathbb{R}^m$. Notice that degree-0 polynomial block tensor reduces to the tensor block model (5). We genealized the tensor block model to degree-ℓ polynomial block tensor, in a way that is analogous to the generalization from k-bin histogram to k-piece-wise polynomial regression in nonparametric statistics [25].

Smoothness of the function f in (3) turns out to play an important role in the block-wise polynomial approximation. The following lemma explains the role of smoothness in the approximation.

Lemma 1 (Tensor block approximation)**.** *Suppose $\Theta \in \mathcal{P}(\alpha)$. Then, for every block number $k \le d$, and degree $\ell \in \{0\} \cup \mathcal{N}_+$, we have the approximation error*

$$\inf_{\mathcal{B} \in \mathscr{B}(k,\ell)} \frac{1}{d^m} \|\Theta - \mathcal{B}\|_F^2 \lesssim \frac{m^2}{k^{2\min(\alpha, \ell+1)}}.$$

This theorem implies that we can always find block-wise polynomial tensor close to the signal tensor generated from α-Hölder smooth function f.

4 Fundamental limits via least-squares estimation

We propose two estimation methods based on the block-wise polynomial approximation. We first introduce a statistically optimal but computationally infeasbile least-squares estimator. The least-squares estimation serves as statistical benchmark because it achieves the minimax lower bound. In Section 5, we will present a polynomial-time algorithm with provably same optimal rates under monotonicity assumptions.

We propose the least-squares estimator for the signal tensor and the permutation (Θ, π) by minimizing the Frobenius loss under block-k, degree-ℓ polynomial tensor family $\mathscr{B}(k, \ell)$,

$$(\hat{\Theta}^{\mathrm{LSE}}, \hat{\pi}^{\mathrm{LSE}}) = \underset{\Theta \in \mathscr{B}(k,\ell),\ \pi \in S(d)}{\arg\min} \|\mathcal{Y} - \Theta \circ \pi\|_F. \tag{6}$$

The least-squares estimator $(\hat{\Theta}^{\mathrm{LSE}}, \hat{\pi}^{\mathrm{LSE}})$ depends on two tuning parameters: the number of blocks k and the polynomial degree ℓ. The optimal choice (k^*, ℓ^*) is provided in our next theorem. The result establishes the upper bound for the mean squared error of the least square estimator (6).

Theorem 1 (Least-squares estimation error)**.** *Consider the order-m ($m \geq 2$) permuted smooth tensor model* (2) *with $\Theta \in \mathcal{P}(\alpha)$. Let $(\hat{\Theta}^{\text{LSE}}, \hat{\pi}^{\text{LSE}})$ denote the least-squares estimates with a given (k, ℓ) in* (6)*. Then, the estimator $\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}}$ satisfies*

$$\frac{1}{d^m}\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \Theta \circ \pi\|_F^2 \lesssim \frac{m^2}{k^{2\min(\alpha,\ell+1)}} + \frac{k^m(\ell+m)^\ell}{d^m} + \frac{\log d}{d^{m-1}}, \tag{7}$$

with very high probability. In particular, setting $\ell^* = \min(\lceil\alpha\rceil, m(m-1)/2) - 1$ and $k^* = \lceil d^{\frac{m}{m+2\min(\alpha,\ell^*+1)}}\rceil$ yields the bound as

$$(7) \lesssim \begin{cases} d^{-\frac{2m\alpha}{m+2\alpha}} & \text{when } \alpha < \frac{m(m-1)}{2}, \\ \frac{\log d}{d^{m-1}} & \text{when } \alpha \geq \frac{m(m-1)}{2}. \end{cases}$$

We discuss the asymptotic error rates as $d \to \infty$ while treating the tensor order m and smoothness α fixed. The least square estimation error has two sources of error: the nonparametric error $d^{-\frac{2m\alpha}{m+2\alpha}}$ and the clustering error $\log d/d^{m-1}$. When the function f is smooth enough, estimating the function f becomes relatively easier compared to estimating the permutation π. This intuition coincides with the fact that the clustering error dominates the nonparametric error when $\alpha \geq m(m-1)/2$.

We now compare our results with existing work in the literature. Based on Theorem 1, the best rate is obtained with the choice of $(\ell^*, k^*) = (0, \lceil d^{\frac{1}{\alpha\wedge 1+1}}\rceil)$ in the matrix case ($m = 2$). This block-wise constant approximation and convergence rate reduce to the results in [8, 15]. Therefore, the least square estimation achieves the minimax optimal rate in matrix case. Furthermore, we solve the conjectured optimal convergence rate in [2] for higher order tensor case ($m \geq 3$). This improvement stems from polynomial tensor approximation in Lemma 1. The work in [2] considers only the block-wise constant approximation ($\ell = 0$). This restriction results in sub-optimality because the optimal degree ℓ^* is shown to be greater than 0 for higher-order tensors. For example, order-3 α-smooth tensors have the optimal degree and block size as $(\ell^*, k^*) = (2, \lceil d^{1/3}\rceil)$ for all $\alpha \geq 2$. This result shows the clear difference from matrices and highlights the challenges with tensors.

We now show that the upper bound of Theorem 1 is not only minimax optimal for the matrices but also for higher-order tensors. The result is based on information-theoretical analysis that combines the minimax rate for nonparametric and permutation estimation. Our minimax lower bound applies to all estimators including, but not limited to, least square estimator and all polynomial-time estimators.

Theorem 2 (Minimax lower bound)**.** *For any given $\alpha \in (0, \infty)$, the estimation problem based on model* (1) *obeys the minimax lower bound*

$$\inf_{(\hat{\Theta},\hat{\pi})} \sup_{\Theta\in\mathcal{P}(\alpha),\pi\in S(d)} \mathbb{P}\left(\frac{1}{d^m}\|\Theta \circ \pi - \hat{\Theta} \circ \hat{\pi}\|_F^2 \gtrsim d^{-\frac{2m\alpha}{m+2\alpha}} + d^{-(m-1)}\log d\right) \geq 0.8.$$

We see that the lower bound matches the upper bound in Theorems 1. Therefore, the least square estimator (6) is statistically optimal.

5 An adaptive and computationally feasible procedure

At this point, we should point out that computing the least square optimizer $(\hat{\Theta}^{\text{LSE}}, \hat{\pi}^{\text{LSE}})$ in (6) with polynomial-time algorithm is unknown. We suspect that the algorithm for (6) may be computationally intractable. In this section, we propose an efficient polynomial-time *Borda count* algorithm. Furthermore, we show that Borda count estimator actually achieves the same convergence rate as the minimax lower bound under the β-monotonicity condition.

5.1 Borda count algorithm

We first introduce β-monotonicity for the generating functions.

Definition 2 (β-monotonicity)**.** A function $f\colon [0,1]^m \to \mathbb{R}$ is called β-monotonic, denoted as $f \in \mathcal{M}(\beta)$, if

$$\left(\frac{i-j}{d}\right)^{1/\beta} \leq g(i) - g(j), \quad \text{for all } i > j \in [d], \tag{8}$$

where we define $g(i) = d^{-(m-1)}\sum_{(i_2,\ldots,i_m)\in[d]^m} f\left(\frac{i}{d}, \frac{i_2}{d}, \ldots, \frac{i_m}{d}\right)$ for all $i \in [d]$.

This β-monotonicity condition can be viewed as an extension of the strict monotonic degree condition in binary-valued networks [4] to general setting. Consider the hypergraph example where the observed tensor $\mathcal{Y}$ is the adjacency tensor representing the connectivity among d-nodes. Then, the function g is the degree function of the nodes. Our β-monotonicity condition is also closely related to isotonic functions [10, 20] which assume the coordinate-wise monotonicity, i.e., $f(x_1, \ldots, x_d) \leq f(x'_1, \ldots, x'_d)$ when $x_i \leq x'_i$ for $i \in [d]$.

This β-monotonicity condition allows us to estimate the permutation π in polynomial-time complexity. Before presenting the theoretical guarantees, we provide the intuition here. The parameter β measures the difficulty of the problem for estimating the permutation π. Consider the noisy observation $\mathcal{Y}$ in (1). We define the scores function $\tau\colon [d] \to \mathbb{R}$ as

$$\tau(i) = \frac{1}{d^{m-1}} \sum_{(i_2,\ldots,i_m)\in[d]^m} \mathcal{Y}_{i,i_2,\ldots,i_m}.$$

Then, the permuted score function $\tau \circ \pi^{-1}$ is equivalent to the function g in (8) for the noiseless case. Therefore, we can find an estimate $\hat{\pi}$ that makes the permuted score function $\tau \circ \hat{\pi}^{-1}$ monotonically increasing. Notice that the estimated permutation $\hat{\pi}$ could be different from the oracle permutation π due to the noise. We find that the larger β guarantees the sharper consistency of $\hat{\pi}$. The large β implies the large gaps of $|g(i) - g(j)|$ for $i \neq j \in [d]$. Therefore, we obtain similar ordering of $\{\tau(i)\}_{i=1}^d$ before and after the addition of the noise. This intuition is well represented by the following lemma.

Lemma 2 (Permutation error)**.** *Let $\hat{\pi}$ be the permutation that makes the permuted score function $\tau \circ \hat{\pi}^{-1}$ monotonically increasing. Then, we have*

$$\mathrm{Loss}(\pi, \hat{\pi}) := \frac{1}{d} \max_{i\in[d]} |\pi(i) - \hat{\pi}(i)| \lesssim \left(\sigma d^{-(m-1)/2} \sqrt{\log d}\right)^{\beta},$$

with high probability.

Now we introduce a Borda count estimator that consists of two stages. The full estimation procedure is illustrated in Figure 2.

Sorting stage: The purpose of the sorting is to rearrange the observed tensor $\mathcal{Y}$ so that the score function τ of sorted tensor is monotonically increasing. We define a permutation $\hat{\pi}^{\mathrm{BC}}$ such that

$$\tau((\hat{\pi}^{\mathrm{BC}})^{-1}(1)) \leq \cdots \leq \tau((\hat{\pi}^{\mathrm{BC}})^{-1}(d)). \tag{9}$$

Then, we obtain a sorted observation $\tilde{\mathcal{Y}}$,

$$\tilde{\mathcal{Y}}_{i_1,\ldots,i_m} = \mathcal{Y}_{(\hat{\pi}^{\mathrm{BC}})^{-1}(i_1),\ldots,(\hat{\pi}^{\mathrm{BC}})^{-1}(i_m)},$$

for all $(i_1, \ldots, i_m) \in [d]^m$. An example of sorted observation is shown in Figure 2.

Block-wise polynomial approximation stage: Given degree ℓ, we estimate the degree-ℓ polynomial block tensor based on the sorted observation $\tilde{\mathcal{Y}}$ solving the following optimization problem,

$$\hat{\Theta}^{\mathrm{BC}} = \underset{\mathcal{B}\in\mathscr{B}(k,\ell)}{\arg\min} \|\tilde{\mathcal{Y}} - \Theta\|_F. \tag{10}$$

The estimate $\hat{\Theta}^{\mathrm{BC}}$ depends on two tuning parameters: the number of blocks k and polynomial degree ℓ. The optimal choice of (k^*, ℓ^*) is provided in Theorem 3. Notice that the least square estimation in (6) requires combinatoric search for the permutation resulting in exponential time complexity. However, (10) only requires to estimate the degree-ℓ polynomial block tensor. Therefore, this step easily reduces to a degree-ℓ polynomial regression problem within each block $\mathcal{E}_k$.

5.2 Computational and statistical complexity

The complexity of the Borda count algorithm can be computed separately in each stage. In the sorting stage, computing the score function τ requires $\mathcal{O}(d^{m-1})$ additions while sorting the $\tau(1), \ldots, \tau(d)$ takes about $\mathcal{O}(d \log d)$ comparisons. In block-wise polynomial approximation stage, we compute k^m different degree-ℓ polynomial tensors. For each degree-ℓ polynomial tensor, $\mathcal{O}((d/k)^m \ell)$ arithmetic operations are needed. Thus, the second step requires $\mathcal{O}(d^m \ell)$ arithmetic operations. Combining these two steps yields the total complexity at most $\mathcal{O}(d^m \log d)$.

We show the consistency of the signal tensor estimation based on Lemma 1-2.

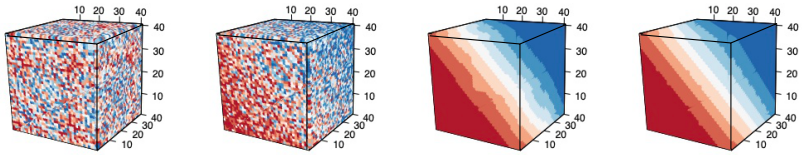

Figure 2: Procedure of Borda count estimation. We first sort the tensor entries using the proposed procedure. Then, we estimate the signal tensor using block-k degree-ℓ polynomial approximation.

Theorem 3 (Estimation error for Borda count)**.** *Suppose that the signal tensor Θ is generated as in* (3) *with $f \in \mathcal{H}(\alpha) \cap \mathcal{M}(\beta)$. Let $(\hat{\Theta}^{\text{BC}}, \hat{\pi}^{\text{BC}})$ be the Borda count estimator in* (9)-(10) *with a given (k, ℓ). Then, for every $k \leq d$ and degree $\ell \in \mathbb{N}_{\geq 0}$, we have*

$$\text{MSE}(\hat{\Theta}^{\text{BC}} \circ \hat{\pi}^{\text{BC}},\ \Theta \circ \pi) \lesssim \frac{m^2}{k^{2\min(\alpha,\ell+1)}} + \frac{k^m(\ell+m)^\ell}{d^m} + \left(\frac{\log d}{d^{m-1}}\right)^{\beta\min(\alpha,1)}, \qquad (11)$$

with high probability.

The three terms in the estimation bound (11) correspond to approximation error (Lemma 1), non-parametric error (Theorem 1), and permutation error (Lemma 2), respectively. We find that the Borda count estimator achieves the same minimax-optimal rate as the least-squares estimator for sufficiently smooth tensors under Lipschitz score condition $\beta = 1$. The least-squares estimator requires a combinatoric search with exponential-time complexity. By contrast, the Borda count estimator is polynomial-time solvable. Therefore, Borda count algorithm enjoys both statistical accuracy and computational efficiency.

6 Numerical comparisons

We simulate symmetric order-3 d-dimensional tensors based on the permuted smooth tensor model (3) with function f in Table 2. Notice that considered functions cover a reasonable range of model complexities from low rank to high rank. We generate the entries of the noise tensor i.i.d. from Gaussian distribution $N(0, 0.5^2)$. The permutation π is randomly sampled from all permutations from $[d]$ to $[d]$. Throughout all experiments, we evaluate the accuracy of the estimation by mean square error (MSE) $= d^{-3}\|\Theta \circ \pi - \hat{\Theta} \circ \hat{\pi}\|_F^2$ across $n_{\text{sim}} = 20$ replications.

Table 2: Smooth functions in simulation. We define the numerical CP/Tucker rank as the minimal rank r for which the relative approximation error is below 10^{-4}. The reported rank in the table is estimated from a $100 \times 100 \times 100$ signal tensor generated by (3).

Model ID	$f(x, y, z)$	CP rank	Tucker rank
1	xyz	1	$(1, 1, 1)$
2	$(1 + \exp(-3x^2 + 3y^2 + 3z^2))^{-1}$	9	$(4, 4, 4)$
3	$\exp\left(-\max(x, y, z) - \sqrt{x} - \sqrt{y} - \sqrt{z}\right)$	≥ 100	$(90, 90, 90)$

The first experiment examines the impact of the block number k and degree of polynomial ℓ for the approximation. We fix the tensor dimension $d = 100$, and vary the number of blocks $k \in \{1, \ldots, 15\}$ and polynomial degree $\ell \in \{0, 1, 2, 3\}$. Figure 3 demonstrates the trade-off in accuracy determined by the number of groups for each polynomial degree. The results are consistent to our bias-variance analysis in Theorem 1. While a large block number k provides less biased approximation, this large k renders the signal tensor estimation difficult within each block due to small sample size. In addition, we find that degree-2 polynomial approximation with the optimal k gives the smallest MSE among all considered polynomial approximation. These two observations are well explained by our theoretical results where the optimal number of blocks and polynomial degree are $(\mathcal{O}(\lceil d^{3/7} \rceil, 2)$.

The second experiment compares our method (**Borda Count**) with several popular alternative methods: (a) Spectral method (**Spectral**) [26] that performs universal singular value thresholding [5] on the unfolded tensor; (b) Least square estimation (**LSE**) [2], which solves the optimization problem (6) with constant block approximation ($\ell = 0$) [8]; (c) Our **Borda Count** algorithm. We choose degree-2

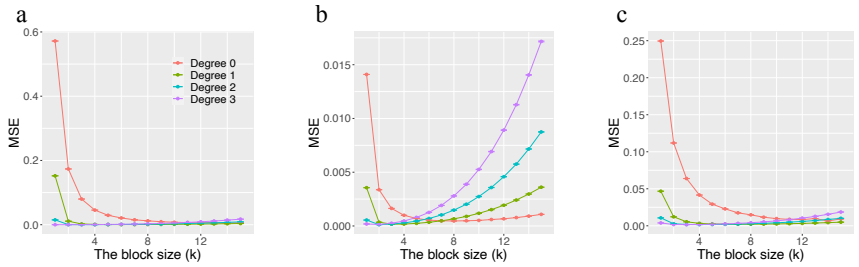

Figure 3: MSE comparison versus the number of blocks for different polynomial approximation. Panels a-c show the results under the models 1-3 respectively.

polynomial approximation as our theorems suggested, and vary tensor dimension $d \in \{10, \ldots, 100\}$ under each model specification. We choose the block number for **Borda Count** and **LSE**, which achieves the best performance based on the intuition in our theorems and Figure 3. Similarly, we set the threshold value that obtains the best performance for **Spectral**.

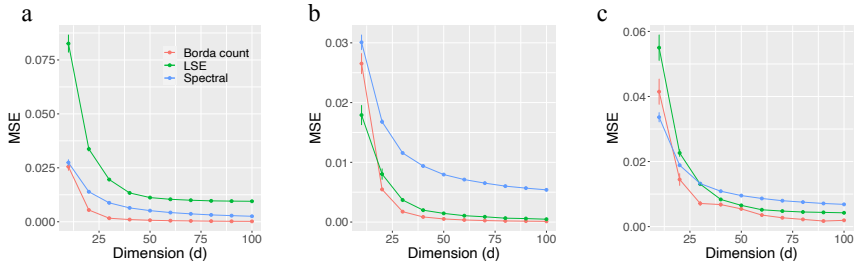

Figure 4: MSE comparison of different methods versus tensor dimension. Panels a-c show the results under models 1-3 respectively.

Figure 4 shows that our algorithm **Borda Count** achieves the best performance in all scenarios as the tensor dimension increases. The poor performance of **Spectral** can be explained by the loss of multilinear structure in the tensor unfolding procedure. The sub-optimality of **LSE** is possibly due to its limits in both statistics and computations. Statistically, our theorems have shown that constant block approximation has sub-optimal rates compared to polynomial approximation for higher-order tensors. Computationally, the least square optimization (6) is highly non-convex and computationally unstable. The outperformance of **Borda count** demonstrates the efficacy of our method.

7 Application to Chicago crime data

Chicago crime dataset consists of crime counts reported in the city of Chicago, ranging from January 1st, 2001 to December 11th, 2017. The observed tensor is an order-3 tensor with entries representing the log counts of crimes from 24 hours, 77 community areas, and 32 crime types. We apply our Borda Count method to Chicago crime dataset. Because the data tensor is asymmetric, we allow different number of blocks across the three modes. Cross validation result suggests the $(k_1, k_2, k_3) = (6, 4, 10)$, representing the block number for crime hours, community areas, and crime types, respectively.

We first investigate the four clustered community areas obtained from our Borda Count algorithm. Figure 5(b) shows the four areas overlaid on a map of Chicago. Interestingly, we find that the clusters conform the actual locations even though our algorithm did not take any geographic information such as longitude or latitude as inputs. In addition, we compare the cluster patterns with benchmark results based on homicides- and shooting incidents-maps in Chicago shown in Figure 5(a). We find that our clusters share similar geographical patterns with Figure 5(a). The benchmark Figure 5(a) covers only homicides and shooting incidents in 2020, whereas our result in Figure 5(b) considers 32 crime types across 2001-2017. The results demonstrate the power of our approach in detecting meaningful pattern from tensor data.

Then, we examine the denoised signal tensor obtained from our method and analyze the trends between crime types and crime hours by the four community areas in Figure 5(b). Figure 6 shows the averaged log counts of crimes according to crime types and crime hours by four areas. We find that the major difference among four areas is the crime rates. Area 4 has the highest crime rates, and the

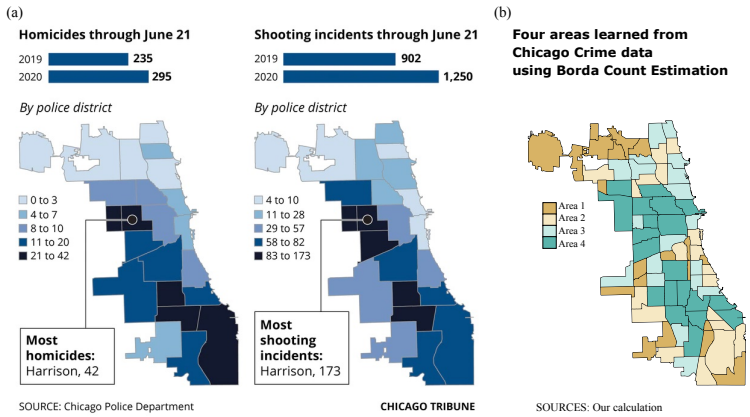

Figure 5: Chicago crime maps. Figure(a) shows homicides and shooting incidents in community areas in Chicago. This figure is from *Chicago Tribune* article in 2020 [13]. Figure(b) shows the four areas estimated by our Borda Count algorithm.

crime rates monotonically decrease from Area 4 to Area 1. The variation in crime rates across hour and type, nevertheless, exhibits similarity among the four areas. For example, Figure 6 shows that the number of crimes increases hourly from 8 p.m., peaks at night hours, and then drops to the lowest at 6 p.m. The identified similarities and differences among the four community areas highlight the interpretability of our method in real data.

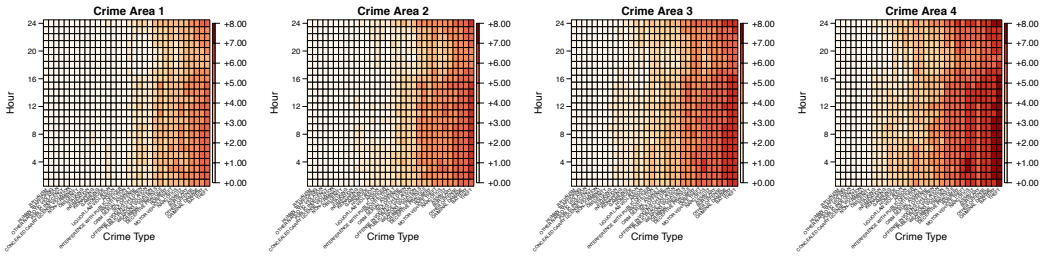

Figure 6: Averaged log counts of crimes according to crime types, hours, and the four areas estimated by our Borda Count algorithm. We plot the estimated signal tensor entries averaged within four areas in the heatmap.

Finally, we compare the block-wise constant approximation versus block-wise degree-2 polynomial approximation. We assess the goodness-of-fit by considering the hypothesis testing

$$H_0 : \Theta \text{ is a block-wise constant tensor} \quad \text{vs.} \quad H_1 : \Theta \text{ is a block-wise polynomial degree-2 tensor.}$$

Notice that the class of block-wise constant tensors is nested within that of block-wise polynomial degree-2 tensors. Therefore, we perform a F-test considering the degree of freedom of each model. The degree of freedom of H_0 is the number of total blocks, $6 \times 4 \times 10$. The degree of H_1 is $10 \times 6 \times 4 \times 10$ because degree-2 polynomials have 10 times more coefficients than the constant block model. The obtained F-statistics from Chicago crime dataset is 19.63 with p-value is $< 10^{-3}$. The result provides the significant evidence for the validity of the polynomial approximation. We emphasize that our method does not necessarily assume the block structure. We present F-test result as an evidence supporting our premises that permuted smooth tensor model with polynomial approximation performs better than common tensor block models in this application.

8 Conclusion

We have developed permuted smooth tensor model and estimation methods with theoretical guarantees. The efficacy of our procedure is demonstrated through both simulations and analysis of Chicago crime dateset.

Acknowledgements

This research is supported in part by NSF grant DMS- 1915978, NSF DMS-2023239, and Wisconsin Alumni Research Foundation.

References

[1] Sameer Agarwal, Kristin Branson, and Serge J. Belongie. Higher order learning with graphs. *Proceedings of the 23rd international conference on Machine learning*, 2006.

[2] Krishnakumar Balasubramanian. Nonparametric modeling of higher-order interactions via hypergraphons. *arXiv preprint arXiv:2105.08678*, 2021.

[3] Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Aykan Aydin, Karl-Heinz Lüke, and Roland Schwaiger. Incarmusic: Context-aware music recommendations in a car. In *International Conference on Electronic Commerce and Web Technologies*, pages 89–100. Springer, 2011.

[4] Stanley Chan and Edoardo Airoldi. A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216, 2014.

[5] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[6] Jian Ding, Zongming Ma, Yihong Wu, and Jiaming Xu. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1):29–115, 2021.

[7] Nicolas Flammarion, Cheng Mao, and Philippe Rigollet. Optimal rates of statistical seriation. *Bernoulli*, 25(1):623–653, 2019.

[8] Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. Optimal estimation and completion of matrices with biclustering structures. *Journal of Machine Learning Research*, 17(1):5602–5630, 2016.

[9] Chao Gao and Zongming Ma. Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *arXiv preprint arXiv:1811.06055*, 2018.

[10] Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019.

[11] Rungang Han, Yuetian Luo, Miaoyan Wang, and Anru R Zhang. Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*, 2020.

[12] Jan-Christian Hütter, Cheng Mao, Philippe Rigollet, and Elina Robeva. Estimation of monge matrices. *Bernoulli*, 26(4):3051–3080, 2020.

[13] Gorner Jeremy. A trying first half of 2020 included spike in shootings and homicides in chicago. *Chicago Tribune*.

[14] Zheng Tracy Ke, Feng Shi, and Dong Xia. Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*, 2019.

[15] Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

[16] Yihua Li, Devavrat Shah, Dogyoon Song, and Christina Lee Yu. Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 66(3):1760–1784, 2019.

[17] Lorenzo Livi and Antonello Rizzi. The graph matching problem. *Pattern Analysis and Applications*, 16(3):253–283, 2013.

[18] Tom Michoel and Bruno Nachtergaele. Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 86 5 Pt 2:056111, 2012.

[19] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*, volume 11, pages 809–816, 2011.

[20] Ashwin Pananjady and Richard J Samworth. Isotonic regression with unknown permutations: Statistics, computation, and adaptation. *arXiv preprint arXiv:2009.02609*, 2020.

[21] N Shah, Sivaraman Balakrishnan, and M Wainwright. Low permutation-rank matrices: Structural properties and noisy completion. *Journal of machine learning research*, 2019.

[22] Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*, 67(6):4162–4184, 2020.

[23] Lu Wang, Zhengwu Zhang, and David Dunson. Common and individual structure of brain networks. *The Annals of Applied Statistics*, 13(1):85–112, 2019.

[24] Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, pages 713–723, 2019.

[25] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[26] Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, pages 5433–5442, 2018.

[27] Chenyu Zhang, Rungang Han, Anru R Zhang, and Paul M Voyles. Denoising atomic resolution 4d scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy*, 219:113123, 2020.

[28] Yanqing Zhang, Xuan Bi, Niansheng Tang, and Annie Qu. Dynamic tensor recommender systems. *Journal of Machine Learning Research*, 22(65):1–35, 2021.

[29] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.