Smooth tensor estimation with unknown permutations

Abstract

We consider the problem of structured tensor denoising in the presence of unknown permutations. Such data problems arise commonly in recommendation system, 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 $(m-2)(m+1)/2$ is sufficient for accurate recovery of order-m tensors, whereas higher smoothness 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 applications.

1 Introduction

Higher-order tensor datasets arise ubiquitously in modern data science applications. 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 [2]. 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 social sciences [4] and recommendation system [6]. In both examples, higher-order tensors represent multi-way interactions in an efficient way.

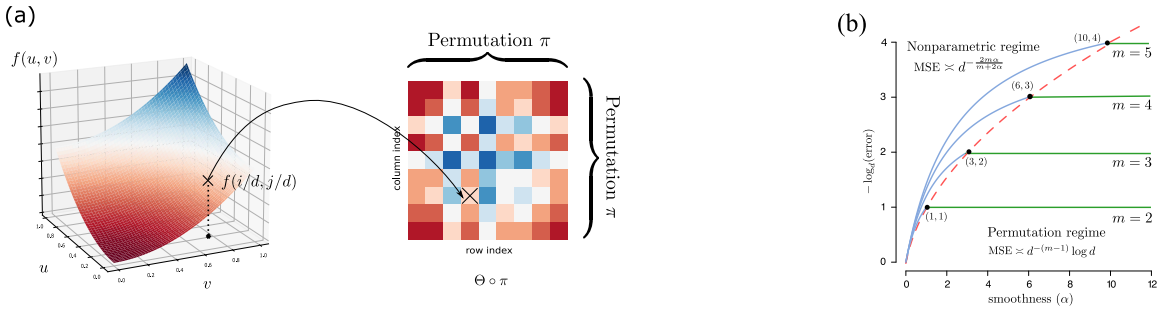


Figure 1: (a): Illustration of order-m d-dimensional permuted smooth tensor models with $m = 2$. (b): Phase transition of mean squared error (MSE) (on -$\log d$ scale) as a function of smoothness α and tensor order m. Bold dots correspond to the critical smoothness level above which higher smoothness exhibits no further benefits to tensor estimation. See Theorems 1-3 in Sections 3-4 for details.

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

$$\mathcal{Y} = \Theta \circ \pi + \mathcal{E}, \quad \text{where} \quad \Theta_{i_1,\ldots,i_m} = f\left(\frac{i_1}{d}, \ldots, \frac{i_m}{d}\right). \tag{1}$$

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 certain notion of smoothness, $\Theta \circ \pi$ denotes the permuted tensor after reordering the indices along each of the m modes, and $\mathcal{E}$ is a symmetric noise tensor consisting of zero-mean, independent sub-Gaussian entries with variance bounded by σ^2. Figure 1(a) shows an example of this generative model for the matrix case $m = 2$.

For ease of presentation, we focus on symmetric tensors; our models and techniques easily generalize to non-symmetric tensors. Our primary goal is to estimate a permuted smooth signal tensor from a noisy observation (1).

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, 8], stochastic transitivity models [11], and crowd labeling [9]. Most of these methods are developed for matrices. The tensor counterparts are far less well understood.

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

	Pananjady et al [10]	Balasubramanian [1]	Li et al [9]	**Ours***
Model structure	monotonic	Lipschitz	Lipschitz	α-smoothness
Minimax lower bound	$\surd$	$\times$	$\times$	$\surd$
Error rate for order-m tensor*	d^{-1}	$d^{-2m/(m+2)}$	$d^{-\lfloor m/3 \rfloor}$	$d^{-(m-1)}$
(e.g., when $m=3$)		$(d^{-6/5})$	(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 1-3 in Sections 3-4.

• We develop a general permuted α-smooth tensor model, where $\alpha \geq 0$ is some natural measure of functional smoothness (formal definition in Section 2). In contrast to earlier work [1, 9], we establish the statistically optimal rate and fully characterize its dependence on tensor order, dimension, and smoothness index. Table 1 summarizes our improvement over previous works on tensor learning with latent permutations.

• We discover a phase transition phenomenon with respect to the smoothness threshold needed for optimal recovery in model (1). Figure 1(b) plots the dependence of estimation error in terms of smoothness level α for tensors of order m. We find that the estimation accuracy improves with smoothness in the regime $\alpha \leq m(m-1)/2$, but then it becomes a constant of α in the regime $\alpha > m(m-1)/2$. The phenomenon is distinctive from matrix problems [8] and classical *non-permuted* smooth function estimation, thereby highlighting the fundamental challenges in our new setting.

• We propose two estimation algorithms with accuracy guarantees: the least-squares estimation and Borda count estimation. The least-squares estimation is minimax optimal but computationally hard. The Borda count algorithm is polynomial-solvable, and we show it provably achieves the same optimal rate under extra monotonicity assumptions. Application to Chicago crime analysis is presented to showcase the usefulness of our method. The software package and all data used have been publicly released at CRAN.

Notation. We use $[d] = \{1, \ldots, d\}$ for d-set with $d \in \mathbb{N}_+$. For a set S, $\mathbb{1}_S$ denotes the indicator function. For positive two sequences $\{a_n\}, \{b_n\}$, we denote $a_n \lesssim b_n$ if $\lim_{n\to\infty} a_n/b_n \leq c$, and $a_n \asymp b_n$ if $c_1 \leq \lim_{n\to\infty} a_n/b_n \leq c_2$ for some constants $c, c_1, c_2 > 0$. 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. An event A is said to occur *with high probability* if $\mathbb{P}(A)$ tends to 1 as the tensor dimension $d \to \infty$.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 and block-wise approximation

Suppose we observe an order-m d-dimensional symmetric data tensor from the permuted tensors in (1). We 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 $\text{Poly}_{\lfloor\alpha\rfloor}(\boldsymbol{x} - \boldsymbol{x}_0)$ of degree $\lfloor\alpha\rfloor$, such that

$$|f(\boldsymbol{x}) - \text{Poly}_{\lfloor\alpha\rfloor}(\boldsymbol{x} - \boldsymbol{x}_0)| \leq C\|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty^\alpha, \text{ for all } \boldsymbol{x}, \boldsymbol{x}_0 \in [0,1]^m \text{ and a constant } C > 0. \tag{2}$$

In addition to the function class $\mathcal{H}(\alpha)$, we define the smooth tensor class based on discretization (1),

$$\mathcal{P}(\alpha) = \left\{\Theta \in \mathbb{R}^{d\times\cdots\times d} \colon \Theta(\omega) = f\left(\frac{\omega}{d}\right) \text{ for all } \omega = (i_1, \ldots, i_m) \in [d]^m \text{ and } f \in \mathcal{H}(\alpha)\right\}. \tag{3}$$

Combining (1) and (2) yields our proposed *permuted smooth tensor model*. The unknown parameters are the smooth tensor $\Theta \in \mathcal{P}(\alpha)$ and latent permutation $\pi \in S(d)$. The model 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 summarized as an order-4 (asymmetric) tensor, with entries encoding the winner of the games. Our model is then given by

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

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 $(i_1, \dots, i_m)$ if the authors $i_1, \dots, i_m$ 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

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

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. Our nonparametric model learns the unknown function f from data.

Our general strategy for estimating the signal tensor in model (3) is based on the block-wise tensor approximation. We first introduce the tensor block model [6]. Then, we extend the idea to block-wise polynomial approximation.

Tensor block model. The tensor block model [6] 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:

$$\Theta_{i_1,\dots,i_m} = \mathcal{S}_{z(i_1),\dots,z(i_m)}, \quad \text{for all } (i_1, \dots, i_m) \in [d]^m. \tag{4}$$

A tensor Θ satisfying (4) is called a block-k tensor. Classical tensor block models aim to explain data with a finite number of blocks; this approach is useful when the sample outsizes the parameters. Our nonparametric models (1), by contrast, 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 (1). In our setting, 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.

Block-wise polynomial approximation. The tensor block model (4) can be viewed as a discrete version of piece-wise *constant* function. This connection motivates us to use block-wise *polynomial* tensors to approximate α-Hölder functions. 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, $z(i) = \lceil ki/d \rceil$, 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

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

We refer to $\Delta \in \mathcal{E}_k$ as the *canonical blocks*. We propose to approximate the signal Θ in (1) by degree-ℓ polynomial tensor within each block $\Delta \in \mathcal{E}_k$. 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 (4). We generalize the tensor block model to degree-ℓ polynomial block tensor, in a way analogous to the generalization from k-bin histogram to k-piece-wise polynomial regression.

3 Fundamental limits via least-squares estimation

We develop two estimation methods based on the block-wise polynomial approximation. We first introduce a minimax optimal but computationally inefficient least-squares estimator as a statistical benchmark. In Section 4, we will present a polynomial-time solvable estimator with provably same optimal rate under monotonicity assumptions.

We propose the least-squares estimation for model (1) by minimizing the Frobenius loss under block-k, degree-ℓ polynomial tensor family $\mathscr{B}(k, \ell)$,

$$(\hat{\Theta}^{\mathrm{LSE}}, \hat{\pi}^{\mathrm{LSE}}) = \underset{\Theta\in\mathscr{B}(k,\ell),\ \pi\in S(d)}{\arg\min} \|\mathcal{Y} - \Theta \circ \pi\|_F. \tag{5}$$

The least-squares estimator $(\hat{\Theta}^{\text{LSE}}, \hat{\pi}^{\text{LSE}})$ 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.

Theorem 1 (Least-squares estimation error)**.** *Consider the order-m ($m \geq 2$) permuted smooth tensor model* (1) *with $\Theta \in \mathcal{P}(\alpha)$. Then, the estimator $\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}}$ in* (5) *satisfies with high probability*

$$\frac{1}{d^m}\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \Theta \circ \pi\|_F^2 \lesssim \begin{cases} d^{-\frac{2m\alpha}{m+2\alpha}} & \text{when } \alpha < m(m-1)/2, \\ \frac{\log d}{d^{m-1}} & \text{when } \alpha \geq m(m-1)/2, \end{cases} \tag{6}$$

under the optimal choice of $\ell^* = \min(\lfloor \alpha \rfloor, (m-2)(m+1)/2)$ and $k^* = \lceil d^{\frac{m}{m+2\min(\alpha, \ell^*+1)}} \rceil$.

Theorem 1 establishes the upper bound for the mean squared error of the least-squares estimator (5). We discuss the asymptotic error rates as $d \to \infty$ while treating the tensor order m and smoothness α fixed. The least-squares 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. In the matrix case ($m = 2$), our block-wise constant approximation and convergence rate reduce to the results in [8]. For higher order tensor case ($m \geq 3$), earlier work [1] conjectures that constant block approximation ($\ell^* = 0$) remains minimax optimal for tensors. Our Theorem 1 disproves this conjecture, and we reveal a much faster rate $d^{-(m-1)}$ compared to the conjectured lower bound $d^{-2m/(m+2)}$ [1]. In fact, permuted α-smooth tensors of order-3 require quadratic approximation ($\ell^* = 2$) with $k^* \asymp d^{1/3}$ blocks, for all $\alpha \geq 2$. The results show the clear difference from matrices and highlight the challenges with tensors.

The next theorem shows that the critical polynomial degree up to $(m-2)(m+1)/2$ is not only sufficient but also necessary for accurate estimation of order-m permuted smooth tensors.

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

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

The above result demonstrates that the upper bound (6) is minimax optimal. Theorem 2 is obtained via information-theoretical analysis and thus applies to all estimators including, but not limited to, the least-squares estimator (5) and Borda count estimator introduced in next section.

4 An adaptive and computationally feasible procedure

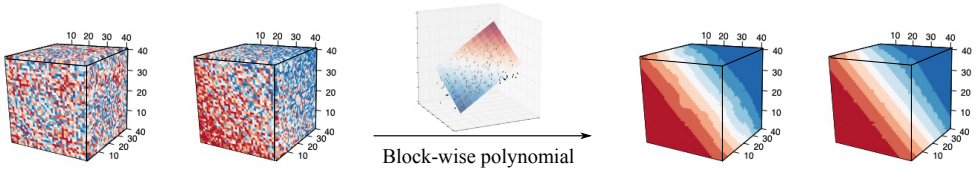
At this point, we should note that the least-squares estimation in (5) is generally computationally hard. In this section, we propose an efficient polynomial-time *Borda count* algorithm with provably same optimal rate under the β-monotonicity condition. We first introduce β-monotonicity condition.

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

$$\left(\frac{i-j}{d}\right)^{1/\beta} \leq g(i) - g(j) \;\text{ for all } i > j \in [d], \quad \text{where } g(i) := \frac{1}{d^{m-1}} \sum_{(i_2,\ldots,i_m) \in [d]^m} f\left(\frac{i}{d}, \frac{i_2}{d}, \ldots, \frac{i_m}{d}\right).$$

Our β-monotonicity condition extends the strictly monotonic degree condition in the graphon literature [3]; the latter is a special case of our definition with $\beta = 1, m = 2$. Our β-monotonicity condition is also related to isotonic functions [5, 10] 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 $i \in [d]$.

Now we introduce a Borda count estimator that consists of two stages: sorting and block-wise polynomial approximation. The simplified version of the algorithm is described in Algorithm 1 and Figure 2.



(a) Observation (b) Sorted observation (c) Estimated signal (d) True signal

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.

Algorithm 1 Borda Count algorithm

Input: Noisy observed data tensor $\mathcal{Y} \in \mathbb{R}^{d\times\cdots\times d}$

1: **Sorting stage:** Compute a permutation $\hat{\pi}^{\mathrm{BC}}$ such that $\tau \circ (\hat{\pi}^{\mathrm{BC}})^{-1}$ is monotonically increasing, where $\tau(i) = \frac{1}{d^{m-1}} \sum_{(i_2,\ldots,i_m)\in[d]^m} \mathcal{Y}_{i,i_2,\ldots,i_m}$.

2: Obtain a rearranged observation $\tilde{\mathcal{Y}}_{i_1,\ldots,i_m} = \mathcal{Y}_{(\hat{\pi}^{\mathrm{BC}})^{-1}(i_1),\ldots,(\hat{\pi}^{\mathrm{BC}})^{-1}(i_m)}$

3: **Block-wise polynomial approximation stage:** Given degree ℓ and block k, solve the following optimization problem, $\hat{\Theta}^{\mathrm{BC}} = \arg\min_{\mathcal{B}\in\mathscr{B}(k,\ell)} \|\tilde{\mathcal{Y}} - \Theta\|_F$.

Output: Estimated signal tensor and permutation $(\hat{\Theta}^{\mathrm{BC}}, \hat{\pi}^{\mathrm{BC}})$.

Theorem 3 (Estimation error for Borda count; simplified version)**.** *Suppose that the signal tensor Θ is generated as in* (1) *with $f \in \mathcal{H}(\alpha) \cap \mathcal{M}(\beta)$. Then estimators $(\hat{\Theta}^{BC}, \hat{\pi}^{BC})$ from Algorithm 1 satisfies*

$$\frac{1}{d^m}\|\hat{\Theta}^{\mathrm{BC}} \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F^2 \lesssim \begin{cases} d^{-\frac{2m\alpha}{m+2\alpha}} & \text{when } \alpha < c(\alpha,\beta,m), \\ \left(\frac{\log d}{d^{m-1}}\right)^{\beta\min(\alpha,1)} & \text{when } \alpha \geq c(\alpha,\beta,m), \end{cases}$$

with high probability under the optimal choice of $\ell^* = \min(\lfloor\alpha\rfloor, \lfloor c(\alpha,\beta,m)\rfloor)$ and $k^* = \lceil d^{\frac{m}{m+2\min(\alpha,\ell^*+1)}}\rceil$. Here $c(\alpha,\beta,m) > 0$ is a constant only depending on α, β, and m.

Theorem 3 shows the estimation consistency of Borda count estimator. We find that the Borda count estimator achieves the same minimax-optimal rate as the least-squares estimator under 1-monotonicity condition. The least-squares estimator requires a combinatoric search with exponential-time complexity. By contrast, Algorithm 1 requires only the estimation of degree-ℓ polynomials within k *canonical blocks*. Therefore, the Borda count estimator is polynomial-time efficient.

5 Numerical experiments and data application

Numerical comparisons. We simulate symmetric order-3 d-dimensional tensors based on the permuted smooth tensor model (1) with diverse functions f. The detailed simulation procedure is described in Appendix. We assess the performance for the four popular tensor methods: (a) Spectral method (**Spectral**) [12] on unfolded tensor; (b) Least-squares estimation (**LSE**) with $\ell = 0$ implied by [4]; (c) Lease square estimation (**BAL**) implied by [1]; (d) Our **Borda Count** algorithm. The performance accuracy is assessed via mean square error (MSE) $= d^{-3}\|\Theta \circ \pi - \hat{\Theta} \circ \hat{\pi}\|_F^2$.

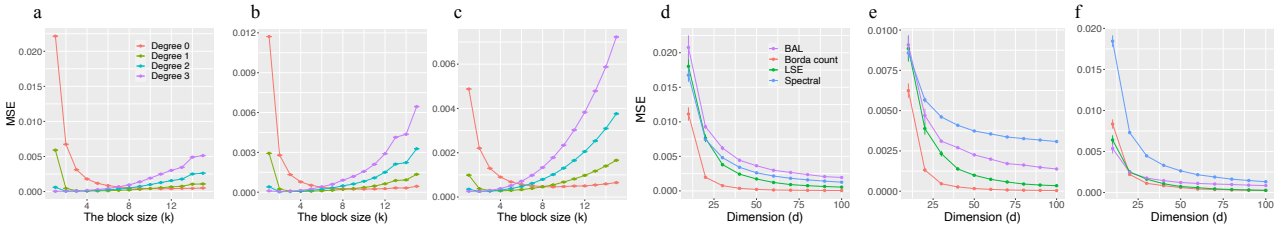


Figure 3: (a-c) MSE comparison versus the number of blocks under simulation models 1,3 and 5 respectively. (d-f) MSE comparison versus tensor dimension under models 1,3 and 5 respectively. MSEs are measured across $n_{\mathrm{sim}} = 20$ replications.

Figure 3a-c examine 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 $\ell \in \{0,1,2,3\}$. The result demonstrates the trade-off in accuracy determined by the number of blocks for each polynomial degree. We find that degree-2 polynomials give the smallest MSE among all considered approximation for order-3 tensors. These observations are consistent with our theoretical results in Sections 3-4. Figure 3d-f 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 the least square estimations is possibly due to its limits in both statistics and computations. Statistically, our theorems have shown that constant block approximation has sub-optimal rates. Computationally, the least-squares estimation (5) is highly non-convex and computationally unstable. The outperformance of **Borda count** demonstrates the efficacy of our method.

Applications to Chicago crime data. Chicago crime tensor dataset is an order-3 tensor with entries representing the log counts of crimes from 24 hours, 77 Chicago community areas, and 32 crime types ranging from January 1st, 2001 to December 11th, 2017. We apply our Borda Count method to Chicago crime dataset. 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 investigate the four clustered community areas obtained from our Borda Count algorithm. Figure 4a-b shows the four

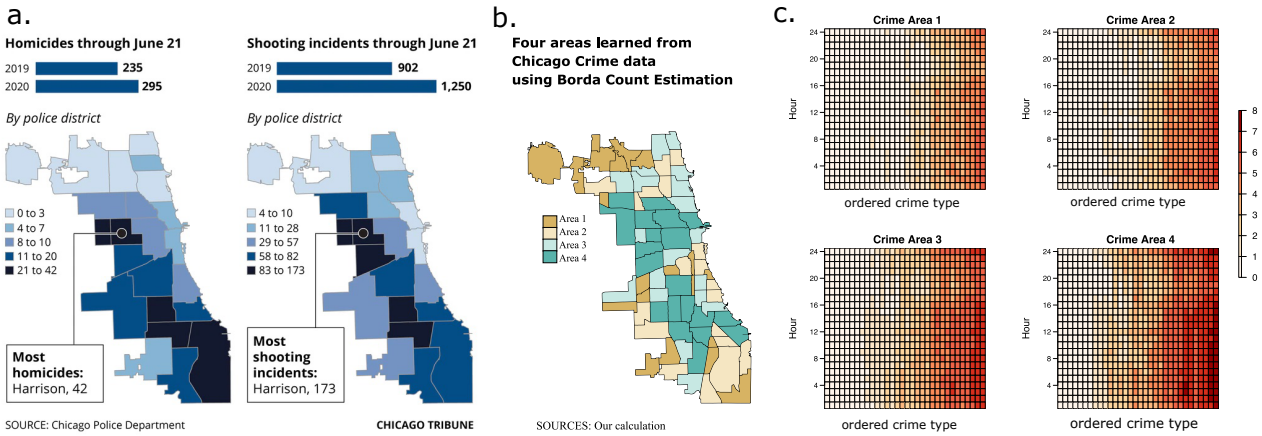


Figure 4: Chicago crime maps based on (a) *Chicago Tribune* article in 2020 [7] and (b) our estimation using Borda Count algorithm. (c) Averaged log counts of crimes according to crime types, hours, and the four areas estimated by our Borda count algorithm. For space consideration, the annotated crime types are described in the Appendix.

areas overlaid on a map of Chicago. We find that our clusters conform the actual locations even though our algorithm did not take any geographic information such as longitude or latitude. In addition, our clusters (Figure 4b) share similar geographical patterns with benchmark result (Figure 4a) based on *Chicago Tribune* article in 2020 [7]. Figure 4c reveals that the major difference among four areas is the crime rates: Area 4 has the highest crime rates, and the 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, 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 interpretable similarities and differences among the four community areas demonstrate the applicability of our method in real data.

6 Conclusion

We have presented a suite of statistical theory, estimation methods, and data applications for permuted smooth tensor models. We believe our results will be of interest to a very broad readership – from those interested in foundations of tensor methods to those in tensor data applications. Our method will help the practitioners efficiently analyze tensor datasets in various areas. Toward this end, the software package and all data used have been publicly released at CRAN.

References

[1] Krishnakumar Balasubramanian. Nonparametric modeling of higher-order interactions via hypergraphons. *Journal of Machine Learning Research*, 22:1–25, 2021.

[2] 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.

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

[4] Chao Gao and Zongming Ma. Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statistical Science*, 36(1):16–33, 2021.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] Ashwin Pananjady and Richard J Samworth. Isotonic regression with unknown permutations: Statistics, computation, and adaptation. *The Annals of Statistics, in press*, 2021.

[11] N Shah, Sivaraman Balakrishnan, and M Wainwright. Low permutation-rank matrices: Structural properties and noisy completion. *Journal of Machine Learning Research*, 20:1–43, 2019.

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

Appendix for "Smooth tensor estimation with unknown permutations"

The appendix includes extra numerical results and proofs to theorems.

A Extra numerical results

A.1 Details in synthetic data experiment

Simulation models. We describe the simulation set up in Section 5 in details. We simulate order-3 d-dimensional tensors based on the permuted smooth tensor model (1). The symmetric tensors are generated based on functions f in Table S1.

Model ID	$f(x, y, z)$	CP rank	Tucker rank
1	xyz	1	$(1, 1, 1)$
2	$(x + y + z)/3$	3	$(2, 2, 2)$
3	$(1 + \exp(-3x^2 + 3y^2 + 3z^2))^{-1}$	9	$(4, 4, 4)$
4	$\log(1 + \max(x, y, z))$	≥ 100	$\geq (50, 50, 50)$
5	$\exp\left(-\max(x, y, z) - \sqrt{x} - \sqrt{y} - \sqrt{z}\right)$	≥ 100	$\geq (50, 50, 50)$

Table S1: 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 (1).

The generative functions involve compositions of operations such as polynomial, logarithm, exponential, square roots, etc. Notice that considered functions cover a reasonable range of model complexities from low rank to high rank. Two types of noise are considered: Gaussian noise and Bernoulli noise. For the Gaussian model, we simulate continuous-valued tensors with i.i.d. noises drawn from $N(0, 0.5^2)$. For the Bernoulli model, we generate binary tensors $\mathcal{Y}$ using the success probability tensor $\Theta \circ \pi$. The permutation π is randomly chosen. For space consideration, only results for Models 1, 3, and 5 are presented in the paper. We first examine impacts of model complexity to estimation accuracy. We then compare Borda count estimation with alternative methods under a range of scenarios.

Impacts of the number of blocks, tensor dimension, and polynomial degree. 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 S1 demonstrates the trade-off in accuracy determined by the number of groups for each polynomial degree. The results confirm 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 approximations. These observations are consistent with our theoretical results that the optimal number of blocks and polynomial degree are $(k^*, \ell^*) = (\mathcal{O}(d^{3/7}), 2)$.

The second experiment investigates the impact of the tensor dimension d for various polynomial degrees. We vary the tensor dimension $d \in \{10, \ldots, 100\}$ and polynomial degree $\ell \in \{0, 1, 2, 3\}$ in each model configuration. We set optimal number of blocks as the one that gives the best accuracy. Figure S2 compares the estimation errors among different polynomial approximations. The result verifies that the degree-2 polynomial approximation performs the best under the sufficient tensor dimension, which is consistent with our theoretical results. We emphasize that this phenomenon is different from the matrix case where the degree-0 polynomial approximation gives the best results [4, 7].

Comparison with alternative methods. We compare our method (**Borda Count**) with several popular alternative methods.

- Spectral method (**Spectral**) [10] that performs universal singular value thresholding [2] on the unfolded tensor.
- Least-squares estimation (**LSE**) [4] which solves the optimization problem (5) with constant block approximation ($\ell = 0$) based on spectral k-means. We extend the matrix-based biclustering algorithm to higher-order tensors [6].
- Least-squares estimation (**BAL**) [1] which solves the optimization problem (5) with constant block approximation ($\ell = 0$). This tensor-based algorithm is only available for binary observations because it uses count-based statistics. Therefore, we only use this algorithm for the Bernoulli model.

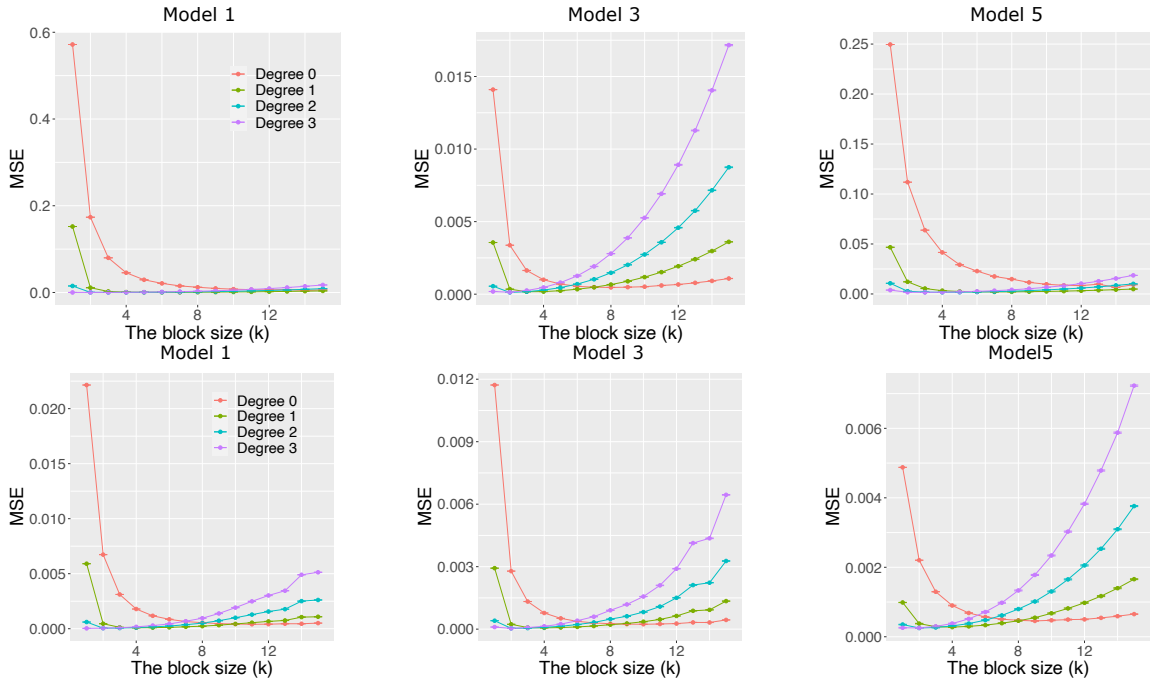


Figure S1: MSE versus the number of blocks based on different polynomial approximations. Columns 1-3 consider the Models 1, 3, and 5 respectively. Panel (a) is for continuous tensors, whereas (b) is for the binary tensors.

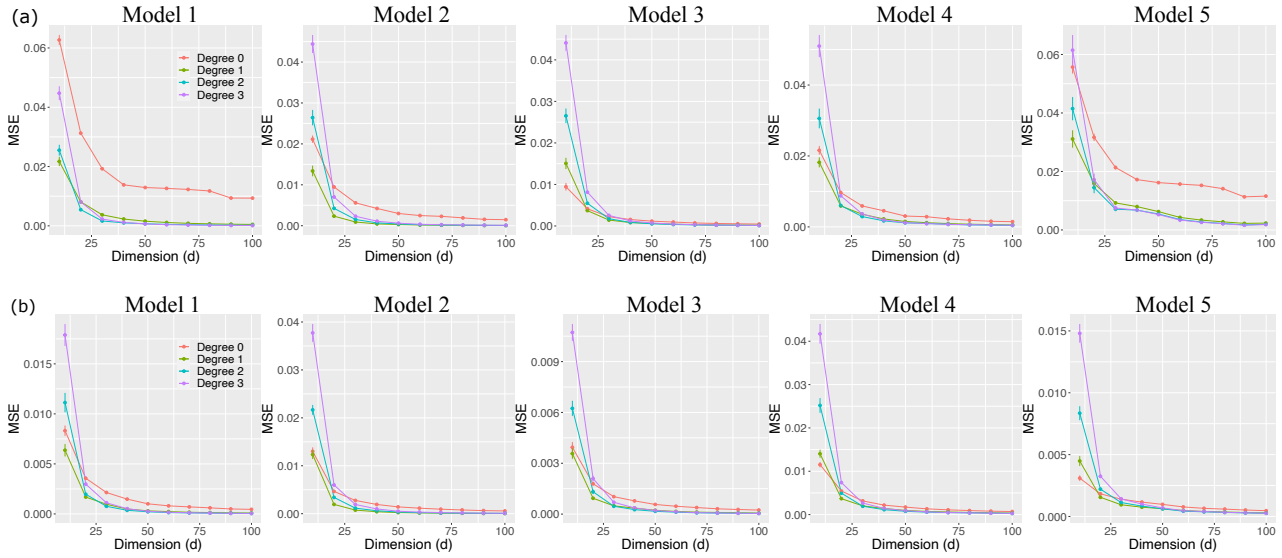


Figure S2: MSE versus the tensor dimension based on different polynomial approximations. Columns 1-5 consider the Models 1-5 in Table S1 respectively. Panel (a) is for continuous tensors, whereas (b) is for the binary tensors.

We choose degree-2 polynomial approximation as our theorems suggested, and vary tensor dimension $d \in \{10, \ldots, 100\}$ under each model configuration. For **Borda Count** and **LSE**, we choose the block numbers that achieve the best performance in the corresponding outputs. For **Spectral** method, we set the hyperparameter (singular-value threshold) that gives the best performance.

Figure S3 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,

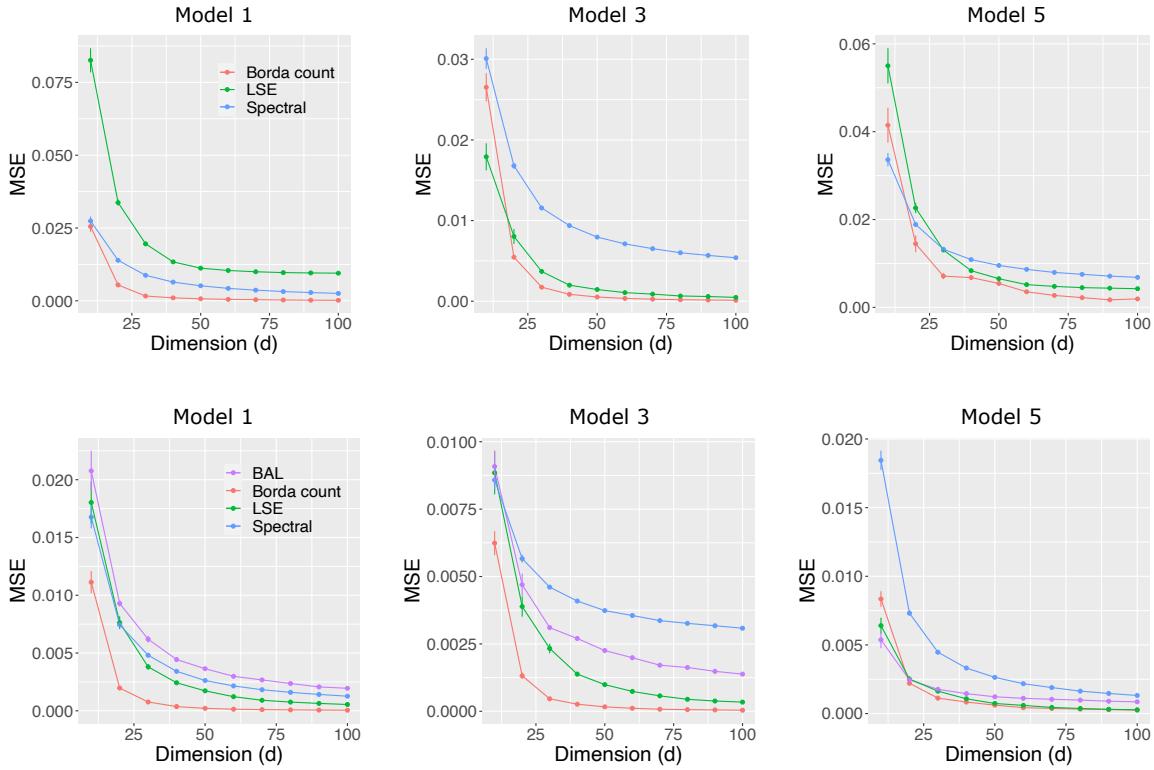


Figure S3: MSE versus the tensor dimension based on different estimation methods. Columns 1-3 consider the Models 1, 3, and 5 in Table S1 respectively. Panel (a) is for continuous tensors, whereas (b) is for the binary tensors.

our theorems have shown that constant block approximation results in sub-optimal rates compared to polynomial approximation. Computationally, the least-squares optimization (5) is highly non-convex and computationally unstable. Figure S4 displays true signal tensors of three models and corresponding observed tensors of dimension $d = 80$ with Gaussian noise. We use oracle permutation π to obtain the estimated signal tensor from the estimated permuted signal tensor $\hat{\Theta} \circ \hat{\pi}$ for the better visualization and comparisons. As shown in the figure, we see clearly that our method achieves the best signal recovery, thereby supporting the numerical results in Figure S3. The outperformance of **Borda count** demonstrates the efficacy of our method.

Investigation of non-symmetric tensors. Our models and techniques easily extend to non-symmetric tensors. We use non-symmetric functions to generate order-3 signal tensors. based on functions in Table S2.

Model ID	$f(x, y, z)$
1	$xy + z$
2	$x^2 + y + yz^2$
3	$x(1 + \exp(-3(x^2 + y^2 + z^2)))^{-1}$
4	$\log(1 + \max(x, y, z) + x^2 + yz)$
5	$\exp\left(-x - \sqrt{y} - z^3\right)$

Table S2: List of non-symmetric smooth functions in simulation.

We fix the tensor dimension $30 \times 40 \times 50$ and assume that the noise tensors are from Gaussian distribution. Similar to other simulations, we evaluate the accuracy of the estimation by MSE and report the summary statistics across $n_{\text{sim}} = 20$ replicates. The hyperparameters are chosen via cross-validation that give the best accuracy for each method. Table S3 summarizes the choice of hyperparameters. Table S4 compares the MSEs from repeated simulations based on different methods under Models 1-5. We find that Borda count estimation outperforms all alternative methods for non-symmetric tensors. The results demonstrate the applicability of our method to general tensors.

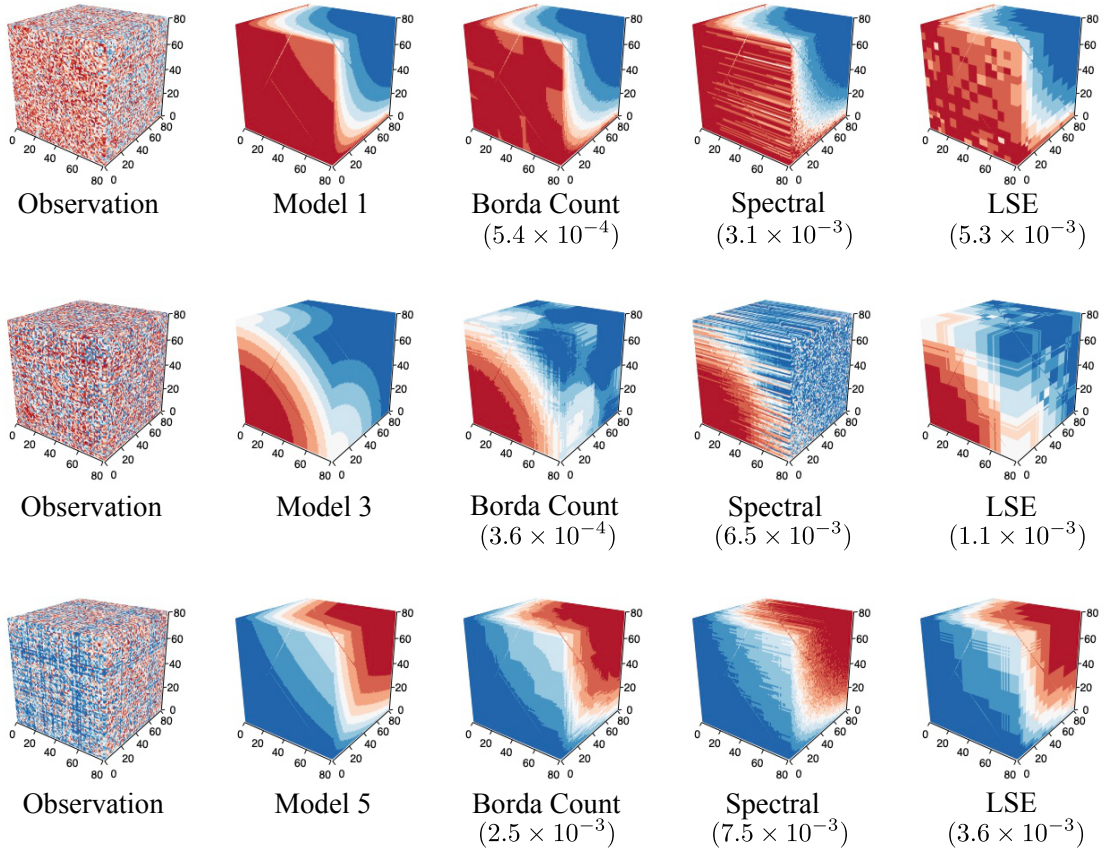


Figure S4: Performance comparison among different methods. The observed data tensors, true signal tensors, and estimated signal tensors are plotted for Models 1, 3 and 5 in Table S1 with fixed dimension $d = 80$. Numbers in parenthesis indicate the mean squared error.

Method	Model 1	Model 2	Model 3	Model 4	Model 5
Borda count	(2,1,2)	(1,2,2)	(1,3,3)	(2,1,2)	(1,4,4)
LSE	(6,2,3)	(8,5,8)	(6,9,6)	(9,5,6)	(7,9,3)
Spectral	(1,24)	(3,48)	(1,48)	(1,28)	(1,22)

Table S3: Hyperparameters for the methods under Models 1-5 in Table S2. For **Borda count** and **LSE** methods, the values in the table indicate the number of blocks. For **Spectral** method, the first value indicates the tensor unfolding mode, while the second one represents the singular value threshold.

Method	Model 1	Model 2	Model 3	Model 4	Model 5
Borda count	**0.57 (0.01)**	**0.51 (0.02)**	**0.87 (0.02)**	**1.02 (0.02)**	**2.56 (0.21)**
LSE	23.58 (0.03)	7.70 (0.04)	9.45 (0.05)	3.29 (0.05)	9.93 (0.03)
Spectral	10.76 (0.06)	10.64 (0.05)	6.27 (0.05)	10.90 (0.06)	5.24 (0.04)

Table S4: MSEs from 20 repeated simulations based on different methods. All numbers are displayed on the scales 10^{-3}. Standard errors are reported in parenthesis.

A.2 Details on Chicago crime data analysis

We compare the prediction performance based on constant block model and our permuted smooth tensor model. Notice that constant block model uses $\ell = 0$ approximation, whereas our permuted smooth tensor model uses $\ell = 2$ approximation. Table S5 shows the mean squared error over five runs of cross-validation, with 20% entries for testing and 80% for training. We find that the permuted smooth tensor model substantially outperforms the classical constant block models. We emphasize that our method does not necessarily assume the block structure. The comparison supports our premises

	Constant block model	Permuted smooth tensor model
MSE	0.399 (0.009)	0.283 (0.006)
Block number	$(7, 11, 10)$	$(6, 4, 10)$

Table S5: Performance comparison in Chicago data analysis. Reported MSEs are averaged over five runs of cross-validation, with 20% entries for testing and 80% for training, with standard errors in parentheses. Block number is set to achieve the best prediction performance.

that permuted smooth tensor model with polynomial approximation performs better than common constant block models in this application.

We also investigate the ten groups of crime types clustered by our method. Table S6 shows that the clustering captures the similar type of crimes. For example, group 2 consists of misdemeanors such as public indecency, non-criminal, and concealed carry license violation, while group 6 represents sex-related offenses such as prostitution, sex offense, and crime sexual assault.

GROUP	I	II	III
CRIME TYPE	RITUALISM, HUMAN TRAFFICKING, OTHER NARCOTIC VIOLATION	PUBLIC INDECENCY, NON-CRIMINAL, CONCEALED CARRY LICENSE VIOLATION	OBSCENITY, STALKING, INTIMIDATION
GROUP	IV	V	VI
CRIME TYPE	KIDNAPPING, GAMBLING, HOMICIDE	LIQUOR LAW VIOLATION, ARSON, INTERFERENCE WITH PUBLIC OFFICER	PROSTITUTION, SEX OFFENSE, CRIM SEXUAL ASSAULT
GROUP	VII	VIII	VIIII
CRIME TYPE	OTHER OFFENSE, CRIMINAL DAMAGE, BATTERY, THEFT, BURGLARY	CRIMINAL TRESPASS, ROBBERY, DECEPTIVE PRACTICE	NARCOTICS, ASSAULT, MOTOR VEHICLE THEFT
GROUP	X		
CRIME TYPE	PUBLIC PEACE VIOLATION, WEAPONS VIOLATION, OFFENSE INVOLVING CHILDREN		

Table S6: Groups of crime types learned based on the Borda count estimation.

B Proofs

B.1 Proof of Theorem 1

Smoothness of the function f in (1) plays an important role in the block-wise polynomial approximation. The following lemma explains the role of smoothness in the approximation.

Lemma 1 (Block-wise polynomial tensor approximation)**.** *Suppose $\Theta \in \mathcal{P}(\alpha, L)$. Then, for every block number $k \leq d$, and degree $\ell \in \mathbb{N}_{\geq 0}$, we have the approximation error*

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

Lemma 1 implies that we can always find a block-wise polynomial tensor close to the signal tensor generated from α-Hölder smooth function f. The approximation error decays with block number k and degree $\min(\alpha, \ell + 1)$.

Proof of Lemma 1. Recall that we denote $\mathcal{E}_k$ as the m-way partition

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

where $z \colon [d] \to [k]$ is the canonical clustering function such that $z(i) = \lceil ki/d \rceil$, for all $i \in [d]$, and we use the shorthand $\bigtimes_{a=1}^{m}$ to denote the Cartesian product of m sets. For a given partition $\bigtimes_{a=1}^{m} z^{-1}(j_a) \in \mathcal{E}_k$, fix any index $(i_1^0, \ldots, i_m^0) \in \bigtimes_{a=1}^{m} z^{-1}(j_a)$. Then, we have

$$\|(i_1, \ldots, i_m) - (i_1^0, \ldots, i_m^0)\|_\infty \leq \frac{d}{k}, \tag{1}$$

for all $(i_1, \ldots, i_m) \in \bigtimes_{a=1}^{m} z^{-1}(j_a)$. We define the block-wise degree-ℓ polynomial tensor $\mathcal{B}$ based on the partition $\mathcal{E}_k$ as

$$\mathcal{B}(i_1, \ldots, i_m) = \mathrm{Poly}^{j_1,\ldots,j_m}_{\min(\lfloor \alpha \rfloor, \ell)} \left(\frac{i_1 - i_1^0}{d}, \ldots \frac{i_m - i_m^0}{d} \right), \quad \text{for all } (i_1, \ldots, i_m) \in \bigtimes_{a=1}^{m} z^{-1}(j_a),$$

where $\text{Poly}^{j_1,\ldots,j_m}_{\min(\lfloor\alpha\rfloor,\ell)}$ denotes a degree-ℓ polynomial function satisfying

$$\left|f\left(\frac{i_1}{d},\ldots,\frac{i_m}{d}\right)-\text{Poly}^{j_1,\ldots,j_m}_{\min(\lfloor\alpha\rfloor,\ell)}\left(\frac{i_1-i_1^0}{d},\ldots\frac{i_m-i_m^0}{d}\right)\right|\le L\left\|\left(\frac{i_1-i_1^0}{d},\ldots,\frac{i_m-i_m^0}{d}\right)\right\|_\infty^{\min(\alpha,\ell+1)}, \quad (2)$$

for all $(i_1,\ldots,i_m)\in\bigtimes_{a=1}^m z^{-1}(j_a)$. Notice that we can always find such polynomial function by α-Hölder smoothness of the generative function f. Based on the construction of block-wise degree-ℓ polynomial tensor $\mathcal{B}$, we have

$$\begin{aligned}
&\frac{1}{d^m}\|\Theta-\mathcal{B}\|_F^2\\
&=\frac{1}{d^m}\sum_{(i_1,\ldots,i_m)\in[d]^m}|\Theta(i_1,\ldots,i_m)-\mathcal{B}(i_1,\ldots,i_m)|^2\\
&=\frac{1}{d^m}\sum_{(j_1,\ldots,j_m)\in[k]^m}\sum_{(i_1,\ldots,i_m)\in\bigtimes_{a=1}^m z^{-1}(j_a)}\left|f\left(\frac{i_1}{d},\ldots,\frac{i_m}{d}\right)-\text{Poly}^{j_1,\ldots,j_m}_{\min(\lfloor\alpha\rfloor,\ell)}\left(\frac{i_1-i_1^0}{d},\ldots,\frac{i_m-i_m^0}{d}\right)\right|^2\\
&\lesssim\frac{L^2}{d^m}\sum_{(j_1,\ldots,j_m)\in[k]^m}\sum_{(i_1,\ldots,i_m)\in\bigtimes_{a=1}^m z^{-1}(j_a)}\left\|\left(\frac{i_1-i_1^0}{d},\ldots,\frac{i_m-i_m^0}{d}\right)\right\|_\infty^{2\min(\alpha,\ell+1)}\\
&\le\frac{L^2}{k^{2\min(\alpha,\ell+1)}},
\end{aligned}$$

where the first inequality uses (2) and the second inequality is from (1). □

Proof of Theorem 1. By Lemma 1, there exists a block-wise polynomial tensor $\mathcal{B}\in\mathscr{B}(k,\ell)$ such that

$$\|\mathcal{B}-\Theta\|_F^2\lesssim\frac{L^2d^m}{k^{2\min(\alpha,\ell)}}. \quad (3)$$

By the triangle inequality,

$$\|\hat\Theta^{\text{LSE}}\circ\hat\pi^{\text{LSE}}-\Theta\circ\pi\|_F^2\le 2\|\hat\Theta^{\text{LSE}}\circ\hat\pi^{\text{LSE}}-\mathcal{B}\circ\pi\|_F^2+2\underbrace{\|\mathcal{B}\circ\pi-\Theta\circ\pi\|_F^2}_{\text{Lemma 1}}. \quad (4)$$

Therefore, it suffices to bound $\|\hat\Theta^{\text{LSE}}\circ\hat\pi^{\text{LSE}}-\mathcal{B}\circ\pi\|_F^2$. By the global optimality of least-square estimator, we have

$$\begin{aligned}
\|\hat\Theta^{\text{LSE}}\circ\hat\pi^{\text{LSE}}-\mathcal{B}\circ\pi\|_F&\le\left\langle\frac{\hat\Theta^{\text{LSE}}\circ\hat\pi^{\text{LSE}}-\mathcal{B}\circ\pi}{\|\hat\Theta^{\text{LSE}}\circ\hat\pi^{\text{LSE}}-\mathcal{B}\circ\pi\|_F},\ \mathcal{E}+(\Theta\circ\pi-\mathcal{B}\circ\pi)\right\rangle\\
&\le\sup_{\pi,\pi'\colon[d]\to[d]}\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)}\left\langle\frac{\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi\|_F},\mathcal{E}\right\rangle+\underbrace{\|\mathcal{B}\circ\pi-\Theta\circ\pi\|_F}_{\text{Lemma 1}}.
\end{aligned}$$

Now we bound inner product term. For fixed π,π', let $\boldsymbol{P}$ and $\boldsymbol{P}'$ be permutation matrices corresponding to permutations π and π' respectively. We express vectorized block-wise degree-ℓ polynomial tensors, $\text{vec}(\mathcal{B})$ and $\text{vec}(\mathcal{B}')$, by discrete polynomial functions. Specifically, denote $\text{vec}(\mathcal{B})=\boldsymbol{X}\boldsymbol{\beta}$ and $\text{vec}(\mathcal{B}')=\boldsymbol{X}\boldsymbol{\beta}'$, where $\boldsymbol{X}\in\mathbb{R}^{d^m\times k^m(k+m)^\ell}$ is a design matrix consisting of m-multivaraite degree-ℓ polynomial basis over grid design $(1/d,\ldots,d/d)$, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'\in\mathbb{R}^{k^m(k+m)^\ell}$ are corresponding coefficient vectors. Notice that the number of coefficients for m-multivariate polynomial of degree-ℓ is $\binom{\ell+m}{\ell}$. We choose to use $(k+m)^\ell$ coefficients for each block for notational simplicity. Therefore, we rewrite the inner product

$$\begin{aligned}
\left\langle\frac{\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi\|_F},\mathcal{E}\right\rangle&=\left\langle\frac{(\boldsymbol{P}')^{\otimes m}\text{vec}(\mathcal{B}')-(\boldsymbol{P})^{\otimes m}\text{vec}(\mathcal{B})}{\|(\boldsymbol{P}')^{\otimes m}\text{vec}(\mathcal{B}')-(\boldsymbol{P})^{\otimes m}\text{vec}(\mathcal{B})\|_F},\mathcal{E}\right\rangle\\
&=\left\langle\frac{(\boldsymbol{P}')^{\otimes m}\boldsymbol{X}\boldsymbol{\beta}'-(\boldsymbol{P})^{\otimes m}\boldsymbol{X}\boldsymbol{\beta}}{\|(\boldsymbol{P}')^{\otimes m}\boldsymbol{X}\boldsymbol{\beta}'-(\boldsymbol{P})^{\otimes m}\boldsymbol{X}\boldsymbol{\beta}\|_F},\mathcal{E}\right\rangle\\
&=\left\langle\frac{\boldsymbol{A}\boldsymbol{c}}{\|\boldsymbol{A}\boldsymbol{c}\|_F},\mathcal{E}\right\rangle,
\end{aligned}$$

where we define $\boldsymbol{A} := (\boldsymbol{P}' \quad -\boldsymbol{P}) \begin{pmatrix} \boldsymbol{X} & 0 \\ 0 & \boldsymbol{X} \end{pmatrix} \in \mathbb{R}^{d^m \times 2k^m(k+m)^\ell}$ and $\boldsymbol{c} := \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{\beta} \end{pmatrix} \in \mathbb{R}^{2k^m(k+m)^\ell}$. By Lemma 5, we have

$$\sup_{\mathcal{B},\mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi}{\|\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi\|_F}, \mathcal{E} \right\rangle \leq \sup_{\boldsymbol{c} \in \mathbb{R}^{2k^m(\ell+m)^\ell}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2}, e \right\rangle, \tag{5}$$

where $e \in \mathbb{R}^{2k^m(k+m)^\ell}$ is a vector consisting of i.i.d. sub-Gaussian entries with variance proxy σ^2. By the union bound of Gaussian maxima over countable set $\{\pi, \pi' \colon [d] \to [d]\}$, we obtain

$$\begin{aligned} &\mathbb{P}\left(\sup_{\pi,\pi' \colon [d]\to[d]} \sup_{\mathcal{B},\mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi}{\|\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi\|_F}, \mathcal{E} \right\rangle \geq t \right) \\ &\leq \sum_{\pi,\pi' \in [d]^d} \mathbb{P}\left(\sup_{\mathcal{B},\mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi}{\|\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi\|_F}, \mathcal{E} \right\rangle \geq t \right) \\ &\leq d^d \mathbb{P}\left(\sup_{\boldsymbol{c} \in \mathbb{R}^{2k^m(\ell+m)^\ell}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2}, e \right\rangle \geq t \right) \\ &\leq \exp\left(-\frac{t^2}{8\sigma^2} + k^m(\ell+m)^\ell \log 6 + d \log d \right), \end{aligned} \tag{6}$$

where the second inequality is from (5) and the last inequality is from Lemma 6. Setting $t = C\sigma\sqrt{k^m(\ell+m)^\ell + d\log d}$ in (7) for sufficiently large $C > 0$ gives

$$\sup_{\pi,\pi' \colon [d]\to[d]} \sup_{\mathcal{B},\mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi}{\|\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi\|_F}, \mathcal{E} \right\rangle \lesssim \sigma\sqrt{k^m(\ell+m)^\ell + d\log d}, \tag{7}$$

with high probability.
Combining the inequalities (3), (4) and (7) yields the desired conclusion

$$\|\hat{\Theta}^{\mathrm{LSE}} \circ \hat{\pi}^{\mathrm{LSE}} - \Theta \circ \pi\|_F^2 \lesssim \sigma^2 \left(k^m(\ell+m)^\ell + d\log d \right) + \frac{L^2 d^m}{k^{2\min(\alpha,\ell)}}. \tag{8}$$

Finally, optimizing (8) with respect to (k, l) gives that

$$(8) \lesssim \begin{cases} L^2 \left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}} d^{-\frac{2m\alpha}{m+2\alpha}}, & \text{when } \alpha < m(m-1)/2, \\ \sigma^2 d^{-(m-1)} \log d, & \text{when } \alpha \geq m(m-1)/2, \end{cases}$$

under the choice

$$\ell^* = \min(\lfloor \alpha \rfloor, (m-2)(m+1)/2), \quad k^* = \left\lceil \left(d^m L^2/\sigma^2\right)^{\frac{1}{m+2\min(\alpha,\ell^*+1)}} \right\rceil.$$

□

B.2 Proof of Theorem 2

Proof of Theorem 2. By the definition of the tensor space, we seek the minimax rate ε^2 in the following expression

$$\inf_{(\hat{\Theta},\hat{\pi})} \sup_{\Theta \in \mathcal{P}(\alpha,L)} \sup_{\pi \in \Pi(d,d)} \mathbb{P}\left(\frac{1}{d^m} \|\Theta \circ \pi - \hat{\Theta} \circ \hat{\pi}\|_F^2 \geq \varepsilon^2 \right).$$

On one hand, if we fix a permutation $\pi \in \Pi(d,d)$, the problem can be viewed as a classical m-dimensional α-smooth nonparametric regression with d^m sample points. The minimax lower bound is known to be $\varepsilon^2 = L^2 \left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}} d^{-\frac{2m\alpha}{m+2\alpha}}$. On the other hand, if we fix $\Theta \in \mathcal{P}(\alpha, L)$, the problem become a new type of convergence rate due to the unknown permutation. We refer to the resulting error as the permutation rate, and we will prove that $\varepsilon^2 = \sigma^2 d^{-(m-1)} \log d$. Since our target is the sum of the two rates, it suffice to prove the two different rates separately. In the following arguments, we will proceed by this strategy.

Nonparametric rate. The nonparametric rate for α-smooth function is readily available in the literature; see [5, Section 3.2] and [9, Section 2]. We state the results here for self-completeness.

Lemma 2 (Minimax rate for α-smooth function estimation)**.** *Consider a sample of N data points, $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_N, Y_N)$, where $\boldsymbol{x}_n = (\frac{i_1}{d}, \ldots, \frac{i_m}{d}) \in [0,1]^m$ is the m-dimensional predictor and $Y_n \in \mathbb{R}$ is the scalar response. Consider the observation model*

$$Y_n = f(\boldsymbol{x}_n) + \varepsilon_n, \quad \textit{with } \varepsilon_n \sim \textit{i.i.d. } N(0,1), \quad \textit{for all } n \in [N].$$

Assume f is in the α-Holder smooth function class, denoted by $\mathcal{H}(\alpha, L)$. Then,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{H}(\alpha, L)} \mathbb{P}\left(\|f - \hat{f}\|_2 \geq \sigma^{\frac{4\alpha}{m+2\alpha}} L^{\frac{2m}{m+2\alpha}} N^{-\frac{2\alpha}{m+2\alpha}} \right) \geq 0.9.$$

Our desired nonparametric rate readily follows from Lemma 2 by taking sample size $N = d^m$ and function norm $\|f - \hat{f}\|_2 = \frac{1}{d^m} \|\Theta - \hat{\Theta}\|_F^2$. In summary, for a given permutation $\pi \in \Pi(d, d)$, we have

$$\inf_{\hat{\Theta}} \sup_{\Theta \in \mathcal{P}(\alpha, L)} \mathbb{P}\left(\frac{1}{d^m} \|\hat{\Theta} \circ \pi - \Theta \circ \pi\|_F^2 \geq L^2 \left(\frac{\sigma}{L} \right)^{\frac{4\alpha}{m+2\alpha}} d^{-\frac{2m\alpha}{m+2\alpha}} \right) \geq 0.9. \tag{9}$$

Permutation rate. Since nonparametric rate dominates permutation rate when $\alpha \leq 1$, it is sufficient to prove the permutation rate lower bound for $\alpha \geq 1$. We first show the minimax permutation rate for k-block degree-0 tensor family $\mathcal{B}(k, 0)$, and then construct a smooth $f \in \mathcal{H}(\alpha, L)$ to mimic the constant block tensors.
Let $\Pi(d, k)$ denote the collection of all possible onto mappings from $[d]$ to $[k]$. Lemma 3 shows the permutation rate over k-block degree-0 tensor family $\mathcal{B}(k, 0)$ is $\sigma^2 d^{-(m-1)} \log k$.

Lemma 3 (Permutation error for tensor block model)**.** *Consider the problem of estimating d-dimensional, block-k signal tensors from sub-Gaussian tensor block models. For every given integer $k \in [d]$, there exists a core tensor $\mathcal{S} \in \mathbb{R}^{k \times \cdots \times k}$ satisfying*

$$\inf_{\hat{\Theta}} \sup_{z \in \Pi(d,k)} \mathbb{P}\left\{ \frac{1}{d^m} \sum_{(i_1, \ldots, i_m) \in [d]^m} \left[\hat{\Theta}(i_1, \ldots, i_m) - \mathcal{S}\left(z(i_1), \ldots, z(i_m)\right) \right]^2 \gtrsim \frac{\sigma^2 \log k}{d^{m-1}} \right\} \geq 0.9. \tag{10}$$

The proof of Lemma 3 is constructive and deferred to Section B.4. We fix a core tensor $\mathcal{S} \in \mathbb{R}^{k \times \cdots \times k}$ satisfying (10), and use it to construct the smooth tensors.
Now we construct a function $f \in \mathcal{H}(\alpha, L)$ that mimics the core tensor $\mathcal{S}$ in block tensor family $\mathcal{B}(k, 0)$. Define $k = d^\delta$ for some $\delta \in (0, 1)$, which will be specified later. Consider a smooth function $K(x)$ that is infinitely differentiable,

$$K(x) = C_k \exp\left(-\frac{1}{1 - 64x^2} \right) \mathbb{1}\left\{ |x| < \frac{1}{8} \right\},$$

where $C_k > 0$ satisfies $\int K(x) dx = 1$. Then, we define a smooth cutoff function as

$$\psi(x) = \int_{-3/8}^{3/8} K(x - y) dy.$$

The smooth cutoff function has support $[-1/2, 1/2]$ and takes value 1 on the interval $[-1/4, 1/4]$. For a given core tensor $\mathcal{S}$ satisfying Lemma 3, we define α-smooth function

$$f(x_1, \ldots, x_m) = \sum_{(a_1, \ldots, a_m) \in [k]^m} \left(\mathcal{S}(a_1, \ldots, a_m) - \frac{1}{2} \right) \prod \psi\left(k x_1 - a_1 + \frac{1}{2} \right) + \frac{1}{2}. \tag{11}$$

One can verify that $f \in \mathcal{H}(\alpha, L)$ as long as we choose sufficiently small δ depending on α and L. Notice that for any $(a_1, \ldots, a_m) \in [k]^m$,

$$f(x_1, \ldots, x_m) = \mathcal{S}(a_1, \ldots, a_m), \quad \text{if } (x_1, \ldots, x_m) \in \bigtimes_{i=1}^m \left[\frac{a_i - 3/4}{k}, \frac{a_i - 1/4}{k} \right].$$

From this observation, we define a sub-domain $I \subset [d]$ such that

$$I = \left(\bigcup_{a=1}^k \left[\frac{d(a - 3/4)}{k}, \frac{d(a - 1/4)}{k} \right] \right) \bigcap [d].$$

Then, $\{f(i_1/d,\dots,i_m/d)\colon i_1,\dots,i_m\in I\}$ forms the block structure with the core tensor $\mathcal{S}\in\mathbb{R}^{k\times\cdots k}$. Define a subset of permutations $\Pi'(d,d)=\{\pi\in\Pi(d,d)\colon \sigma(i)=i \text{ for } i\in[d]\setminus I\}\subset\Pi(d,d)$, which collects permutations on I while fixing indices on $[d]\setminus I$. Then we have

$$\begin{aligned}
&\inf_{(\hat\Theta,\hat\pi)}\sup_{\pi\in\Pi(d,d)}\mathbb{P}\left(\frac{1}{d^m}\|\hat\Theta\circ\hat\pi-\Theta\circ\pi\|_F^2\ge\varepsilon^2\right)\\
&\overset{(1)}{=}\inf_{\hat\Theta}\sup_{\pi\in\Pi(d,d)}\mathbb{P}\left(\frac{1}{d^m}\|\hat\Theta-\Theta\circ\pi\|_F^2\ge\varepsilon^2\right)\\
&\overset{(2)}{\ge}\inf_{\hat\Theta}\sup_{\pi\in\Pi'(d,d)}\mathbb{P}\left(\frac{1}{d^m}\sum_{(i_1,\dots,i_m)\in[d]^m}\Big[\hat\Theta(i_1,\dots,i_m)-f(\pi(i_1)/d,\dots,\pi(i_m)/d)\Big]^2\ge\varepsilon^2\right)\\
&\ge\inf_{\hat\Theta}\sup_{\pi\in\Pi'(d,d)}\mathbb{P}\left(\frac{1}{d^m}\sum_{(i_1,\dots,i_m)\in I^m}\Big[\hat\Theta(i_1,\dots,i_m)-f(\pi(i_1)/d,\dots,\pi(i_m)/d)\Big]^2\ge\varepsilon^2\right),
\end{aligned}\tag{12}$$

where (1) absorbs the estimate $\hat\pi$ into the estimate $\hat\Theta$, and (2) uses the constructed function (11) and the permutation collections $\Pi'(d,d)$. For any $\pi\in\Pi'(d,d)$, define clustering function $z\colon I\to[k]$ such that $z(i)=\lceil k\pi(i)/d\rceil$ for all $i\in I$. Then, we have

$$f\left(\frac{\pi(i_1)}{d},\dots,\frac{\pi(i_m)}{d}\right)=\mathcal{S}(z(i_1),\dots,z(i_m)),\quad\text{for all } i_1,\dots,i_m\in I.\tag{13}$$

Finally, combining (12), (13), and Lemma 3 yields

$$\inf_{(\hat\Theta,\hat\pi)}\sup_{\pi\in\Pi(d,d)}\mathbb{P}\left(\frac{1}{d^m}\|\hat\Theta\circ\hat\pi-\Theta\circ\pi\|_F^2\gtrsim\frac{\sigma^2\log d}{d^{m-1}}\right)\ge 0.9,\tag{14}$$

where k is replaced by n^δ.

Combining two rates. Now, we combine (9) and (14) to get the desired lower bound. For any Θ generated as in (1) with $f\in\mathcal{H}(\alpha,L)$, by union bound, we have

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{d^m}\|\hat\Theta-\Theta\|_F^2\gtrsim L^2\left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}}d^{-\frac{2m\alpha}{m+2\alpha}}+\frac{\sigma^2\log d}{d^{m-1}}\right\}\\
\ge\;&\mathbb{P}\left\{\frac{1}{d^m}\|\hat\Theta-\Theta\|_F^2\gtrsim L^2\left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}}d^{-\frac{2m\alpha}{m+2\alpha}}\right\}+\mathbb{P}\left\{\frac{1}{d^m}\|\hat\Theta-\Theta\|_F^2\gtrsim\frac{\sigma^2\log d}{d^{m-1}}\right\}-1.
\end{aligned}$$

Taking sup on both sides with the property

$$\sup_{\substack{\Theta\in\mathcal{P}(\alpha,L)\\ \pi\in\Pi(d,d)}}(f(\pi)+g(\Theta))=\sup_{\pi\in\Pi(d,d)}f(\pi)+\sup_{\Theta\in\mathcal{P}(\alpha,L)}g(\Theta)$$

yields the desired rate (2). □

B.3 Proof of Theorem 3

The β-monotonicity condition allows us to efficiently estimate the permutation π. Before presenting the theoretical guarantees, we provide the intuition here. The exponent β measures the difficulty for estimating the permutation π. Consider the noisy observation $\mathcal{Y}$ from model (**??**). We define the empirical score function $\tau\colon[d]\to\mathbb{R}$ as

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

The permuted score function $\tau\circ\pi^{-1}$ reduces to the function g in (2) under the noiseless setting. Therefore, a good estimate $\hat\pi$ should make 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 a larger β guarantees a faster consistency rate of $\hat\pi$. A large β implies large gaps of $|g(i)-g(j)|$ for $i\neq j$. Therefore, we obtain similar orderings of $\{\tau(i)\}_{i=1}^d$ before and after the addition of noise. This intuition is well represented by the following lemma.

Lemma 4 (Permutation error)**.** *Consider the permuted smooth tensor model with $f\in\mathcal{M}(\beta)$. Let $\hat\pi$ be the permutation such that the permuted empirical score function $\tau\circ\hat\pi^{-1}$ is monotonically increasing. Then, with high probability,*

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

Proof of Lemma 4. Without loss of generality, assume that π is the identity permutation. Notice that $g(i) - \tau(i)$ is the sample average of roughly (excluding repetitions from symmetricity) d^{m-1} independent mean-zero sub-Gaussian random variables with the variance proxy σ. Based on the independence of sub-Gaussian random variables, we have

$$|g(i) - \tau(i)| < 2\sigma d^{-(m-1)/2}\sqrt{\log d}, \tag{15}$$

with probability $1 - \frac{2}{d^2}$ for all $i \in [d]$.
By the weakly β-monotonicity of the function g, we have

$$g(1) \pm \delta \le g(2) \pm \delta \le \cdots \le g(d-1) \pm \delta \le g(d) \pm \delta, \tag{16}$$

where $\delta \lesssim d^{-(m-1)/2}$ is the small tolerance. The estimated permutation $\hat{\pi}$ is defined for which

$$\tau(\hat{\pi}^{-1}(1)) \le \tau(\hat{\pi}^{-1}(2)) \le \cdots \le \tau(\hat{\pi}^{-1}(d-1)) \le \tau(\hat{\pi}^{-1}(d)). \tag{17}$$

For any given index i, we examine the error $|i - \hat{\pi}(i)|$. By (16) and (17), we have

$$i = \underbrace{|\{j \colon g(j) \le g(i)\}|}_{=:\mathrm{I}}, \quad \text{and} \quad \hat{\pi}(i) = \underbrace{|\{j \colon \tau(j) \le \tau(i)\}|}_{=:\mathrm{II}},$$

where $|\cdot|$ denotes the cardinality of the set. We claim that the sets I and II differ only in at most $d^{(m-1)\beta/2}$ elements. To prove this, we partition the indices in $[d]$ in two cases.

1. Long-distance indices in $\{j \colon |j - i| \ge C\left(\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta}\}$ for some sufficient large constant $C > 0$. In this case, the ordering of (i, j) remains the same in (16) and (17), i.e.,

$$g(i) < g(j) \Longleftrightarrow \tau(i) < \tau(j). \tag{18}$$

We only prove the right side direction in (18) here. The other direction can be similarly proved. Suppose that $g(i) < g(j)$. Then we have

$$\begin{aligned}
\tau(j) - \tau(i) &\ge -|g(j) - \tau(j)| - |g(i) - \tau(i)| + g(j) - g(i) \\
&> -4\sigma d^{(m-1)/2}\sqrt{\log d} + g(j) - g(i) \\
&\ge 0,
\end{aligned}$$

where the second inequality is from (15) with probability at least $(1 - 2/d^2)^d$ and the last inequality uses weakly β-monotonousity of $g(\cdot)$, the tolerance condition $\delta \lesssim d^{-(m-1)/2}$, and the assumption $|j - i| \ge C\left(\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta}$. Therefore we show that $g(i) < g(j)$ implies $\tau(i) < \tau(j)$. In this case, we conclude that none of long-distance indices belongs to IΔII.

2. Short-distance indices in $\{j \colon |j - i| < \left(\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta}\}$. In this case, (16) and (17) may yield different ordering of (i, j).

Combining the above two cases gives that

$$\left\{j \colon \frac{1}{d}|j - i| \le \left(4\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta}\right\} \supset \mathrm{I}\Delta\mathrm{II}.$$

Finally, we have

$$\mathrm{Loss}(\pi, \hat{\pi}) := \frac{1}{d}\max_{i \in [d]}|\pi(i) - \hat{\pi}(i)| \le \frac{1}{d}\mathrm{I}\Delta\mathrm{II} \le \left(4\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta},$$

with high probability. □

Proof of Theorem 3. By Lemma 1, there exists a block-wise polynomial tensor $\mathcal{B} \in \mathscr{B}(k, \ell)$ satisfying (3). By the triangle inequality, we decompose estimation error into three terms,

$$\begin{aligned}
&\|\hat{\Theta}^{\mathrm{BC}} \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F \\
&\le \|\hat{\Theta}^{\mathrm{BC}} \circ \hat{\pi}^{\mathrm{BC}} - \mathcal{B} \circ \hat{\pi}^{\mathrm{BC}}\|_F + \|\mathcal{B} \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \hat{\pi}^{\mathrm{BC}}\|_F + \|\Theta \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F \\
&= \underbrace{\|\Theta \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F}_{\text{Permutation error}} + \underbrace{\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F}_{\text{Nonparametric error}} + \underbrace{\|\mathcal{B} - \Theta\|_F}_{\text{Lemma 1}}.
\end{aligned} \tag{19}$$

Therefore, it suffices to bound two terms $\|\Theta \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F$ and $\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F$ separately.

Permutation error. For any $(i_1, \dots, i_m) \in [d]^m$, we have

$$\begin{aligned}
&\left|\Theta\left(\hat{\pi}^{\mathrm{BC}}(i_1), \dots, \hat{\pi}^{\mathrm{BC}}(i_m)\right) - \Theta\left(\pi(i_1), \dots, \pi(i_m)\right)\right| \\
&\le \left\| \left(\frac{\hat{\pi}^{\mathrm{BC}}(i_1)}{d}, \dots, \frac{\hat{\pi}^{\mathrm{BC}}(i_m)}{d}\right) - \left(\frac{\pi(i_1)}{d}, \dots, \frac{\pi(i_m)}{d}\right) \right\|_\infty^{\min(\alpha,1)} \\
&\le \left[\frac{1}{d} \max_{i\in[d]} \left|\hat{\pi}^{\mathrm{BC}}(i) - \pi(i)\right|\right]^{\min(\alpha,1)} \\
&\lesssim \left(\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta\min(\alpha,1)},
\end{aligned}$$

where the first inequality is from the α-Hölder smoothness of Θ, and the last inequality is from Lemma 4. Therefore, we obtain the upper bound of the permutation error

$$\frac{1}{d^m}\|\Theta \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F^2 \lesssim \left(\sigma^2 \frac{\log d}{d^{m-1}}\right)^{\beta\min(\alpha,1)}. \tag{20}$$

Nonparametric error. Recall that Borda count estimation is defined by $\hat{\Theta}^{\mathrm{BC}} := \arg\min_{\Theta\in\mathscr{B}(k,\ell)} \|\tilde{\mathcal{Y}} - \Theta\|_F^2$, where $\tilde{\mathcal{Y}} = \mathcal{Y} \circ (\hat{\pi}^{\mathrm{BC}})^{-1}$. By the optimality of least-square estimator, we have

$$\begin{aligned}
\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F &\le \left\langle \frac{\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}}{\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F},\ \mathcal{Y} \circ \pi \circ (\hat{\pi}^{\mathrm{BC}})^{-1} - \mathcal{B} \right\rangle \\
&\equiv \left\langle \frac{\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}}{\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F}, \mathcal{E} + \left(\Theta \circ \pi \circ (\hat{\pi}^{\mathrm{BC}})^{-1} - \mathcal{B}\right) \right\rangle \\
&\le \sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B}' - \mathcal{B}\|_F}, \mathcal{E} \right\rangle + \|\Theta \circ \pi - \mathcal{B} \circ \hat{\pi}^{\mathrm{BC}}\|_F \\
&\le \sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B}' - \mathcal{B}\|_F}, \mathcal{E} \right\rangle + \underbrace{\|\Theta \circ \pi - \Theta \circ \hat{\pi}^{\mathrm{BC}}\|_F}_{\text{Permutation error (20)}} + \underbrace{\|\Theta - \mathcal{B}\|_F}_{\text{Lemma 1}}
\end{aligned}$$

Now we bound the inner product term. By the same argument in the proof of Theorem 1, the space embedding $\mathscr{B}(k,\ell) \subset \mathbb{R}^{(\ell+m)^\ell k^m}$ implies the space embedding $\{(\mathcal{B}' - \mathcal{B})\colon \mathcal{B}, \mathcal{B}' \in \mathscr{B}(k,\ell)\} \subset \mathbb{R}^{2(\ell+m)^\ell k^m}$. Therefore, we have

$$\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B}' - \mathcal{B}\|_F}, \mathcal{E} \right\rangle \le \sup_{\boldsymbol{c}\in\mathbb{R}^{2k^m(\ell+m)^\ell}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\ e \right\rangle, \tag{21}$$

where $e \in \mathbb{R}^{2k^m(k+m)^\ell}$ is a vector consisting of i.i.d. sub-Gaussian entries with variance proxy σ^2. Combining (21) and Lemma 6 yields

$$\begin{aligned}
\mathbb{P}\left(\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B}' - \mathcal{B}\|_F}, \mathcal{E} \right\rangle \ge t\right) &\le \mathbb{P}\left(\sup_{\boldsymbol{c}\in\mathbb{R}^{2k^m(\ell+m)^\ell}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\ e \right\rangle \ge t\right) \\
&\le \exp\left(-\frac{t^2}{8\sigma^2} + k^m(\ell+m)^\ell \log 6\right),
\end{aligned}$$

Setting $t = C\sigma\sqrt{k^m(\ell+m)^\ell}$ for sufficiently large $C > 0$ gives

$$\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B}' - \mathcal{B}\|_F}, \mathcal{E} \right\rangle \lesssim \sigma\sqrt{k^m(\ell+m)^\ell}, \tag{22}$$

with high probability.
Finally, combining all sources of error from Lemma 1 and inequalities (20), (22), (19) yields

$$\frac{1}{d^m}\|\hat{\Theta}^{\mathrm{BC}} \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F \lesssim \left(\sigma^2 \frac{\log d}{d^{m-1}}\right)^{\beta\min(\alpha,1)} + \sigma^2 \frac{k^m(\ell+m)^\ell}{d^m} + \frac{L^2}{k^{2\min(\alpha,\ell+1)}}. \tag{23}$$

Finally, optimizing (23) with respect to (k, l) gives that

$$(23) \lesssim \begin{cases} L^2 \left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}} d^{-\frac{2m\alpha}{m+2\alpha}}, & \text{when } \alpha < c(\alpha, \beta, m), \\ \left(\frac{\sigma^2 \log d}{d^{m-1}}\right)^{\beta \min(\alpha, 1)}, & \text{when } \alpha \geq c(\alpha, \beta, m), \end{cases}$$

under the choice

$$\ell^* = \min\left(\lfloor \alpha \rfloor, \lfloor c(\alpha, \beta, m) \rfloor\right), \quad k^* = c_1 d^{m/(m+2\min(\alpha, \ell^*+1))},$$

where $c(\alpha, \beta, m) := \frac{m(m-1)\beta \min(\alpha,1)}{\max(0,\, 2(m-(m-1)\beta \min(\alpha,1)))}$.

□

B.4 Auxiliary Lemmas

Proof of Lemma 3. We provide the proof for $m = 3$ only. The extension to higher orders ($m \geq 4$) uses exactly the same techniques and thus is omitted. Let us pick $\omega_1, \ldots, \omega_{k/3} \in \{0,1\}^{k^2/9}$ such that $\rho_H(\omega_p, \omega_q) \geq k^2/36$ for all $p \neq q \in [k/3]$. This selection is possible by lemma 7. Fixing such $\omega_1, \ldots, \omega_{k/3}$, we define a symmetric core tensor $\mathcal{S} \in \mathbb{R}^{k \times k \times k}$ for $p < q < r$,

$$\mathcal{S}(p,q,r) = \begin{cases} s_{p,q,r} & \text{if } p \in \{1, \ldots, k/3\}, q \in \{k/3+1, \ldots, 2k/3\}, r \in \{2k/3+1, \ldots, k\}, \\ 0 & \text{Otherwise,} \end{cases}$$

where $\{s_{p,q,r} \colon p \in \{1, \ldots, k/3\}, q \in \{k/3+1, \ldots, 2k/3\}, r \in \{2k/3+1, \ldots, k\}\}$ satisfies

$$\begin{aligned} \boldsymbol{s}(r) &:= \operatorname{vec}\left(\mathcal{S}\left(1 : \frac{k}{3},\ \frac{k}{3}+1 : \frac{2k}{3},\ r\right)\right) \\ &= \sqrt{\frac{c\sigma^2 \log k}{d^2}} \omega_{r-2k/3} \quad \text{for any } r \in \{2k/3+1, \ldots, k\}. \end{aligned} \tag{24}$$

The choice of constant $c > 0$ is deferred to a later part of the proof. Notice that for any $r_1, r_2 \in \{2k/3+1, \ldots, k\}$, we have

$$\|\boldsymbol{s}(r_1) - \boldsymbol{s}(r_2)\|_F^2 \geq \frac{c\sigma^2 k^2 \log k}{36 d^2}. \tag{25}$$

Define a subset of permutation set $\Pi(d, k)$ by

$$\mathcal{Z} = \left\{ z \in \Pi(d,k) \colon |z^{-1}(p)| = \frac{d}{k} \text{ for } a \in [k], z^{-1}(a) = \left\{ \frac{(p-1)d}{k} + 1, \ldots, \frac{pd}{k} \text{ for } p \in [2k/3] \right\} \right\}.$$

Each $z \in \mathcal{Z}$ induces a block tensor in $\mathcal{B}(k, 0)$. We consider the collection of block tensors induced by $\mathcal{Z}$; i.e.,

$$\mathcal{B}(\mathcal{Z}) = \{\Theta^z \in \mathbb{R}^{d \times d \times d} \colon \Theta^z(i,j,k) = \mathcal{S}(z(i), z(j), z(k)) \text{ for } z \in \mathcal{Z}\}.$$

To apply Proposition 1, we find upper bound $\sup_{\Theta, \Theta' \in \mathcal{B}(\mathcal{Z})} D(\mathbb{P}_\Theta | \mathbb{P}_{\Theta'})$ and lower bound $\log \mathcal{M}(\epsilon, \mathcal{B}(\mathcal{Z}), \rho)$, where ρ is defined by $\rho(\Theta, \Theta') = \frac{1}{n^3}\|\Theta - \Theta'\|_F^2$. For sub-Gaussian signal plus noise model, we have

$$D(\mathbb{P}_\Theta | \mathbb{P}_{\Theta'}) \leq \frac{1}{2\sigma^2}\|\Theta - \Theta'\|_F \leq \frac{1}{2\sigma^2} d^3 \frac{c\sigma^2 \log k}{d^2} = \frac{cd \log k}{2}, \tag{26}$$

where the first inequality holds for any $\Theta, \Theta' \in \mathcal{B}(\mathcal{Z})$ by [3, Proposition 4.2]. Now we provide a lower bound of the packing number $\log \mathcal{M}(\epsilon, \mathcal{B}(\mathcal{Z}), \epsilon)$ with $\epsilon^2 \asymp \frac{\sigma^2 \log k}{d^2}$. From the construction of $\mathcal{S}$ in (24), we have one to one correspondence between $\mathcal{Z}$ and $\mathcal{B}(\mathcal{Z})$. Thus $\mathcal{M}(\epsilon, \mathcal{B}(\mathcal{Z}), \rho) = \mathcal{M}(\epsilon, \mathcal{Z}, \rho')$ for some metric ρ' on $\mathcal{Z}$ defined by $\rho'(z_1, z_2) = \rho(\Theta^{z_1}, \Theta^{z_2})$. Let P be the packing set in $\mathcal{Z}$ with the same cardinality of $\mathcal{M}(\epsilon, \mathcal{Z}, \rho')$. Given any $z \in \mathcal{Z}$, define its ϵ-neighbor by $\mathcal{N}(z, \epsilon) = \{z' \in \mathcal{Z} \colon \rho'(z, z') \leq \epsilon\}$. Then, we have $\cup_{z \in P} \mathcal{N}(z, \epsilon) = \mathcal{Z}$, because the cardinality of P is same as packing number $\mathcal{M}(\epsilon, \mathcal{Z}, \rho')$. Therefore, we have

$$|\mathcal{Z}| \leq \sum_{z \in P} |\mathcal{N}(z, \epsilon)| \leq |P| \max_{z \in P} |\mathcal{N}(z, \epsilon)|. \tag{27}$$

It remains to find the upper bound of $\max_{z\in P}|\mathcal{N}(z,\epsilon)|$. For any $z_1, z_2 \in \mathcal{Z}$, $z_1(i) = z_2(i)$ for $i \in [2d/3]$ and $|z_1^{-1}(p)| = d/k$ for all $p \in [k]$. Therefore,

$$\begin{aligned}
\rho'^2(z_1,z_2) &\geq \frac{1}{d^3}\sum_{1\leq i_1\leq d/3<i_2\leq 2d/3\leq i_3\leq d}\left(\mathcal{S}\left(z_1(i_1),z_1(i_2),z_1(i_3)\right)-\mathcal{S}\left(z_2(i_1),z_2(i_2),z_2(i_3)\right)\right)^2\\
&= \frac{1}{d^3}\sum_{2n/3<i_3\leq n}\sum_{1\leq p\leq k/3<q\leq 2k/3}\sum_{i_1\in z_1^{-1}(p),i_2\in z_1^{-1}(q)}\left(\mathcal{S}\left(p,q,z_1(i_3)\right)-\mathcal{S}\left(p,q,z_2(i_3)\right)\right)^2\\
&= \frac{1}{d^3}\sum_{2n/3<i_3\leq n}\sum_{1\leq p\leq k/3<q\leq 2k/3}\left(\frac{d}{k}\right)^2\left(\mathcal{S}\left(p,q,z_1(i_3)\right)-\mathcal{S}\left(p,q,z_2(i_3)\right)\right)^2\\
&= \frac{1}{d^3}\sum_{2n/3<i_3\leq n}\left(\frac{d}{k}\right)^2\|\boldsymbol{s}(z_1(i_3)-\boldsymbol{s}(z_2(i_3))\|_F^2\\
&\geq \frac{c\sigma^2\log k}{36d^3}|\{j\colon z_1(j)\neq z_2(j)\}|,
\end{aligned}$$

where the last inequality is from (25). Hence with the choice of $\epsilon^2 = \frac{c\sigma^2\log k}{288d^2}$, we have $|\{j\colon z(j)\neq z'(j)\}|\leq d/8$ for any $z'\in\mathcal{N}(z,\epsilon)$. This implies

$$|\mathcal{N}(z,\epsilon)| \leq \binom{d}{d/8}k^{d/8} \leq (8e)^{d/8}k^{d/8} \leq \exp\left(\frac{1}{5}d\log k\right), \tag{28}$$

for sufficiently large k. Now we find the lower bound of $|\mathcal{Z}|$ based on Stirling's formula,

$$|\mathcal{Z}| = \frac{(d/3)!}{[(d/k)!]^{k/3}} = \exp\left(\frac{1}{3}d\log k + o(d\log k)\right) \geq \exp\left(\frac{1}{4}d\log k\right). \tag{29}$$

Plugging (28) and (29) into (27) yields

$$\mathcal{M}(\epsilon,\mathcal{B}(\mathcal{Z}),\rho) = |P| \geq \frac{\max_{z\in P}\mathcal{N}(z,\epsilon)}{|\mathcal{Z}|} \geq \exp\left(\frac{1}{20}d\log k\right). \tag{30}$$

Finally, applying Proposition 1 based on (26) and (30) gives

$$\inf_{\hat{\Theta}}\sup_{\Theta\in\mathcal{B}(\mathcal{Z})}\mathbb{P}\left(\frac{1}{d^3}\|\hat{\Theta}-\Theta\|_F^2\geq\frac{C\sigma^2\log k}{d^2}\right) = \inf_{\hat{\Theta}}\sup_{z\in\mathcal{Z}}\mathbb{P}\left(\frac{1}{d^3}\|\hat{\Theta}-\Theta\|_F^2\geq\frac{C\sigma^2\log k}{d^2}\right)\geq 0.9,$$

with some constant $C>0$ for sufficiently small $c>0$ in (24). □

Lemma 5 (Sub-Gaussian maxima under full embedding)**.** *Let $\boldsymbol{A}\in\mathbb{R}^{d_1\times d_2}$ be a deterministic matrix with rank $r\leq\min(d_1,d_2)$. Let $\boldsymbol{y}\in\mathbb{R}^{d_1}$ be a sub-Gaussian random vector with variance proxy σ^2. Then, there exists a sub-Gaussian random vector $\boldsymbol{x}\in\mathbb{R}^r$ with variance prosy σ^2 such that*

$$\max_{\boldsymbol{p}\in\mathbb{R}^{d_2}}\left\langle\frac{\boldsymbol{Ap}}{\|\boldsymbol{Ap}\|_2},\boldsymbol{y}\right\rangle = \max_{\boldsymbol{q}\in\mathbb{R}^r}\left\langle\frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2},\boldsymbol{x}\right\rangle.$$

Proof. Let $\boldsymbol{u}_i\in\mathbb{R}^{d_1},\boldsymbol{v}_j\in\mathbb{R}^{d_2}$ singular vectors and $\lambda_i\in\mathbb{R}$ be singular values of of $\boldsymbol{A}$ such that $\boldsymbol{A}=\sum_{i=1}^r\lambda_i\boldsymbol{u}_i\boldsymbol{v}_i^T$. Then for any $\boldsymbol{p}\in\mathbb{R}^{d_2}$, we have

$$\boldsymbol{Ap} = \sum_{i=1}^r\lambda_i\boldsymbol{u}_i\boldsymbol{v}_i^T\boldsymbol{p} = \sum_{i=1}^r\lambda_i(\boldsymbol{v}_i^T\boldsymbol{p})\boldsymbol{u}_i = \sum_{i=1}^r\alpha_i\boldsymbol{u}_i,$$

where $\boldsymbol{\alpha}(\boldsymbol{p}) = (\alpha_1,\ldots,\alpha_r)^T := \left(\lambda_1(\boldsymbol{v}_1^T\boldsymbol{p}),\ldots,\lambda_r(\boldsymbol{v}_r^T\boldsymbol{p})\right)^T\in\mathbb{R}^r$. Notice that $\boldsymbol{\alpha}(\boldsymbol{p})$ covers $\mathbb{R}^r$ in the sense that $\{\boldsymbol{\alpha}(\boldsymbol{p})\colon\boldsymbol{p}\in\mathbb{R}^{d_2}\}=\mathbb{R}^r$. Therefore, we have

$$\begin{aligned}
\max_{\boldsymbol{p}\in\mathbb{R}^{d_2}}\left\langle\frac{\boldsymbol{Ap}}{\|\boldsymbol{Ap}\|_2},\boldsymbol{y}\right\rangle &= \max_{\boldsymbol{p}\in\mathbb{R}^{d_2}}\sum_{i=1}^r\frac{\alpha_i}{\|\boldsymbol{\alpha}(\boldsymbol{p})\|_2}\boldsymbol{u}_i^T\boldsymbol{y}\\
&= \max_{\boldsymbol{p}\in\mathbb{R}^{d_2}}\left\langle\frac{\boldsymbol{\alpha}(\boldsymbol{p})}{\|\boldsymbol{\alpha}(\boldsymbol{p})\|_2},\boldsymbol{x}\right\rangle\\
&= \max_{\boldsymbol{q}\in\mathbb{R}^r}\left\langle\frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2},\boldsymbol{x}\right\rangle,
\end{aligned}$$

where we define $\boldsymbol{x}=(\boldsymbol{u}_1^T\boldsymbol{y},\ldots,\boldsymbol{u}_r^T\boldsymbol{y})^T\in\mathbb{R}^r$. Since $\boldsymbol{u}_i^T\boldsymbol{y}$ is sub-Gaussian with variance proxy σ^2 because of orthonormality of $\boldsymbol{u}_i$, the proof is completed. □

Remark 1. In particular, if $\boldsymbol{x} \in \mathbb{R}^r, \boldsymbol{y} \in \mathbb{R}^{d_1}$ are two Gaussian random vectors with i.i.d. entries drawn from $N(0, \sigma^2)$. Define two Gaussian maximums

$$F(\boldsymbol{x}) \stackrel{\text{def}}{=} \max_{\boldsymbol{q} \in \mathbb{R}^r} \left\langle \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2}, \boldsymbol{x} \right\rangle, \qquad G(\boldsymbol{x}) \stackrel{\text{def}}{=} \max_{\boldsymbol{p} \in \mathbb{R}^{d_2}} \left\langle \frac{\boldsymbol{A}\boldsymbol{p}}{\|\boldsymbol{A}\boldsymbol{p}\|_2}, \boldsymbol{y} \right\rangle.$$

Then $F(\boldsymbol{x}) = G(\boldsymbol{y})$ in distribution. This equality holds because $(\boldsymbol{u}_1^T \boldsymbol{y}, \ldots, \boldsymbol{u}_r^T \boldsymbol{y})$ is again Gaussian random vectors whose entries are i.i.d. drawn from $N(0, \sigma^2)$.

Lemma 6 (Theorem 1.19 in [8])**.** *Let $e \in \mathbb{R}^d$ be a sub-Gaussian vector with variance proxy σ^2. Then,*

$$\mathbb{P}\left(\max_{\boldsymbol{c} \in \mathbb{R}^d} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2}, e \right\rangle \geq t \right) \leq \exp\left(-\frac{t^2}{8\sigma^2} + d \log 6 \right).$$

Proposition 1 (Proposition 4.1 in [3])**.** *Let (Ξ, ρ) be a metric space and $\{\mathbb{P}_\xi : \xi \in \Xi\}$ be a collection of probability measure. For any totally bounded $T \subset \Xi$, define the Kullback-Leibler diameter of T by $d_{KL}(T) = \sup_{\xi, \xi' \in T} D(\mathbb{P}_\xi | \mathbb{P}_{\xi'})$. Then,*

$$\inf_{\hat{\xi}} \sup_{\xi \in \Xi} \mathbb{P}_\xi \left\{ \rho(\hat{\xi}, \xi) \geq \frac{\epsilon^2}{4} \right\} \geq 1 - \frac{d_{KL}(T) + \log 2}{\log \mathcal{M}(\epsilon, T, \rho)},$$

where $\mathcal{M}(\epsilon, T, \rho)$ is a packing number of T with respect to the metric ρ.

Lemma 7 (Varshamov-Gilbert bound)**.** *There exists a sequence of subset $\omega_1, \ldots, \omega_N \in \{0,1\}^d$ such that*

$$\rho_H(\omega_i, \omega_j) := \|\omega_i - \omega_j\|_F^2 \geq \frac{d}{4} \text{ for any } i \neq j \in [N],$$

for some $N \geq \exp(d/8)$.

References

[1] Krishnakumar Balasubramanian. Nonparametric modeling of higher-order interactions via hypergraphons. *Journal of Machine Learning Research*, 22:1–25, 2021.

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

[3] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

[4] Chao Gao and Zongming Ma. Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statistical Science*, 36(1):16–33, 2021.

[5] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

[6] 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.

[7] 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.

[8] Jan-Christian Hitter Phillippe Rigollet. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

[9] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

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