Received: 16 December 2020 | Accepted: 10 July 2022 | Published on: 30 October 2022

DOI: 10.1111/rssb.12547

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY Series B Statistical Methodology

Exact clustering in tensor block model: Statistical optimality and computational limit

Rungang Han[1,2] | **Yuetian Luo**[1] | **Miaoyan Wang**[1] | **Anru R. Zhang**[1,2,3,4,5]

[1]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

[2]Department of Statistical Science, Duke University, Durham, NC, USA

[3]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA

[4]Department of Computer Science, Duke University, Durham, NC, USA

[5]Department of Mathematics, Duke University, Durham, NC, USA

Correspondence
Anru R. Zhang, Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA.
Email: anru.zhang@duke.edu

Funding information
NIH Grant, Grant/Award Number: R01 GM131399; NSF Grants, Grant/Award Numbers: CAREER-2203741, CAREER-2141865, DMS-1915978, DMS-2023239, EF-2133740; Wisconsin Alumni Research Foundation

Abstract

High-order clustering aims to identify heterogeneous substructures in multiway datasets that arise commonly in neuroimaging, genomics, social network studies, etc. The non-convex and discontinuous nature of this problem pose significant challenges in both statistics and computation. In this paper, we propose a tensor block model and the computationally efficient methods, *high-order Lloyd algorithm* (HLloyd), and *high-order spectral clustering* (HSC), for high-order clustering. The convergence guarantees and statistical optimality are established for the proposed procedure under a mild sub-Gaussian noise assumption. Under the Gaussian tensor block model, we completely characterise the statistical-computational trade-off for achieving high-order exact clustering based on three different signal-to-noise ratio regimes. The analysis relies on new techniques of high-order spectral perturbation analysis and a 'singular-value-gap-free' error bound in tensor estimation, which are substantially different from the matrix spectral analyses in the literature. Finally, we show the merits of the proposed procedures via extensive experiments on both synthetic and real datasets.

KEYWORDS

clustering, computational limit, high-order Lloyd algorithm, statistical optimality, tensor block model

1 | INTRODUCTION

High-order tensors have received increasing recent attention in many fields including social networks (Kolda & Bader, 2006), computer vision (Koniusz & Cherian, 2016), neuroscience (Zhang et al., 2019) and genomics (Hore et al., 2016). Tensors provide an effective representation of the hidden structures in multi-way data. One of the popular structure utilised in tensor data analysis is the so-called low-rankness, which decomposes the signal tensor into a low-dimensional core tensor and multiple matrix factors, one on each modes (Kolda, 2001). Despite many celebrated results in tensor data analysis under the low-rank formulation, such as tensor regression (Zhou et al., 2013), tensor completion (Xia et al., 2021), tensor PCA (Zhang & Xia, 2018), and generalised tensor learning (Han et al., 2022), another important model, *multi-way tensor block model*, has not been well studied yet.

Figure 1 shows an example of multi-way tensor block model for an order-3 tensor in which each of the modes is partitioned into several clusters. The goal is to identify the block structure (clustering), as well as to recover the whole tensor data (estimation), from observed data. In comparison with the low-rankness, the discrete block structure is more interpretable, because of the membership information encoded by loading matrices. The tensor block model and high-order clustering arise commonly in practical applications. For example,

- *Multi-tissue gene expression analysis*. Gene expression profiles such as scRNA-seq and micro-arrays are collected from multiple individuals across numbers of tissues (Melé et al., 2015; Wang et al., 2019). Genes involved in the same biological function typically exhibit similar expressions for some group of tissues and individuals, while these expression values vary from group to group. Similarly, tissues/individuals exhibit clustering patterns due to the similarity therein. Investigating the complex interactions among these three entities is of great scientific interest.
- *Multi-layer network analysis*. Multi-layer networks arise commonly in longitudinal studies of network (Lei et al., 2020) and multi-relational data (Nickel et al., 2011). A multi-layer network consists of multiple directed/undirected graphs (or adjacency matrices), where each graph represents the connection among the same set of vertices. The multi-layer network data can be organised as an order-3 tensor with the first two modes being vertices and the third mode being the contexts under which the graph is observed. Depending on how the connection edges are encoded, the tensor entries can be either binary (indicating presence/absence of connection) or continuous (representing weighted connection strength).
- *Online click-through prediction*. In e-commerce, predicting click-through for user-item pairs in a time-specific way plays an important role in online recommendation system (Shan

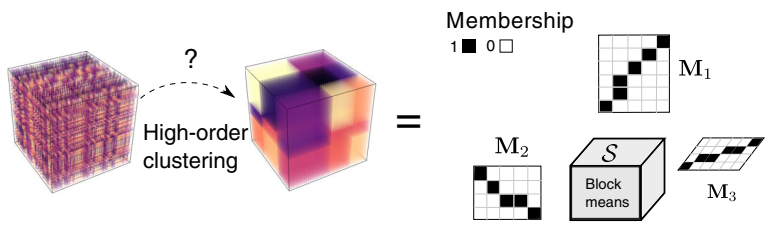


FIGURE 1 High-order clustering aims to identify the block structure represented by membership matrices on each of the modes. [Colour figure can be viewed at wileyonlinelibrary.com]

et al., 2016; Sun et al., 2015). The click-through data in a specific day can be organised as an order-3 tensor, where each entry indexed by (users, items, time) represents the total number of user–item interactions in a time period (e.g., 24 different hours in that day). The users/items often exhibit clustering structures due to similar preferences/attributes. In addition, the shopping behaviour also varies in time, and this heterogeneity depends on the specific group of users and items.

Additional applications of tensor block model include hypergraph clustering (Chien et al., 2019; Ke et al., 2019), collaborative filtering (Zhang et al., 2021) and signal detection in three-(3D)/four-dimensional (4D) imaging (Zhang et al., 2020), among others.

The multi-way structure of a data tensor imposes unique challenges to clustering analysis. In the vector case, clustering methods find subgroups of the observations based on entrywise similarity; while in the matrix case, bi-clustering algorithms seek to simultaneously identify the block structure of observations (rows) and features (columns) (Busygin et al., 2008). Recent developments of tensor-based high-order clustering methods roughly fall into two types. The first approach utilises the maximum likelihood estimation (MLE) (Wang & Zeng, 2019) to search for the tensor block structure. The MLE, however, is a mixed-integer programming and therefore NP-hard to compute in general (Aloise et al., 2009). The computational intractability renders the statistical inference less useful in practice. The second approach adopts polynomial-time algorithms for surrogate objectives. Efforts in this vein include convex relaxation (Chi et al., 2020) and spectral relaxation (Wu et al., 2016; Zha et al., 2002). Despite the popularity, these methods often sacrifice the statistical accuracy for computational feasibility. To our best knowledge, a provable scheme that achieves both statistical and computational efficiency has yet to be developed.

In this paper, we develop a computationally efficient procedure for the task of high-order clustering in tensor block model. The procedure operates in steps: *High-order Spectral Clustering* (HSC) and *High-order Lloyd* (HLloyd). The proposed HSC algorithm involves a power iteration procedure. While the statistical property of the power iteration has been recently established under a strong singular value gap condition (Luo et al., 2021; Zhang & Xia, 2018), the previous result is not applicable to our analysis, because of the possible lack of singular value gap in block tensors even under the model identifiablility conditions. This difference originates from the unique 'discrete' low-rank structure in tensor block model as opposed to the 'continuous' low-rank structure considered in Zhang and Xia (2018). The discrete setting requires new theoretical analysis for high-order clustering methods (see Section 2.2 and Remark 3). We then establish the clustering error rate for HSC algorithm under modest conditions (see Section 4.2). The second component of the proposed procedure, HLloyd algorithm, can be seen as a high-order extension of Lloyd algorithm for one-dimensional *k-means* to order-d clustering. Compared to the analysis of Lloyd algorithm for vector clustering with a single discrete structure (Lu & Zhou, 2016), the multiple discrete structures in high-order clustering make the analysis more challenging (see details in Remark 4). We prove that, under warm initialisation, our HLloyd algorithm solves the high-order clustering problem with optimality guarantees in tensor block model.

Apart from the newly proposed algorithm, we discover an intriguing interplay between statistical optimality and computational efficiency of high-order clustering in tensor block model. Specifically, we introduce a notion of signal-to-noise ratio (SNR) for tensor block model that quantifies the minimum gaps between block means (see formal definition in Section 4.1) over the noise level. This notion completely characterises the hardness of the high-order clustering in tensor block model. Our main phase transition results can be informally summarised as follows.

Theorem 1 (Informal results). *Consider the high-order clustering on an order-d dimension-p tensor under the Gaussian tensor block model (see (3) in Section 2.2). Suppose* $\mathrm{SNR} = p^{\gamma}$ *and* $p \to \infty$.

- *When* $\gamma > -d/2$, *the proposed HLloyd + HSC algorithm performs exact clustering (Theorem 4);*
- *When* $\gamma < -(d-1)$, *no algorithm can achieve exact clustering (Theorem 6);*
- *When* $-(d-1) < \gamma < -d/2$, *MLE achieves exact clustering at the cost of being computationally intractable; and no polynomial-time algorithm can achieve exact clustering under a computational hardness assumption for hypergraphic planted clique (HPC) detection (Theorem 7).*

Here, the exact clustering means that the clustering labels are precisely recovered with high probability.

Figure 2 summarises the phase transitions in high-order clustering under the tensor block model. In the strong SNR region A ($\mathrm{SNR} = p^{\gamma}, \gamma > -d/2$), we prove that the combination of HSC and HLloyd achieves exact clustering in polynomial time. We find that the estimation error bound of the target tensor is free of the tensor dimension (Theorem 5), which is distinct from the tensor estimation error bounds in the literature under the continuous low-rank structure (see Remark 7). In the weak SNR region B ($\gamma < -(d-1)$), we develop a minimax lower bound to show that no algorithm succeeds in high-order clustering for tensor block model. In the modest SNR region C ($-(d-1) < \gamma < -d/2$), we show that the problem is statistically possible while computationally infeasible. That is, computing any estimate that achieves exact clustering is as hard as solving a version HPC detection problem which is conjectured to be polynomial-time unsolvable (see details in Section 5). Note that the former two SNR regions apply to matrix biclustering ($d = 2$) while the latter statistical-computational gap region exists only for high-order tensors with $d \geq 3$. To the best of our knowledge, we are among the first to establish both of the statistical and computational limits for high-order clustering in tensor block models.

1.1 | Related literature

Our work is related to but also clearly distinctive from several lines of existing work. Classic clustering algorithms such as k-means (Jain, 2010) and spectral clustering (Von Luxburg, 2007) have been widely used in statistics and machine learning. In the order-1 (vector) case, the clustering problem is usually formulated under Gaussian mixture model. The optimal statistical guarantee has been developed for the state-of-art clustering algorithms,

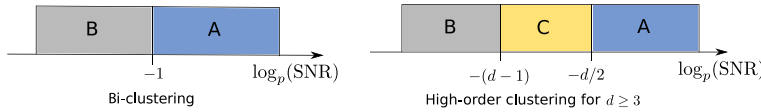


FIGURE 2 Phase transition diagram for high-order clustering in order-d tensor block model with dimension $(p, \ldots, p)$. A statistical-computational gap (Region C) arises only for tensors of order 3 or greater. [Colour figure can be viewed at wileyonlinelibrary.com]

including spectral clustering (Löffler et al., 2021), EM algorithm (Wu & Zhou, 2019) and Lloyd algorithm (Lu & Zhou, 2016). In the order-2 (matrix) case, clustering methods have been studied under the stochastic block model (Abbe, 2017), bi-clustering (Gao et al., 2016), and bipartite community detection (Zhou & Amini, 2019). These existing methods and theories are inapplicable to high-order clustering due to the distinct algebraic properties of general order-d tensors. The high-order clustering also exhibits many distinct features compared to the clustering for vectors or bi-clustering for matrices, for example, the statistical-computational gap arises only from tensors of order 3 or greater as shown in Figure 2.

In addition, our work is related to the recent development on low-rank tensor decomposition, in which the main goal is to find the best low-rank approximation of a tensor. Numerous algorithms have been proposed, such as truncated power iteration (Anandkumar et al., 2014), high-order singular value decomposition (SVD) (De Lathauwer et al., 2000b), high-order orthogonal iteration (HOOI) (De Lathauwer et al., 2000a), sparse tensor SVD (Zhang & Han, 2019). Although the block structure implies low-rankness (see Section 2.2), the classic low-rank tensor spectral methods fail to fully utilise the structural information of tensor block model. Moreover, the singular value gap condition, which was commonly imposed in the literature on tensor decomposition, does not generally hold in the tensor block model, and thus new technical tools are required for our high-order clustering problem.

Another related topic is on the statistical and computational trade-off in high-dimensional statistics. This topic focuses on the gap between SNR requirements under which the problem is information-theoretically solvable versus polynomial-time solvable. Some common structures yielding these trade-offs include sparsity (Chen & Xu, 2016; Ma & Wu, 2015), robustness (Diakonikolas et al., 2017), tensors (Barak & Moitra, 2016; Richard & Montanari, 2014), etc. In the last decade, a number of schemes have been proposed to provide rigorous evidence for the gap between computational and statistical limits, such as average-case reduction (Brennan et al., 2018; Wang et al., 2016), sum of squares (Barak et al., 2019), statistical query (Feldman et al., 2017), low-degree polynomials (Hopkins & Steurer, 2017), etc. We refer readers to the recent survey (Wu & Xu, 2021) for a review. In this work, we reveal the statistical and computational trade-off for high-order clustering in the tensor block model and give rigorous evidence for the computational limit via average-case reduction.

1.2 | Organisations

The rest of the paper is organised as follows. After a brief review of basic tensor algebra in Section 2.1, we introduce the tensor block model and the high-order tensor clustering problem in Section 2.2. Then, the HSC and HLloyd algorithms are proposed to solve this problem in Section 3. We present the statistical proprieties of the algorithms in Section 4. We study the fundamental statistical and computational limits of high-order clustering in tensor block model in Section 5. Extensive numerical analyses on simulated data, flight route network data, and online click-through data are presented in Sections 6 and 7, respectively. Conclusion and discussion are given in Section 8. Proof sketches for the main results are provided in Section 9, and all detailed proofs are collected in Appendix S1. Codes for running the simulations and real data analyses of this paper are available at https://github.com/Rungang/HLloyd.

2 | MULTI-WAY TENSOR BLOCK MODEL

2.1 | Notations and preliminaries

We use lowercase letters ($a, b, x \ldots$) to denote scalars/vectors. For any $a, b \in \mathbb{R}$, let $a \wedge b$ and $a \vee b$ be the minimum and maximum of a and b, respectively. Suppose $\{a_n\}, \{b_n\}$ are two sequences of positive numbers. We denote $a_n \lesssim b_n$ or $a_n = \mathcal{O}(b_n)$ (respectively, $a_n \gtrsim b_n$) if there exists a constant $C > 0$ such that $a_n \le Cb_n$ (respectively, $a_n \ge Cb_n$) for all n; and we denote $a_n \asymp b_n$ if there exists c, C such that $ca_n \le b_n \le Ca_n$ for all n. We denote matrices by bold uppercase letters ($\mathbf{A}, \mathbf{B}, \mathbf{U}, \ldots$). Let $\mathbb{O}_{p,r}$ be the collection of all p-by-r matrices with orthonormal columns: $\mathbb{O}_{p,r} = \{\mathbf{U} \in \mathbb{R}^{p\times r}: \mathbf{U}^\top\mathbf{U} = \mathbf{I}\}$, where $\mathbf{I}$ is the identity matrix. We use $\mathbf{A}_{ij}, \mathbf{A}_{i:}$, and $\mathbf{A}_{:j}$ to denote the (i,j) th entry, the i th row, and the j th column of $\mathbf{A}$, respectively. Let $\lambda_1(\mathbf{A}) \ge \lambda_2(\mathbf{A}) \ge \cdots \ge 0$ be the singular values of $\mathbf{A}$ in descending order and we use $\mathrm{SVD}_r(\mathbf{A})$ to denote the matrix comprised of the top r left singular vectors of $\mathbf{A}$. We use $\|\mathbf{A}\| = \lambda_1(\mathbf{A})$ to denote the spectral norm of $\mathbf{A}$ and use $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\mathbf{A}_{ij}^2} = \sqrt{\sum_{i=1}^{p_1\wedge p_2}\lambda_i^2(\mathbf{A})}$ to denote its Frobenius norm. For any matrix $\mathbf{A} = [a_1, \ldots, a_J] \in \mathbb{R}^{I\times J}$ and $\mathbf{B} \in \mathbb{R}^{K\times L}$, the *Kronecker product* is defined as the (IK)-by-(JL) matrix $\mathbf{A} \otimes \mathbf{B} = [a_1 \otimes \mathbf{B} \cdots a_J \otimes \mathbf{B}]$.

Recall that tensors are multi-way arrays. We call the number of modes of a tensor as its order. We use calligraphy letters ($\mathcal{A}, \mathcal{X}, \mathcal{Y} \ldots$) to denote tensors of order 3 or higher. For example, if a tensor $\mathcal{X}$ represents a d-way array of size $p_1 \times p_2 \times \cdots \times p_d$, we say $\mathcal{X}$ is an order-d tensor and write $\mathcal{X} \in \mathbb{R}^{p_1\times\cdots\times p_d}$. The $(i_1, \ldots, i_d)$ th element of a tensor $\mathcal{X}$ is written as $\mathcal{X}_{i_1,\ldots,i_d}$. Let $\mathcal{X}$ and $\mathcal{Y}$ be two tensors of the same dimension, the inner product of them is defined as $\langle\mathcal{X}, \mathcal{Y}\rangle = \sum_{i_1,\ldots,i_d}\mathcal{X}_{i_1,\ldots,i_d}\mathcal{Y}_{i_1,\ldots,i_d}$. The Frobenius norm of the tensor $\mathcal{X}$ is defined as $\|\mathcal{X}\|_{\mathrm{F}} = \langle\mathcal{X}, \mathcal{X}\rangle^{1/2}$. The multi-linear multiplication of a tensor $\mathcal{S} \in \mathbb{R}^{r_1\times\cdots\times r_d}$ by matrices $\mathbf{U}_k \in \mathbb{R}^{p_k\times r_k}$ is defined as

$$(\mathcal{S} \times_1 \mathbf{U}_1 \times \cdots \times_d \mathbf{U}_d)_{i_1,\ldots,i_d} = \sum_{j_1=1}^{r_1}\cdots\sum_{j_d=1}^{r_d}\mathcal{S}_{j_1,\ldots j_d}(\mathbf{U}_1)_{i_1j_1}\cdots(\mathbf{U}_d)_{i_dj_d},$$

which results in an order-d $(p_1, \ldots, p_d)$-dimensional tensor. We also introduce the matricization operator that transforms tensors to matrices. Particularly the mode-1 matricization for $\mathcal{X} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ is defined as

$$\mathcal{M}_1(\mathcal{X}) \in \mathbb{R}^{p_1\times(p_2\cdots p_d)}, \text{ where } [\mathcal{M}_1(\mathcal{X})]_{j_1,j_2+p_2(j_3-1)+\cdots+p_2\cdots p_{d-1}(j_d-1)} = \mathcal{X}_{j_1,\ldots j_d}.$$

Each row of $\mathcal{M}_k(\mathcal{X})$ is the vectorisation of a mode-k slice. The following identity that relates the matrix-tensor product and matricization will be used extensively in our analysis:

$$\mathcal{M}_k(\mathcal{S} \times_1 \mathbf{U}_1 \times \cdots \times_d \mathbf{U}_d) = \mathbf{U}_k\mathcal{M}_k(\mathcal{S})(\mathbf{U}_{k+1} \otimes \cdots \otimes \mathbf{U}_d \otimes \mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_{k-1})^\top.$$

We say a tensor $\mathcal{X} \in \mathbb{P}^{p_1\times\cdots\times p_d}$ has Tucker-rank $(r_1, \ldots, r_d)$ if $r_k = \mathrm{rank}(\mathcal{M}_k(\mathcal{X}))$. In this case $\mathcal{X}$ admits a *Tucker decomposition*:

$$\mathcal{X} = \mathcal{S} \times_1 \mathbf{U}_1 \times \cdots \times_d \mathbf{U}_d \tag{1}$$

for some $\mathcal{S} \in \mathbb{R}^{r_1\times\cdots\times r_d}$ and $\mathbf{U}_k \in \mathbb{R}^{p_k\times r_k}$. The readers are referred to Kolda and Bader (2009) for a more comprehensive tutorial on tensor decomposition.

A clustering that partitions p entities into r clusters is represented by a vector $z \in [r]^p$ such that the i th entry of z equals to s if and only if the i th entity belongs to the s th cluster. Here, we use shorthand $[r] := \{1, 2, \dots, r\}$ to denote the r-set. For two clusters $a = (a_1, \dots, a_p)^\top$, $b = (b_1, \dots, b_p)^\top \in [r]^p$ on the same set, we denote the *misclassification rate* as:

$$h(a, b) = \min_{\pi \in \Pi_r} \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}\{a_i \neq \pi(b_i)\},$$

where Π_r is the collection of all permutations π on $[r]$ and $\mathbb{I}(\cdot)$ is the indicator function. Let $\hat{z}$ be an estimator of $z \in [r]^p$. As sample size goes to infinity, we say $\hat{z}$ is consistent if

$$\mathbb{P}(h(\hat{z}, z) > \varepsilon) \to 0, \qquad \forall \varepsilon > 0;$$

and we say $\hat{z}$ exactly recovers z if

$$\mathbb{P}(h(\hat{z}, z) = 0) \to 1.$$

Finally, we use $C, C_0, C_1, \dots$ and $c, c_0, c_1, \dots$ to represent generic large and small positive constants, respectively. The actual values of these generic symbols may differ from line to line. We introduce the following notions for $p_1 \times \cdots \times p_d$ tensors with rank $(r_1, \cdots, r_d)$:

$$\overline{p} = \max_{k \in [d]} p_k, \ \underline{p} = \min_{k \in [d]} p_k, \ p_* = \prod_{k \in [d]} p_k, \ p_{-k} = p_*/p_k,$$
$$\overline{r} = \max_{k \in [d]} r_k, r_* = \prod_{k \in [d]} r_k, \ r_{-k} = r_*/r_k.$$

2.2 | Tensor block model

Let $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ be an order-d $(p_1, \dots, p_d)$-dimensional data tensor of interest. The tensor block model assumes an underlying checkerbox structure in the signal tensor (see Figure 1). Specifically, suppose there are r_k clusters in the k th mode of the signal tensor for all $k \in [d]$, and we represent the clustering along k th mode by a vector $z_k \in [r_k]^{p_k}$. Then, the entries $\mathcal{Y}_{j_1, \dots j_d}$ are realisations from the following block model:

$$\mathcal{Y}_{j_1, \dots j_d} = \mathcal{S}_{(z_1)_{j_1}, \dots, (z_d)_{j_d}} + \mathcal{E}_{j_1, \dots j_d}, \quad \forall (j_1, \dots, j_d) \in [p_1] \times \cdots \times [p_d], \tag{2}$$

where $\mathcal{S} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$ is the core tensor with collected block means, and $\mathcal{E}_{j_1, \dots j_d}$ s are some mean-zero observational noises. Model (2) can be equivalently written in a form of tensor-matrix product:

$$\mathcal{Y} = \mathcal{S} \times_1 \mathbf{M}_1 \times \cdots \times_d \mathbf{M}_d + \mathcal{E}, \tag{3}$$

where $\mathcal{E} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is the noise tensor, $\mathbf{M}_k \in \{0, 1\}^{p_k \times r_k}$ is the membership matrix associated with z_k such that $(\mathbf{M}_k)_{ij} = 1$ if and only if $(z_k)_i = j$. That is, $\mathbf{M}_k$ has one copy of 1 and $(r_k - 1)$ copies of 0s in each row. We will use forms (2) and (3) interchangeably throughout the paper. Note that $\mathbb{E}\mathcal{Y}$ admits a Tucker low-rank structure (1) with Tucker-rank bounded by, but may not equal to, $(r_1, \dots, r_d)$. The discrete structure in $\mathbf{M}_k$ makes the model more informative and brings new challenges as we mentioned in Section 1.

We impose the following distributional assumption on the noise tensor $\mathcal{E}$:

Assumption 1 (Sub-Gaussian noise). Suppose each entry of $\mathcal{E}$ follows an independent zero-mean sub-Gaussian distribution with sub-Gaussian norm bounded by σ:

$$\mathbb{E}\exp\left(\lambda\mathcal{E}_{j_1,\dots j_d}\right) \le e^{\lambda^2\sigma^2/2}, \qquad \forall\lambda\in\mathbb{R}.$$

Assumption 1 holds in some specific problems arising in applications:

- Gaussian tensor block models (GTBM): each entry $\mathcal{Y}_{j_1,\dots j_d}$ is Gaussian-distributed with mean $\mathcal{S}_{(z_1)_{j_1},\dots,(z_d)_{j_d}}$ and variance σ^2. This setting is suitable to model a tensor $\mathcal{Y}$ with continuous entries on $\mathbb{R}$;
- Stochastic tensor block models (STBM): the core tensor $\mathcal{S}\in[0,1]^{r_1\times\cdots\times r_d}$ and each entry $\mathcal{Y}_{(z_1)_{j_1},\dots,(z_d)_{j_d}} \sim \text{Bernoulli}(\mathcal{S}_{(z_1)_{j_1},\dots,(z_d)_{j_d}})$. Let $\mathcal{E}=\mathcal{Y}-\mathbb{E}\mathcal{Y}$. Then Assumption 1 holds for $\sigma\le 1/4$. This setting is suitable to model a tensor $\mathcal{Y}$ with binary entries on $\{0,1\}$.

In this paper, we mainly focus on two tasks on the inference for tensor block model:

- Clustering. Recover the membership matrix $\mathbf{M}_k$, or equivalently the label vector z_k, for each mode.
- Estimation. Estimate the underlying signal tensor $\mathcal{X} := \mathbb{E}(\mathcal{Y})$.

3 | ALGORITHMS FOR HIGH-ORDER CLUSTERING

We introduce the procedure for high-order clustering in this section. The procedure includes two algorithms, HSC and HLloyd, which will be elaborated in the next two subsections.

3.1 | High-order Lloyd algorithm

As a starting point, it is natural to consider the following least squares estimator of $\mathcal{S}$ and z_k:

$$\left(\hat{\mathcal{S}},\hat{z}_1,\dots\hat{z}_d\right) = \underset{\mathcal{S},z_k\in[r_k]^{p_k},k=1,\dots,d}{\arg\min} \sum_{j_1,\dots j_d}\left(\mathcal{Y}_{j_1,\dots j_d}-\mathcal{S}_{(z_1)_{j_1},\dots,(z_d)_{j_d}}\right)^2. \tag{4}$$

This scheme is a mixed-integer programming with one continuous ($\mathcal{S}$) and d discrete ($z_1,\dots,z_d$) arguments. In general, (4) is non-convex and computationally intractable.

Therefore, we propose a new iterative method to solve this problem. Suppose at step t, we have estimators $\mathcal{S}^{(t)}, z_1^{(t)},\cdots,z_d^{(t)}$ and want to update them at step $t+1$. On one hand, given the block membership vectors $(z_1^{(t)},\cdots,z_d^{(t)})$, the optimisation objective function (4) becomes quadratic of $\mathcal{S}$ and the optimal solution can be computed via a block-wise average:

$$\begin{aligned}\mathcal{S}^{(t+1)} &= \underset{\mathcal{S}}{\arg\min}\sum_{j_1,\dots j_d}\left(\mathcal{Y}_{j_1,\dots j_d}-\mathcal{S}_{(z_1^{(t)})_{j_1},\dots,(z_d^{(t)})_{j_d}}\right)^2\\ \Longrightarrow\quad \mathcal{S}^{(t+1)}_{i_1,\dots,i_d} &= \text{Average}\left(\left\{\mathcal{Y}_{j_1,\dots j_d} : (z_k^{(t)})_{j_k}=i_k,\forall k\in[d]\right\}\right),\end{aligned} \tag{5}$$

where Average$(\cdot)$ computes the sample mean given a set of values.

On the other hand, given $\mathcal{S}^{(t)}$ and $(d-1)$ block memberships $(z_1^{(t)},\dots,z_{k-1}^{(t)},z_{k+1}^{(t)},\dots,z_d^{(t)})$, the update of membership vector $z_k^{(t+1)}$ can be obtained by performing a nearest neighbour

search in a dimension-reduced space. Specifically, we first aggregate all mode-k slices of $\mathcal{Y}$ to a dimension-reduced space using $\{z_{k'}^{(t)}\}_{k' \neq k}$, the block information from the other $d-1$ clusters. That is, we calculate $\mathcal{Y}_k^{(t)} \in \mathbb{R}^{r_1 \times \cdots \times r_{k-1} \times p_k \times r_{k+1} \times \cdots \times r_d}$ as

$$(\mathcal{Y}_k^{(t)})_{i_1,\cdots,i_{k-1},j,i_{k+1},\ldots,i_d} = \text{Average}\left(\left\{\mathcal{Y}_{j_1,\cdots,j_{k-1},j,j_{k+1},\ldots,j_d} : (z_l^{(t)})_{j_l} = i_l, \forall l \in [d]/k\right\}\right). \tag{6}$$

Intuitively speaking, when the other $d-1$ memberships $(z_1^{(t)}, \ldots, z_{k-1}^{(t)}, z_{k+1}^{(t)}, \ldots, z_d^{(t)})$ are close to the truth, such an aggregation can significantly reduce the noise level within $\mathcal{Y}_k^{(t)}$. Then, we perform the nearest neighbour search to update the estimate for $(z_k)_j$:

$$(z_k^{(t+1)})_j = \underset{a \in [r_k]}{\arg\min} \left\| \left(\mathcal{M}_k(\mathcal{Y}_k^{(t)})\right)_{j:} - \left(\mathcal{M}_k(\mathcal{S}^{(t)})\right)_{a:} \right\|_2^2. \tag{7}$$

The full procedure is presented in Algorithm 1 and is referred to as the HLloyd algorithm.

Remark 1. Our updating scheme of $z_k^{(j+1)}$ is slightly different from the updating scheme in the classic Lloyd algorithm. In particular, (7) is not an exact local greedy solution for (5) for fixed $\mathcal{S}^{(t)}$ and $(z_1^{(t)}, \ldots, z_{k-1}^{(t)}, z_{k+1}^{(t)}, \ldots, z_d^{(t)})$. In comparison, the computational complexity for the proposed updating scheme, that is, (6) and (7), is $O(p_* + p_k r_*)$, while the complexity of local greedy method is $O(p_* r_k)$, since the proposed updating scheme performs nearest neighbour search only on an r_{-k}-dimensional space rather than the original p_{-k}-dimensional space.

Algorithm 1. High-order Lloyd algorithm

Require: Data tensor $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, initialisation labels $\{z_k^{(0)} \in [r_k]^{p_k}\}$, iteration number T

for all $t = 0$ to $T-1$ **do**

Update the block means $\mathcal{S}^{(t)}$ via

$$\mathcal{S}^{(t)}_{i_1,\ldots,i_d} = \text{Average}\left(\left\{\mathcal{Y}_{j_1,\ldots,j_d} : \left(z_k^{(t)}\right)_{j_k} = i_k, \forall k \in [d]\right\}\right).$$

for all $k = 1, \ldots, d$ **do**

for all $j = 1, \ldots, p_k$ **do**

Calculate $\mathcal{Y}_k^{(t)} \in \mathbb{R}^{r_1 \times \cdots \times r_{k-1} \times p_k \times r_{k+1} \times \cdots \times r_d}$ such that

$$(\mathcal{Y}_k^{(t)})_{i_1,\cdots,i_{k-1},j,i_{k+1},\ldots,i_d} = \text{Average}\left(\left\{\mathcal{Y}_{j_1,\cdots,j_{k-1},j,j_{k+1},\ldots,j_d} : \left(z_l^{(t)}\right)_{j_l} = i_l, \forall l \in [d]/k\right\}\right).$$

Update the mode-k membership for the jth entity $\left(z_k^{(t+1)}\right)_j$ via

$$\left(z_k^{(t+1)}\right)_j = \underset{a \in [r_k]}{\arg\min} \left\| \left(\mathcal{M}_k(\mathcal{Y}_k^{(t)})\right)_{j:} - \left(\mathcal{M}_k(\mathcal{S}^{(t)})\right)_{a:} \right\|_2^2.$$

end for

end for

end for

return Estimated block memberships $\{z_k^{(T)}, k = 1, \ldots, d\}$

3.2 | High-order spectral clustering

While HLloyd provides an iterative optimisation strategy for tensor clustering, initialisation of labels $z_k^{(0)}$ remains unaddressed and turns out to be crucial to the clustering performance. In this section, we propose the HSC algorithm for initialisation. This algorithm is a generalisation of the classic spectral clustering method but with a number of fundamental novelties as we describe below. Note that the main step for spectral clustering is to estimate the singular sub-space of the membership matrix $\mathbf{M}_k$. By tensor algebra, we have

$$\mathbb{E}\mathcal{M}_k(\mathcal{Y}) = \mathcal{M}_k(\mathcal{X}) = \mathbf{M}_k \mathcal{M}_k(\mathcal{S}) \left(\mathbf{M}_{k+1} \otimes \cdots \otimes \mathbf{M}_d \otimes \mathbf{M}_1 \otimes \cdots \otimes \mathbf{M}_{k-1}\right).$$

It is tempting to estimate the singular sub-space of $\mathbf{M}_k$ by

$$\tilde{\mathbf{U}}_k = \mathrm{SVD}_{r_k}(\mathcal{M}_k(\mathcal{Y})), \quad k = 1, \ldots, d. \tag{8}$$

This method is referred to as the high-order singular value decomposition (HOSVD) (De Lathauwer et al., 2000a). However, we find that $\tilde{\mathbf{U}}_k$ here does not fully utilise the low-rank structure in all modes of the data. We propose to further improve the estimate by making the projection of $\mathcal{Y}$ onto the pre-estimated sub-spaces of the other $(d-1)$ modes, that is,

$$\hat{\mathbf{U}}_k = \mathrm{SVD}_{\min\{r_k, r_{-k}\}} \left(\mathcal{M}_k(\mathcal{Y} \times_1 \tilde{\mathbf{U}}_1^\top \times \cdots \times_{k-1} \tilde{\mathbf{U}}_{k-1}^\top \times_{k+1} \tilde{\mathbf{U}}_{k+1}^\top \times \cdots \times_d \tilde{\mathbf{U}}_d^\top)\right). \tag{9}$$

After obtaining $\{\hat{\mathbf{U}}_k\}_{k=1}^d$, we get the initialisation labels $z_k^{(0)}$ by performing k-means on p_k rows of the following projected matrix:

$$\hat{\mathbf{Y}}_k = \hat{\mathbf{U}}_k \hat{\mathbf{U}}_k^\top \mathcal{M}_k \left(\mathcal{Y} \times_1 \hat{\mathbf{U}}_1^\top \times \cdots \times_{k-1} \hat{\mathbf{U}}_{k-1}^\top \times_{k+1} \hat{\mathbf{U}}_{k+1}^\top \times \cdots \times_d \hat{\mathbf{U}}_d^\top\right). \tag{10}$$

The pseudocode of HSC is given in Algorithm 2. In particular, since the exact k-means may be computationally difficult, we use a relaxed k-means in Step (11), which can be efficiently solved by approximation algorithms, such as k-means++ with relaxation factor $M = \mathcal{O}(\log \bar{r})$ (Arthur & Vassilvitskii, 2006).

Our algorithm takes r_k 's as inputs. In our theory, r_k's are allowed to grow with the tensor dimension. In our simulation studies, we assume the true r_k's are known for simplicity; while in practice, we recommend rank selection using data-driven criteria (see (29) in Section 7) such as the Bayesian information criterion (BIC) and/or prior knowledge.

Remark 2. In the matrix bi-clustering setting ($d = 2$), the step of refined SVD (10) can be removed from the algorithm, since $\hat{\mathbf{U}}_k$ and $\tilde{\mathbf{U}}_k$ are provably the same. Moreover, $\hat{\mathbf{U}}_k$ has guaranteed accuracy due to the Eckart–Young–Mirsky theorem (Eckart & Young, 1936; Mirsky, 1960). In contrast, computing the best low-rank approximation of $\mathcal{Y}$ when $d \geq 3$ is NP-hard in general (Hillar & Lim, 2013), and the estimator $\tilde{\mathbf{U}}_k$ does not yield the desired accuracy. We thus introduce an indispensable refinement (10), because the additional projection of $\mathcal{Y}$ on the pre-estimated multi-linear sub-spaces substantially reduces the noise in $\hat{\mathbf{U}}_k$. This step makes the proposed HSC method distinct from a simple extension of the classic spectral methods in the matrix setting.

Algorithm 2. High-order spectral clustering

Require: $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, $r_1, \ldots, r_d$, relaxation factor in k-means: $M > 1$

Compute $\tilde{\mathbf{U}}_k = \mathrm{SVD}_{r_k}(\mathcal{M}_k(\mathcal{Y}))$ for $k = 1, \ldots, d$

for all $k = 1$ to d **do**

Compute the singular space estimator $\hat{\mathbf{U}}_k$ via

$$\hat{\mathbf{U}}_k = \mathrm{SVD}_{\min\{r_k, r_{-k}\}}(\mathcal{M}_k(\mathcal{Y} \times_1 \tilde{\mathbf{U}}_1^\top \times \cdots \times_{k-1} \tilde{\mathbf{U}}_{k-1}^\top \times_{k+1} \tilde{\mathbf{U}}_{k+1}^\top \times \cdots \times_d \tilde{\mathbf{U}}_d^\top)).$$

end for

for all $k = 1$ to d **do**

Calculate $\hat{\mathbf{Y}}_k = \hat{\mathbf{U}}_k \hat{\mathbf{U}}_k^\top \mathcal{M}_k \left(\mathcal{Y} \times_1 \hat{\mathbf{U}}_1^\top \times \cdots \times_{k-1} \hat{\mathbf{U}}_{k-1}^\top \times_{k+1} \hat{\mathbf{U}}_{k+1}^\top \times \cdots \times_d \hat{\mathbf{U}}_d^\top \right)$

Find $z_k^{(0)} \in [r_k]^{p_k}$ and centroids $\hat{x}_1, \ldots, \hat{x}_{r_k} \in \mathbb{R}^{r_{-k}}$ such that

$$\sum_{j=1}^{p_k} \left\| \left(\hat{\mathbf{Y}}_k\right)_{j:}^\top - \hat{x}_{(z_k^{(0)})_j} \right\|_2^2 \le M \min_{\substack{x_1, \ldots, x_{r_k} \in \mathbb{R}^{r_{-k}} \\ z_k \in [r_k]^{p_k}}} \sum_{j=1}^{p_k} \left\| \left(\hat{\mathbf{Y}}_k\right)_{j:}^\top - x_{(z_k)_j} \right\|_2^2, \tag{11}$$

end for

return $\{z_k^{(0)} \in [r_k]^{p_k}, k = 1, \ldots, d\}$

4 | STATISTICAL THEORY

In this section, we study the statistical properties of the proposed algorithms.

4.1 | Assumptions

We first assume a non-degenerate condition on the separation among block means (i.e. core tensor) to ensure the identifiability for clustering:

$$\Delta_k^2 = \Delta_k^2(\mathcal{S}) := \min_{i_1 \neq i_2} \left\| (\mathcal{M}_k(\mathcal{S}))_{i_1:} - (\mathcal{M}_k(\mathcal{S}))_{i_2:} \right\|_2^2 > 0. \tag{12}$$

In particular, we set $\Delta_k^2 = \infty$ if $r_k = 1$. Roughly speaking, this condition means all mode-k slices of the core tensor $\mathcal{S}$ are distinct; otherwise the number of blocks, $(r_1, \ldots, r_d)$, should be reduced to smaller. Generally speaking, clustering is easier to achieve when the separation is larger or the noise level is smaller. Recall σ is the sub-Gaussian norm of the noise distribution in Assumption 1. Therefore, we define the SNR as

$$\mathrm{SNR} := \Delta_{\min}^2 / \sigma^2, \quad \text{where } \Delta_{\min}^2 := \min_{k \in [d]} \Delta_k^2. \tag{13}$$

Remark 3. It is worth mentioning that an identifiable core in tensor block model may have degenerate ranks, that is, $\mathrm{rank}(\mathcal{M}_k(\mathcal{S})) < r_k$. This is significantly different from most literatures on low-rank tensor decomposition, where the singular value gap $\lambda_{r_k}(\mathcal{M}_k(\mathcal{S}))$ was assumed to be sufficiently large (Richard & Montanari, 2014; Xia et al., 2021; Zhang & Xia, 2018). For example, consider the following core tensor $\mathcal{S}$ representing 2-by-2-by-2 clusters:

$$\mathcal{S}_{1::} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad \mathcal{S}_{2::} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Note that the rank of $\mathcal{M}_k(\mathcal{S})$ is 1, which is smaller than the number of clusters at mode-k. The above example has non-zero separation $\Delta^2_{\min}(\mathcal{S}) = 16$, so the two clusters on each mode are still identifiable.

Since permuting the cluster labels does not alter the clustering result (e.g. naming $\{1, 3\}$, $\{2, 4\}$ as Cluster I/II is equivalent to naming them as Cluster II/I), the cluster label vector z_k on mode-k is estimable only up to a permutation of cluster labels. Given the initialisation label $z_k^{(0)}$, let $\pi_k^{(0)} : [r_k] \to [r_k]$ be the optimal permutation that minimises the mismatches between $z_k^{(0)}$ and z_k, that is,

$$\pi_k^{(0)} := \arg\min_{\pi \in \Pi_{r_k}} \frac{1}{p_k} \sum_{j=1}^{p_k} \mathbb{I}\left\{ (z_k^{(0)})_j \neq (\pi \circ z_k)_j \right\}, \quad \text{where } (\pi \circ z_k)_j := \pi((z_k)_j).$$

Let t be the iteration index in HLloyd algorithm. We define $h_k^{(t)}$ as the mode-k *misclassification rate* at the t th iteration of HLloyd algorithm:

$$h_k^{(t)} := \frac{1}{p_k} \sum_{j=1}^{p_k} \mathbb{I}\left\{ (z_k^{(t)})_j \neq (\pi_k^{(0)} \circ z_k)_j \right\}. \tag{14}$$

We also impose the following 'balanced cluster size' assumption for technical convenience. Such an assumption is widely used in the literature of mixture model clustering (Gao & Zhang, 2022; Löffler et al., 2021; Wu & Yang, 2020).

Assumption 2. There exists universal positive constants $0 < \alpha < 1 < \beta$ such that

$$\alpha p_k / r_k \leq \left| \{ j \in [p_k] : (z_k)_j = a \} \right| \leq \beta p_k / r_k, \qquad \forall a \in [r_k],\ k \in [d], \tag{15}$$

where $|\cdot|$ is the cardinality of a given set.

4.2 | Algorithmic theoretical guarantees

Now we are in a position to establish the theoretical guarantees for HLloyd and HSC. In order to prove our main theoretical result in Theorem 4, we introduce the following more convenient measure of misclassification loss in addition to the classification error rate $h_k^{(t)}$:

$$l_k^{(t)} := \frac{1}{p_k} \sum_{j=1}^{p_k} \left\| (\mathcal{M}_k(\mathcal{S}))_{(z_k^{(t)})_j :} - (\mathcal{M}_k(\mathcal{S}))_{(\pi_k^{(0)} \circ z_k)_j :} \right\|_2^2. \tag{16}$$

The following lemma establishes a relationship between $h_k^{(t)}$ and $l_k^{(t)}$, which implies that it suffices to bound $l_k^{(t)}$ in order to develop the target upper bound for $h_k^{(t)}$.

Lemma 1. *Define $h_k^{(t)}$ and $l_k^{(t)}$ as in (14) and (16). Then, $h_k^{(t)} \leq l_k^{(t)} / \Delta_k^2$.*

Our first result is the local convergence of HLloyd.

Theorem 2 (Local convergence of HLloyd). *Suppose Assumptions 1 and 2 hold. Let $\{z_k^{(0)}\}_{k=1}^d$ be the initialisation of HLloyd algorithm and $\{z_k^{(t)}\}_{k=1}^d$ be the estimations at step t. Assume the initialisation satisfies*

$$l_k^{(0)} \le c\Delta_{\min}^2 / r_k, \tag{17}$$

and the SNR satisfies

$$\Delta_{\min}^2/\sigma^2 \ge \frac{C\overline{p}r_*^2\overline{r}\log\overline{p}}{p_*}. \tag{18}$$

Then with probability at least $1-\exp(-c\underline{p})-\exp\left(-\frac{cp_*}{r_*\overline{p}}\frac{\Delta_{\min}^2}{\sigma^2}\right)$, *for all* $t \ge 0$,

$$l_k^{(t)} \le C\sigma^2\exp\left(-\frac{cp_*\Delta_{\min}^2}{r_*\overline{p}\sigma^2}\right)+\frac{\Delta_{\min}^2}{2^t}, \qquad \forall k\in[d]. \tag{19}$$

Remark 4. Recently, Gao and Zhang (2022) provided the convergence analysis of iterative algorithms for the inference problems with a single discrete structure, including the Lloyd algorithm for Gaussian mixture model as a special instance. Their techniques do not apply to our problem as the tensor block model admits multiple discrete structures (i.e. clusters in each mode). One specific technical challenge for proving Theorem 2 is to characterise the impact to misclassification rate in mode-k by those in the other $d-1$ modes. See more discussions in the proof sketch in Section 9.

The inequality (19) and Lemma 1 together imply that when $T \ge 2\lceil\log\overline{p}\rceil$,

$$h_k^{(T)} \le C\cdot \mathrm{SNR}^{-1}\exp\left(-\frac{cp_*\mathrm{SNR}}{r_*\overline{p}}\right)+2^{-T} < \frac{1}{p_k}. \tag{20}$$

Therefore, HLloyd algorithm guarantees exact recovery of clustering under the SNR condition (18).

We also need a good initialisation satisfying (17) in order to apply Theorem 2. Our initialisation accuracy is guaranteed by the proposed HSC (Algorithm 2).

Theorem 3 (Upper bound on misclassification rate of HSC). *Suppose Assumptions* 1 *and* 2 *hold and each entry of* $\mathcal{E}$ *has equal variance. If the SNR satisfies*

$$\Delta_{\min}^2/\sigma^2 \ge CM\left(\overline{p}r_*^2\overline{r}\log\overline{p}/p_* + r_*\overline{r}/p_*^{1/2}\right), \tag{21}$$

then, with probability at least $1-C\exp(-c\underline{p})$, $\forall k\in[d]$,

$$l_k^{(0)} \le CM\cdot\sigma^2\cdot(r_{-k}/p_*)\left(r_*+\overline{pr}^2+p_*^{1/2}\overline{r}\right) \le c\Delta_{\min}^2/r_k,$$
$$h_k^{(0)} \le CM\cdot(r_{-k}/p_*)\left(r_*+\overline{pr}^2+p_*^{1/2}\overline{r}\right)\sigma^2/\Delta_{\min}^2.$$

The combination of Theorems 2 and 3 yields the following main result of our paper.

Theorem 4. *Denote* $\{z_k^{(t)}\}_{k=1}^d$ *as the membership vectors in the iteration* t *of HLloyd algorithm with* $\{z_k^{(0)}\}_{k=1}^d$ *being the output of HSC. Under the same conditions of Theorem* 3, *with probability at*

least $1-\exp(-c\underline{p})-\exp\left(-\frac{cp_*}{4r_*\overline{p}}\frac{\Delta_{\min}^2}{\sigma^2}\right)$, *we have exact recovery of* $\{z_k\}_{k=1}^d$ *when* $T \geq 2\lceil \log \overline{p} \rceil$; *that is, there exist a set of permutations* $\{\pi_k\}_{k=1}^d$ *such that*

$$z_k^{(T)} = \pi_k \circ z_k, \qquad \forall k = 1, \ldots, d.$$

In particular, if the numbers of clusters r_k *are fixed,* $T \geq \lceil 2\log p \rceil$, $p_1 \asymp \cdots \asymp p_d \asymp p$, *and the SNR satisfies*

$$\Delta_{\min}^2/\sigma^2 \geq C\left(p^{-d/2} \vee p^{-(d-1)}\log p\right), \tag{22}$$

then $z_k^{(T)}$ *achieves exact clustering with high probability.*

In comparison, if one only applies the proposed HSC algorithm without HLloyd refinement, the same condition (22) would only guarantee consistent clustering according to Theorem 3.

Remark 5 (Comparison with theory for matrix spectral method). A key intermediate step for the proof of Theorem 3 is to evaluate the estimation error for the projected tensor observation:

$$\|\tilde{\mathcal{Y}} - \mathcal{X}\|_{\mathrm{F}}^2, \qquad \text{where} \quad \tilde{\mathcal{Y}} := \mathcal{Y} \times_1 \hat{\mathbf{U}}_1\hat{\mathbf{U}}_1^\top \times \cdots \times_d \hat{\mathbf{U}}_d\hat{\mathbf{U}}_d^\top \quad \text{for } \hat{\mathbf{U}}_k \text{ in (9).} \tag{23}$$

In matrix case ($d = 2$), the estimation error bound can be simply derived by the algebraic property of SVD. However, for high-order tensors ($d \geq 3$), such an error bound was usually established under strong singular value gap condition (Zhang & Xia, 2018) that does not hold in tensor block model. To overcome this issue, we develop the following new singular-value-gap-free bound on tensor estimation for $\tilde{\mathcal{Y}}$.

Proposition 1 (A singular-value-gap-free tensor estimation error bound). *Suppose* $\mathcal{Y} = \mathcal{X} + \mathcal{E} \in \mathbb{R}^{p_1\times\cdots\times p_d}$, $\mathcal{X}$ *has Tucker-rank* $(r_1, \ldots, r_d)$, *and* $\mathcal{E}$ *satisfies Assumption 1 with equal variance* $\mathrm{Var}(\mathcal{E}_{j_1,\ldots j_d}) = \sigma^2$, $\forall (j_1, \ldots, j_d) \in [p_1]\times\cdots\times[p_d]$. *Let* $\tilde{\mathcal{Y}}$ *be defined according to* (23). *Then, with probability at least* $1 - C\exp(-c\underline{p})$,

$$\|\tilde{\mathcal{Y}} - \mathcal{X}\|_{\mathrm{F}}^2 \leq C\sigma^2\left(p_*^{1/2}\overline{r} + \overline{p}\overline{r}^2 + r_*\right).$$

The proof sketch of Theorem 3 and Proposition 1 are presented in Section 9.

Remark 6 (Heteroskedastic errors). The guarantee for local convergence in Theorem 2 applies to the general sub-Gaussian noise under Assumption 1, which covers both GTBMs and STBMs. Our theoretical guarantee for initialisation (Theorem 3) is established under an additional homoscedasticity condition, which excludes some interesting STBMs with Bernoulli distributions. Nevertheless, our simulation study in Section 6 illustrates that our algorithm still performs well in practice for a wide range of STBMs. The theoretical error bound for heteroskedastic HSC will be left as future work.

In addition to recovering cluster memberships in the tensor block model, another important task is to recover the block means $\mathcal{S}$, or equivalently, to denoise the observed tensor $\mathcal{Y}$ and obtain an estimate of $\mathcal{X}$. Given the estimates of labels $z_1^{(T)}, \ldots, z_d^{(T)}$, a natural estimator for $\mathcal{X}$ is the aggregated mean in each estimated block, that is,

$$\hat{\mathcal{X}}_{j_1,\ldots j_d} = \mathrm{Average}\left(\left\{\mathcal{Y}_{j_1',\ldots,j_d'} : (z_k^{(T)})_{j_k'} = (z_k^{(T)})_{j_k}, j_k' \in [p_k], \forall k \in [d]\right\}\right).$$

We have the following guarantee for $\hat{\mathcal{X}}$.

Theorem 5 (Upper bound of estimation error for HLloyd + HSC). *Under the same conditions of Theorem 3, we have with probability at least* $1-\exp(-cr_*)-\exp(-c\underline{p})-\exp\left(-\frac{cp_*}{4r_*\bar{p}}\frac{\Delta_{\min}^2}{\sigma^2}\right)$ *that*

$$\left\|\hat{\mathcal{X}}-\mathcal{X}\right\|_{\mathrm{F}}^2 \le C\sigma^2 r_*. \tag{24}$$

Remark 7 (Comparison with HOOI). Recall the tensor block model (3) naturally admits a Tucker low-rank structure. A tempting strategy is to apply HOOI on $\mathcal{Y}$ to estimate $\mathcal{X}$. By theorem 1 of Zhang and Xia (2018), the HOOI estimator $\hat{\mathcal{X}}_{\mathrm{HOOI}}$ achieves the following statistical rate under a strong singular gap condition:

$$\mathbb{E}\left\|\hat{\mathcal{X}}_{\mathrm{HOOI}}-\mathcal{X}\right\|_{\mathrm{F}}^2 \le C\sigma^2\left(r_*+\sum_{k=1}^d p_k r_k\right). \tag{25}$$

Compared to Theorem 5, (25) has an additional dimension-dependent term. The intuition behind this phenomenon is that HOOI only fits a low-rank model while fails to capture the discrete structure in the tensor block model. This theoretical result is also supported by our numeric experiments in Section 6.1.

5 | STATISTICAL AND COMPUTATIONAL TRADE-OFFS

In this section, we study the statistical and computational limits of high-order clustering in the tensor block model (3). We assume the distribution of $\mathcal{E}_{j_1,\dots j_d}$ is i.i.d. Gaussian.

We specifically focus on the following parameter space,

$$\Theta\left(\{\Delta_k\}_{k=1}^d,\alpha,\beta\right)=\left\{(\mathcal{S},z_1,\dots,z_d):\begin{array}{l}\mathcal{S}\in\mathbb{R}^{r_1\times\cdots\times r_d},\quad \Delta_k(\mathcal{S})\ge\Delta_k,\quad z_k\in[r_k]^{p_k}\\ \alpha\frac{p_k}{r_k}\le\left|\{j\in[p_k]:(z_k)_j=a\}\right|\le\beta\frac{p_k}{r_k},\ \forall a\in[r_k],k\in[d]\end{array}\right\}. \tag{26}$$

Here, the constraints in the parameter space (26) correspond to assumptions in Section 4.1. We further introduce the following parameter regime

$$\Delta_{\min}^2/\sigma^2=p^\gamma \quad\text{and}\quad p_1\asymp\cdots\asymp p_d\asymp p. \tag{27}$$

5.1 | Statistical limit

The following theorem establishes the SNR lower bound for exact label recovery, which reveals the statistical limit of high-order clustering in the tensor block model.

Theorem 6 (Statistical lower bound). *Consider the tensor block model (3). Suppose* $r_k=o(p_k^{1/3})$, *and* $\frac{\Delta_k^2}{\sigma^2}\frac{p_{-k}}{r_{-k}}<c_0$ *for some constant* $c_0>0$. *Then, for any estimator* $\hat{z}_k$,

$$\sup_{(\mathcal{S},z_1,\dots,z_d)\in\Theta}\mathbb{E}\min_{\pi_k\in\Pi_{r_k}}\sum_{j=1}^{p_k}\mathbb{I}\{(\hat{z}_k)_j\ne(\pi_k\circ z_k)_j\}\ge 1. \tag{28}$$

Theorem 6 suggests the impossibility of exact clustering under the parameter regime (27), when r_k 's are constants and $\gamma < -(d-1)$. Furthermore, it was shown that the MLE (4), while being computationally intractable, achieves consistent clustering given $\gamma > -(d-1)$ (Wang & Zeng, 2019). Therefore, $\gamma_{\text{stat}} := -(d-1)$ serves as the statistical limit for the clustering in tensor block model: when $\gamma > \gamma_{\text{stat}}$, there exists an algorithm that can successfully recover the clustering labels; when $\gamma < \gamma_{\text{stat}}$, all algorithms, regardless the computational complexity, fail to do so.

In comparison, according to the discussion in Section 4.2, the combination of HSC and HLloyd algorithms achieves exact clustering when $\gamma > -d/2 =: \gamma_{\text{comp}}$, which seems more stringent than the statistical limit γ_{stat} for $d \geq 3$. However, we should point out that, unlike MLE, our algorithm (HSC and HLloyd) is polynomial-time implementable. In the next section, we show that $\gamma > \gamma_{\text{comp}}$ is indeed necessary for any polynomial-time algorithm to succeed.

5.2 | Computational limit

In this section, we establish the computational limit for high-order clustering under model (3). We first introduce the HPC detection problem and its hardness conjecture, which are building blocks for the main results on computational limit.

A d-hypergraph can be seen as an order-d extension of regular graph. In a d-hypergraph $G = (V(G), E(G))$, each hyperedge $e \in E$ includes a set of d different vertices in V. Define $\mathcal{G}_d(N, 1/2)$ as the Erdős-Rényi d-hypergraph with N vertices, where each hyperedge $(i_1, \ldots, i_d)$ is independently included in E with probability $1/2$. Also we define $\mathcal{G}_d(N, 1/2, \kappa)$ as the HPC model with the clique size κ. To generate $G \sim \mathcal{G}_d(N, 1/2, \kappa)$, we sample a random hypergraph from $\mathcal{G}_d(N, 1/2)$, pick κ vertices uniformly at random from $[N]$, denote them as K, and connect all hyperedges e if all vertices of e are in K. The HPC detection can be formulated as the following hypothesis testing problem:

$$H_0 : G \sim \mathcal{G}_d(N, 1/2) \quad \text{v.s.} \quad H_1 : G \sim \mathcal{G}_d(N, 1/2, \kappa). \tag{29}$$

We consider the following version of HPC detection conjecture, which was introduced and studied in the literature (Brennan & Bresler, 2020; Luo & Zhang, 2020; Zhang & Xia, 2018).

Conjecture 1 (HPC detection conjecture). *Suppose $d \geq 2$ is a fixed integer. Suppose*

$$\limsup_{N\to\infty} \log \kappa / \log \sqrt{N} \leq 1 - \varepsilon \quad \text{for any } \varepsilon > 0.$$

Then, for any sequence of polynomial-time tests $\{\phi\}_N : G \to \{0, 1\}$, $\liminf_{N\to\infty} \mathbb{P}_{H_0}(\phi(G) = 1) + \mathbb{P}_{H_1}(\phi(G) = 0) > 1/2$.

Zhang and Xia (2018) observed that spectral method solves HPC detection efficiently if $\kappa = \Omega(\sqrt{N})$ but fails when $\kappa = N^{1/2-\varepsilon}$ for any $\varepsilon > 0$. Recently, it has also been shown that many classes of powerful algorithms, including metropolis algorithms and low-degree polynomial algorithms, fail to solve the HPC detection problem in polynomial time under the conjectured hard regime (Brennan & Bresler, 2020; Luo & Zhang, 2022). Several open questions on HPC detection—in particular, whether HPC detection is equivalently hard as PC detection—are discussed in Luo and Zhang et al. (2020).

With the HPC detection hardness conjecture, we have the following computational lower bound for high-order clustering in the tensor block model.

Theorem 7 (Computational lower bound). *Consider the tensor block model* (3) *under the parameter regime* (27) *and Conjecture* 1. *If* $\gamma < -d/2 =: \gamma_{\mathrm{comp}}$, *then for any polynomial-time estimator* $(\hat{z}_1, \ldots, \hat{z}_d)$, *we have*

$$\liminf_{p\to\infty} \sup_{(S,z_1,\ldots,z_d)\in\Theta} \mathbb{P}\left(\exists k \in [d] \text{ s.t. } \min_{\pi\in\Pi_{r_k}} \sum_{j=1}^{p_k} \mathbb{I}\left\{(\hat{z}_k)_j \neq (\pi\circ z_k)_j\right\} \geq 1 \right) \geq 1/2.$$

Combining Theorems 2,6, and 7, we have finished the proof for the informal statement of Theorem 1 and established the phase transition diagram of Figure 2 in the introduction section. These results render the whole picture of the statistical and computational limits of the high-order clustering in the tensor block model.

6 | NUMERICAL STUDIES

In this section, we first study the performance of the proposed HLloyd and HSC algorithms. Then we compare the proposed algorithms with other state-of-the-art methods. Unless otherwise noted, we consider order-d tensor block models with $p_1 = \cdots = p_d = p$, $r_1 = \cdots = r_d = r$ and balanced cluster sizes across all modes throughout the simulations. In each experiment, we report the averaged statistics and the standard error across 100 replications.

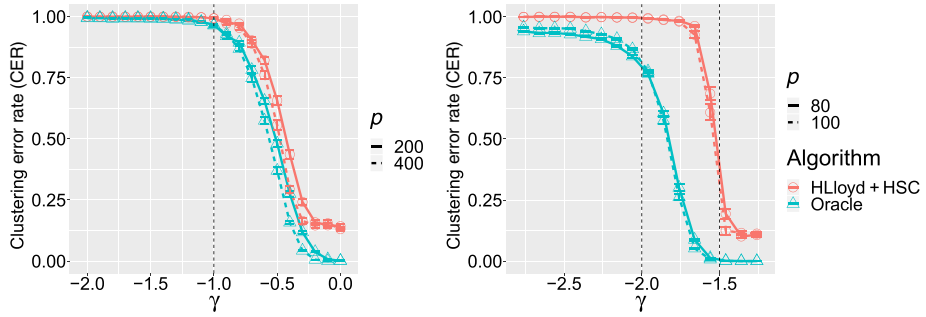
6.1 | Properties of HLloyd and HSC

We first study the clustering performance of the proposed HLloyd and HSC algorithms using clustering error rate (CER). The CER is calculated using the disagreement between estimated and true partitions, that is, one minus adjusted random index (Milligan & Cooper, 1986). A lower CER implies a better clustering and CER = 0 means exact clustering. In the first four experiments, we consider the Gaussian tensor block models with variance $\sigma^2 = 1$, and in the fifth experiment, we study the performance of the proposed algorithms on stochastic tensor block models.

6.1.1 | Statistical and computational phase transition

The first experiment investigates the phase transition of high-order clustering with respect to SNR. We perform the proposed polynomial-time HLloyd algorithm with HSC initialisation on both matrix and order-3 tensor block models. We set $r = 5, p \in \{200, 400\}$ for the matrix case, and $p \in \{80, 100\}$ for the order-3 tensor case. We assess the SNR phases by plotting CER as a function of the signal strength γ for which $\Delta_{\min}^2 = \mathcal{O}(p^\gamma)$. The performance of our estimator is compared to MLE (4), that is, the global optimum. Since MLE is NP-hard to compute, we approximate the global optimum using oracle initialisation refined by HLloyd and call the estimate the 'oracle estimate'. The oracle initialisation is specified as the perturbed ground truth contaminated by 20% random labelling.

Figure 3 shows the SNR phases in the clustering problem for matrices and for order-3 tensors. We find that, in the matrix case, our method and oracle estimates undergo similar SNR phase transitions. A flat, high clustering error is observed for both estimates when $\gamma < -1$, and the error



(a) Matrix Clustering (b) Tensor Clustering

FIGURE 3 Signal-to-noise ratio phase transition for clustering in tensor block model under the setting $r = 5, \Delta_{\min}^2 = \mathcal{O}(p^\gamma)$. The error bar on each point represents the SE. (a) Matrix clustering $p = 200, 400, \gamma \in (-2, 0)$; (b) Tensor clustering $d = 3, p = 80, 100, \gamma \in (-2.8, -1.2)$ [Colour figure can be viewed at wileyonlinelibrary.com]

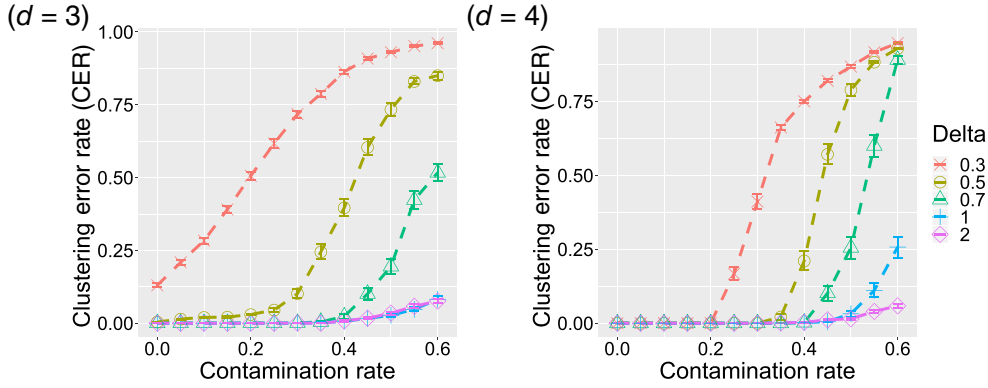


FIGURE 4 Impact of initialisation to the performance of high-order Lloyd algorithm. The clustering error rate and the SE are plotted against the contamination rate from 0 to 0.6 under the settings $\Delta_{\min} \in \{0.3, 0.5, 0.7, 1, 2\}, p = 50, r = 5$. [Colour figure can be viewed at wileyonlinelibrary.com]

immediately decreases as $\gamma > -1$. Note that our theory in Section 4.2 has implied an optimal SNR $\mathcal{O}(p^{-1})$ for order-2 cases, and this critical ratio is indeed achieved by the proposed algorithm. In contrast to the matrix case, the order-3 tensor clustering reveals a striking gap between HLloyd and oracle estimates. In particular, the phase transition occurs around $\gamma = -2$ for oracle estimates, whereas $\gamma = -1.5$ for HLloyd estimates. This gap reflects the statistical-computational gap $p^{-(d-1)} \ll \text{SNR} \ll p^{-d/2}$ for clustering in tensor block model when $d \geq 3$, which corroborates our theoretical results in Section 5.

6.1.2 | Impact of initialisation to HLloyd

The second simulation examines the impact of initialisation to the performance of the HLloyd. We consider CER for $p = 50, r = 5, d \in \{3, 4\}$, and $\Delta_{\min} \in \{0.3, 0.5, 0.7, 1, 2\}$. Given a contamination rate ε, we generate initialisation labels by randomly shuffling $100\varepsilon\%$ labels of the ground truth and then run HLloyd. Figure 4 shows that the clustering error decreases as the signal strength

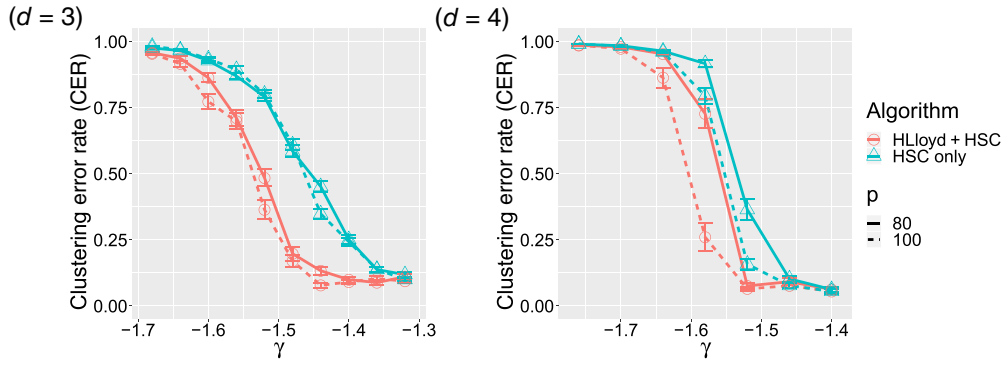


FIGURE 5 Clustering of high-order spectral clustering and spectral initialisation with high-order Lloyd refinement, for $r = 5$, $p \in \{80, 100\}$, and varying γ [Colour figure can be viewed at wileyonlinelibrary.com]

$\Delta_{\min}$ increases or the contamination rate decreases in both $d = 3$ and 4 settings. This shows that stronger signal and better initialisation enhance the clustering performance of the HLloyd algorithm. The experiment also indicates that a proper initialisation is crucial for the success of HLloyd. We will show next that the proposed HSC algorithm can achieve a proper initialisation.

6.1.3 | Clustering via HLloyd + HSC algorithms

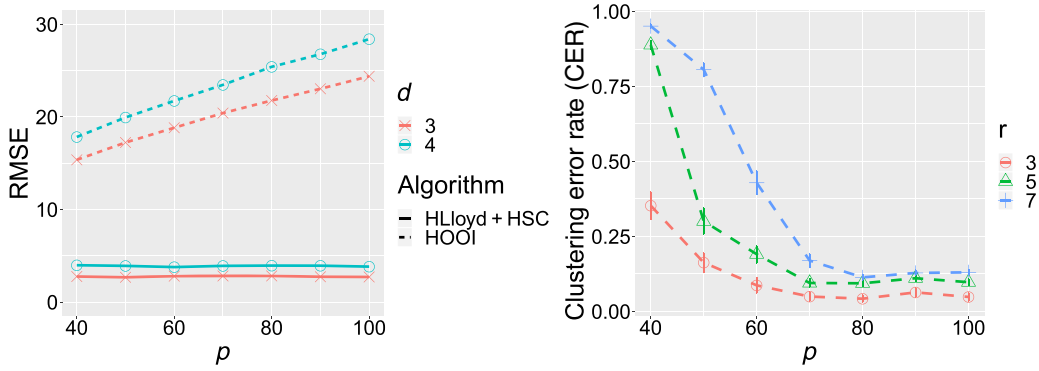
In the third simulation, we assess the clustering accuracy for the proposed approaches: HSC-only algorithm and the combined algorithm (HLloyd + HSC). We consider the settings $r = 5$, $p \in \{80, 100\}$, and $\Delta_{\min}^2 = \mathcal{O}(p^{-\gamma})$ for a range of γ. Figure 5 shows that both approaches achieve nearly exact recovery as the signal level increases in the tensor block model. For a wide range of settings, given the initialisation by HSC, HLloyd can greatly improve the HSC-only algorithm and achieves more accurate clustering. This confirms our theoretical results that HLloyd algorithm effectively boosts the clustering performance of HSC.

6.1.4 | Tensor estimation via HLloyd + HSC algorithms

We use the following experiment to compare the tensor estimation errors of HLloyd + HSC and HOOI. Figure 6a shows the root mean squared error (RMSE) $\|\hat{\mathcal{X}} - \mathcal{X}\|_{\mathrm{F}}$ for $p \in \{40, 50, \ldots, 100\}$, $r = 2$, $d \in \{3, 4\}$, and $\Delta_{\min} = 2$. As p increases, the tensor estimation error of HLloyd + HSC is almost flat over the range of p. This matches our theoretical results in Theorem 5 that the tensor estimation bound of the proposed algorithm is free of dimension p. In contrast, the estimation error of HOOI grows almost linearly with respect to p. This demonstrates the benefit of HLLoyd + HSC over previously tensor algorithm HOOI on tensor estimation in tensor block model.

6.1.5 | Simulations on stochastic tensor block models

We also assess our algorithm on stochastic tensor block models. Let $r \in \{3, 5, 7\}$, $p \in \{40, 50, \ldots, 100\}$, $d = 3$. The memberships $z_1, \ldots, z_d$ are generated in the same way as



(a) Tensor Estimation Error (b) Clustering in STBM

FIGURE 6 (a) Comparison of tensor estimation error of HLloyd + HSC and high-order orthogonal iteration for $r = 2, p \in \{40, 50, \ldots, 100\}$; (b) Clustering of high-order Lloyd (HLloyd) with spectral initialisation for stochastic tensor block model with $d = 3$, $r \in \{3, 5, 7\}$, and $p \in \{40, 50, \ldots, 100\}$ [Colour figure can be viewed at wileyonlinelibrary.com]

before. Each entry in the core tensor $\mathcal{S}$, which encodes the connection probability in each cluster, is generated uniformly at random from $[0, 0.1]$. Then we generate data tensor $\mathcal{Y}$ with independent Bernoulli entries $\mathcal{Y}_{(z_1)_{j_1},\ldots,(z_d)_{j_d}} \sim \text{Bernoulli}(\mathcal{S}_{(z_1)_{j_1},\ldots,(z_d)_{j_d}})$. We consider low-probability because the binary tensors in practice often have low average connection probability (see estimated connection probability in real data in the forthcoming Figures 11 and 12). The simulation results are provided in Figure 6b. We can see as p increases, the clustering error of the proposed algorithm decreases as expected. The results validate our main results in Section 4. In particular, our algorithm performs well under a general class of sub-Gaussian noises in tensor block model.

6.2 | Comparison with other algorithms

In this section, we compare HLloyd + HSC with two other classic tensor clustering algorithms:

(a) HOSVD + Kmeans: apply high-order SVD (HOSVD) on $\mathcal{Y}$ (De Lathauwer et al., 2000a; Ghoshdastidar & Dukkipati, 2015), then perform k-means on the outcome factors of HOSVD;
(b) CP + Kmeans: apply CANDECOMP/PARAFAC (CP) decomposition on $\mathcal{Y}$ (Carroll & Chang, 1970), then perform k-means on the outcome factors of CP decomposition.

In the first three simulations, we compare these algorithms under the Gaussian tensor block model; in the fourth simulation, we compare these algorithms under the stochastic tensor block model.

In the first simulation, we let $r = 5, p = 80$ and $\Delta^2_{\min} \asymp p^\gamma$ with varying γ. The comparison results in Figure 7 show that HLloyd greatly improves the other two methods for most of γ considered. When $d = 4$ and γ is large, the HOSVD-based clustering method has a similar performance to HLloyd, and both of them achieve better accuracy than the CP-decomposition-based algorithm. However, the performance of CP-based algorithm becomes even worse as d increases, which is

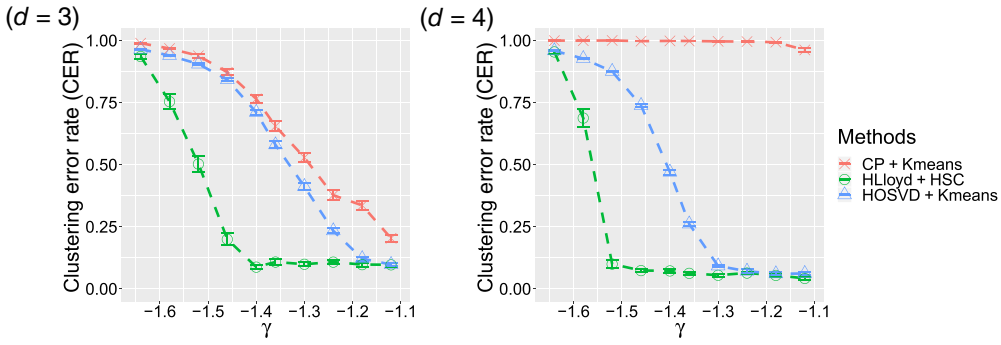


FIGURE 7 Comparison of HLloyd + HSC with HOSVD-/CP-decomposition-based clustering methods for $p = 80, r = 5$ and varying γ [Colour figure can be viewed at wileyonlinelibrary.com]

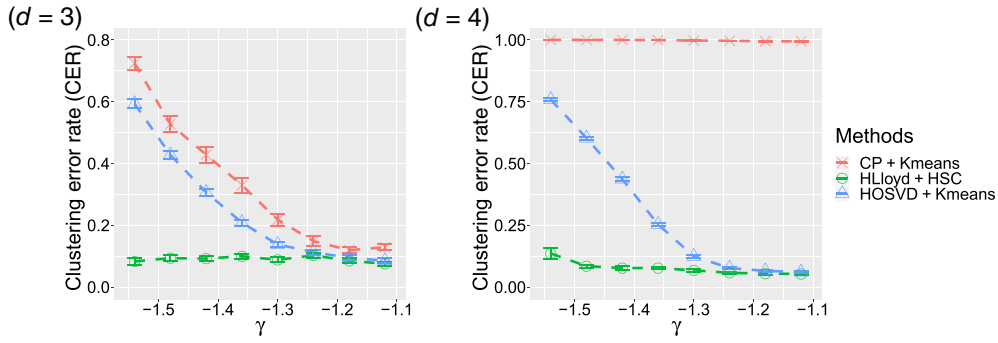


FIGURE 8 Comparison of HLloyd + HSC with HOSVD-/CP-decomposition-based clustering methods with different numbers of clusters along each mode [Colour figure can be viewed at wileyonlinelibrary.com]

because the CP decomposition can hardly capture dense core tensor structure in the tensor block model.

Next, we compare the performance of these algorithms when the number of clusters along each mode differs. We set p and $\Delta_{\min}$ to be the same as the previous experiment and $r_1 = 3, r_2 = 5, r_3 = 7$ when $d = 3$; $r_1 = 3, r_2 = 5, r_3 = 7, r_4 = 9$ when $d = 4$. Figure 8 shows the averaged CER for these methods. We find that HLloyd + HSC still outperforms the other two algorithms and its CER SE is low and stable for all the ranges of γ considered.

Third, we examine the effect of imbalanced cluster size on the performance of these algorithms. In this setting, we let each mode have two clusters, that is, $r = 2$, and the proportion of cluster 1 in each mode is ξ. We let $p, \Delta_{\min}$ be the same as before with fixed $\gamma = -1.4$. Figure 9 illustrates the clustering performance of these algorithms as one gradually increases ξ from 0.05 to 0.5 and it can be observed that the proposed algorithm is more robust against the imbalanced cluster size than the other two baseline methods.

Finally, we compare the performance of these algorithms in the stochastic tensor block models. Consider the setting $r = 5, p \in \{40, 50, \ldots, 100\}$ and every entry in the core tensor $\mathcal{S}$ is generated uniformly at random from $[0, 0.1]$. The comparison result is given in Figure 10. We can see that in both settings $d = 3$ and $d = 4$, the proposed algorithm performs much better than the other two algorithms.

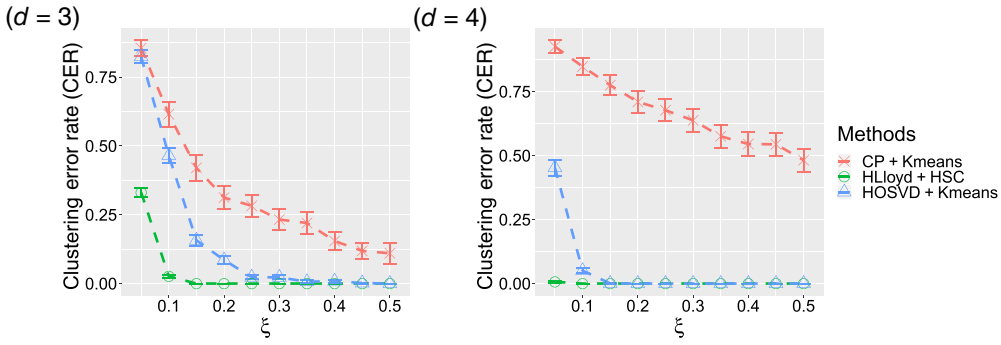


FIGURE 9 Comparison of HLloyd + HSC with HOSVD-/CP-decomposition-based clustering methods when cluster sizes are imbalanced [Colour figure can be viewed at wileyonlinelibrary.com]

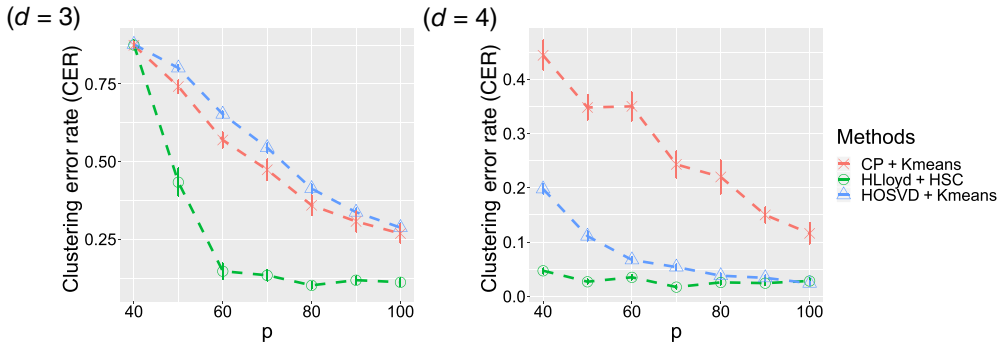


FIGURE 10 Comparison of HLloyd+HSC with HOSVD-/CP-decomposition-based clustering methods in stochastic tensor block model [Colour figure can be viewed at wileyonlinelibrary.com]

7 | REAL DATA ANALYSIS

7.1 | Flight route network

In the first application, we study a subset of the worldwide air routes networks[1] based on 66,765 global flight routes from 568 airlines and 3409 airports. We focus on the top 50 airports with the highest numbers of flight routes and obtain an order-3 tensor $\mathcal{Y}$ of size $39 \times 50 \times 50$, where the entry $\mathcal{Y}_{ijk}$ equals to 1 if there exists a flight route from airport j to airport k in airline i and equals to 0 otherwise. We perform tensor clustering using HLloyd initialised by HSC with rank $(r_1, r_2, r_2) = (5, 5, 5)$. Here, r_1 and r_2 are chosen from $\{3, 4, 5, 6\}$ that they (i) do not result in a singleton in any cluster and (ii) minimise the following BIC for block models (Wang & Zeng, 2019),

$$\mathrm{BIC}(r_1, \ldots, r_d) = p_* \log(\|\hat{\mathcal{X}} - \mathcal{Y}\|_{\mathrm{F}}^2) + \left(r_* + \sum_{k=1}^{d} p_k \log r_k \right) \log p_*. \tag{30}$$

[1]Available at: https://openflights.org/data.html\#route.

TABLE 1 Clustering of airlines based on the global flight routes network

	Airlines
Cluster 1	AA, UA, US (USA)
Cluster 2	BA, AY, IB (Europe)
Cluster 3	SU, AB, AI, AM, NH, AC, AS, FL, DE, etc (Mixture)
Cluster 4	CA, MU, CZ, HU, 3U, ZH (China)
Cluster 5	AF, AZ, KL (Europe), DL (USA)

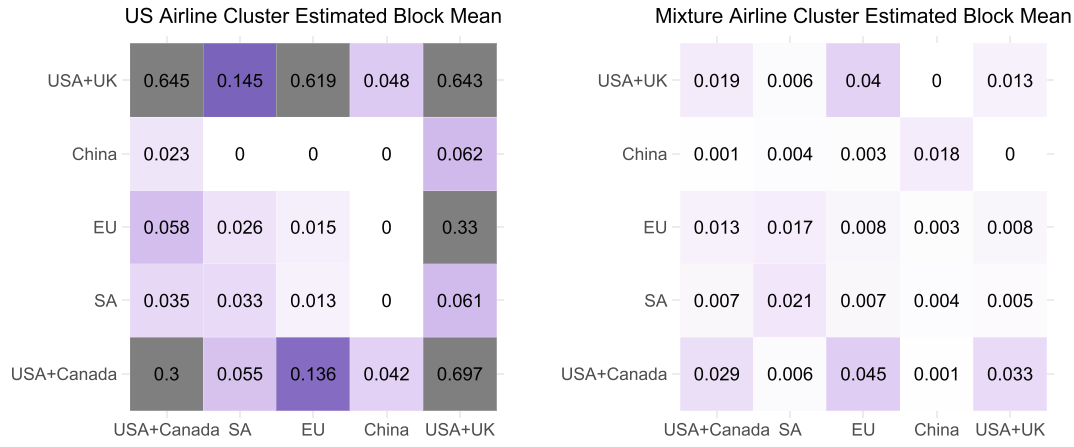
TABLE 2 Clustering of airports based on global flight route network

	Airports
Cluster 1	PHX, SFO, LAX, EWR, IAH, ATL, DEN, LAS, YYZ, MEX (North America)
Cluster 2	TPE, HKG, DEL, KUL, SIN, BKK, ICN, DME, etc (Southeast Asian)
Cluster 3	BRU, FRA, DUS, MUC, MAN, AMS, BCN, MAD, FCO, ZRH (Europe)
Cluster 4	PEK, CAN, XIY, KMG, HGH, CKG, CTU, PVG(China)
Cluster 5	MIA, DEW, PHL, JFK, ORD, CLT (United States), LHR (United Kingdom)

Tables 1 and 2 show the clustering results for airlines and airports, respectively. The clusters well capture the underlying geographic and traffic information. Meaningful regions, such as the United States, Europe, China, Southeast Asia, were identified in the airline clusters and/or airport clusters. We also find several mixture clusters: Airline Cluster 3 is a mixture cluster consisting relatively small airlines around the world; Airline Cluster 5 consists of three airlines from Europe and Delta Airlines from United States.

Figure 11 shows the estimated block mean $\hat{\mathcal{S}}$ corresponding to the US airline cluster (Airline Cluster 1) and the mixture airline cluster (Airline Cluster 3), respectively. The rows and columns of each matrix represent Airport Clusters 1–5 listed in Table 2. The value in each matrix represents the connectivity of airports from each pair of clusters. We find that the US airline block mean matrix shows multiple zeros/small values among South Asian, European and Chinese airports. The matrix also shows high values among airports connected to United States. This result reveals that US airlines operate few flights between airports in Europe or Asia. In contrast, for the mixture airline cluster, the mean matrix has many non-zero but small entries. This reflects that this cluster consists of many small-scale airports scattered around the world.

We also compare the proposed HLloyd + HSC method with the HOSVD- and CP-decomposition-based clustering algorithms (see Section 6.2). We find the proposed method yields more meaningful results than the other two. For example, Table 3 collects the airline clustering result of the HOSVD-based clustering algorithm. Comparing Tables 1 and 3, the proposed algorithm has a better performance on clustering US and China airlines. More clustering results of the HOSVD-/CP-decomposition-based algorithms can be found in Appendix A.



(a) US airlines (b) Mixture airlines

FIGURE 11 Heatmaps of two matrix slices in the estimated block mean corresponding to the clusters of US airlines and mixture airlines in clustering results of global flight routes network. Here 'EU' is short for Europe and 'SA' is short for Southeast Asia. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Clustering of airlines of the high-order singular value decomposition-based algorithm

	Airlines
Cluster 1	AA, US (USA)
Cluster 2	UA (USA)
Cluster 3	3U (China), SU, AB, AI, AM, NH, AC, AS, FL, DE, etc (Mixture)
Cluster 4	CA, MU, CZ, HU, ZH (China)
Cluster 5	AF, AZ, KL (Europe), DL (United States)

TABLE 4 Clustering of US airlines based on the US flight route network

	Airlines
Cluster 1	UA(547), DL(633), WN(664), FL(420)
Cluster 2	AA(640), US(590)
Cluster 3	G4(6), B6(136), NK(134)

TABLE 5 Clustering of US airports based on US flight route network

	Airports
Cluster 1	PHX(137), LAX(148), DFW(130), ORD(151), PHL(110), JFK(108), CLT(159)
Cluster 2	BOS(91), MCI(68), SFO(76), CLE(52), CVG(50), EWR(73), etc
Cluster 3	OAK(28), MEM(46), HOU(47), SAT(48), IAD(57), HNL(22), SJC(33), etc

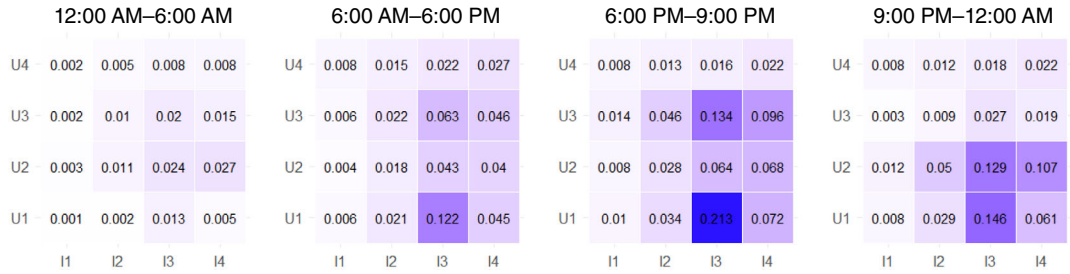


FIGURE 12 Clustering outcomes of online click-through data. Each matrix represents a mode-3 slice of $\hat{S}$ for different daily time cluster. The users clusters $U1, U2, U3, U4$ have 13, 25, 22, 40 members; the item clusters $I1, I2, I3, I4$ have 31, 11, 3, 5 members, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

Finally, we perform a similar analysis for the US flight routes network. Specifically, we pick the top 50 airports and nine airlines with the most traffic and present the clustering results based on proposed algorithm in Tables 4 and 5. We find the airports are grouped mainly based on their traffic sizes rather than their geographic information. We also find the three main airlines—United Airlines, Delta Airlines, American Airlines—are in different clusters, although they share similar numbers of flight routes. One possible reason is that these major airline companies are competitors, and the geometric distributions of their flight routes complement each other.

7.2 | Online click-through data

In this section, we illustrate the application of proposed algorithms to time-dependent user–item collaborative filtering on an e-commerce dataset. The goal is to identify user clusters and item clusters in a longitudinal study. Specifically, we use the users' online click-through behaviour data on *Taobao.com*, one of the most popular online shopping website in China. The data[2] of user–item interaction records are collected over eight consecutive days from 25 November to 2 December 2017. Due to the high dimensionality of the original dataset ($\approx 10^6$ users and $\approx 10^4$ item categories), we only select the most active 100 users and the most popular 50 items in our analysis. For the m th day, we construct a binary tensor $\mathcal{Y}_m \in \{0,1\}^{100\times 50\times 24}$, where the (i,j,k) th entry of $\mathcal{Y}_m$ equals to one if and only if the i th user has an interaction with the j th item (i.e. make a click) in the k th hour in that day. Let $\mathcal{Y} = \frac{1}{8}\sum_{m=1}^{8}\mathcal{Y}_i$ be the averaged observation, and we apply the proposed method to $\mathcal{Y}$. For the hour-mode, we set the number of clusters $r_3 = 4$, as we expect the behaviours might be separated into four time periods including dawn, morning, afternoon and evening; for the other two modes, we set the cluster numbers r_1, r_2 to be the largest possible values that do not result in a singleton in any cluster. This leads to $r_1 = 4$ and $r_2 = 4$.

Figure 12 shows the estimation of block means $\hat{S} \in [0,1]^{4\times 4\times 4}$, where $\hat{S}_{ijk}$ is the estimated probability that a user from the i th group click some item from the j th group in the daily time period of the k th group. The clustering results show that the time mode is well separated into four consecutive periods: 12am-6am (before dawn), 6:00 AM to 6:00 PM (daytime), 6:00 PM to 9:00 PM (evening) and 9:00 PM to 12:00 AM (late night). We find the average activity is extremely low in

[2]Available at: https://tianchi.aliyun.com/dataset/dataDetail?dataId=649.

Period I (before dawn) and high in Period III (evening), as most people are sleeping in Period I and could spend more time on online shopping after their daily work in Period III. For fixed user/item groups, we find some specific time-dependent behaviours. For example, by comparing the last heatplot with the previous two in Figure 12, we identify a particular group of users U3, whose activities almost vanish after 9:00 PM; this might corresponds to middle-aged people or seniors who sleep and rise early. On the other hand, U1 could be the group of young people since they have the most clicks in the late-night while being the least active in the early morning before 6:00 AM. Due to the lack of users/items features, we can not exactly verify our analysis. Nevertheless, the identified similarities among entries without external annotations illustrate the applicability of our method to clustering analysis.

8 | DISCUSSION

This paper develops a polynomial-time high-order clustering algorithm consisting of HLloyd iterations and HSC initialisation. Critical thresholds for SNR are established, revealing the intrinsic distinctions between (vector) clustering, (matrix) bi-clustering and (tensor) tri-clustering/high-order clustering. In particular, we provide both statistical and computational limits for high-order clustering in the tensor block model.

It is worth mentioning that while our focus is on clustering for tensor block models, the developed results are useful in a broader variety of applications. For example, the block-wise constant structure in our tensor model is closely related to non-parametric graphon modelling (Amini & Levina, 2018; Klopp et al., 2017). In the high-order case, our framework lends itself well to tri-clustering (Hore et al., 2016) and multi-layer pattern recognition (Lee & Wang, 2020; Lei et al., 2020). The tensor block model is a building block for more complex structures including latent space models (Wang & Li, 2020), low-rank models (Young et al., 2018) and isotonic models (Pananjady & Samworth, 2022). In principle, complex tensor data can still be fitted by block models, at a cost of a large r. In this regard, we review our block models as the first step towards theoretical understanding of more complicated models. Building flexible models for general tensor data is an important future question.

While this paper mainly considers the tensor block model under Gaussian noise, our algorithm and proof techniques can also be generalised to other statistical settings, such as missing observations, non-continuous observations (e.g. dichotomous or count data), and heavy-tailed noises. For example, when the tensor entries are observed under Huber's ε-contamination model, one can change the aggregated mean procedure in HLloyd algorithm to aggregated median for robustness. On the other hand, it is also interesting to combine the discrete block assumption with other popular low-dimensional structures such as sparsity, monotonicity, and smoothness. In addition, our current lower bounds focus on settings with i.i.d. Gaussian noise; the statistical and computational lower bounds for stochastic tensor block models with Bernoulli distribution remain a challenging and open problem.

9 | PROOF SKETCHES AND TECHNICAL OVERVIEW OF MAIN RESULTS

In this section, we briefly discuss the high-level ideas and sketches of proofs for Theorem 3, Proposition 1, and Theorem 2. The complete proofs are deferred to Section B in Appendix S1.

9.1 | Proof sketch of Theorem 3

Our analysis is conducted on the misclassification loss $l_k^{(t)}$ defined in (16). We specifically discuss how to bound $l_1^{(0)}$ while the same argument applies to the other $l_k^{(0)}$ s. Recall $z_1^{(0)}$ is the clustering outcome of applying k-means on rows of $\hat{\mathbf{Y}}_1$ ($\hat{\mathbf{Y}}_1$ is defined in (10)). One can check applying k-means on the rows of $\hat{\mathbf{Y}}_1$ is equivalent to applying k-means on the rows of

$$\tilde{\mathbf{Y}}_1 := \mathbb{P}_{\hat{\mathbf{U}}_1}\mathcal{M}_1\left(\mathcal{Y}\times_2\mathbb{P}_{\hat{\mathbf{U}}_2}\times\cdots\times_d\mathbb{P}_{\hat{\mathbf{U}}_d}\right)\in\mathbb{R}^{p_1\times p_{-1}},$$

where $\mathbb{P}_{\hat{\mathbf{U}}} := \hat{\mathbf{U}}\hat{\mathbf{U}}^\top$ is the projection operator. Under proper regularity conditions, we can prove that

$$l_1^{(0)} \lesssim \frac{Mr_{-1}}{p_*}\left\|\tilde{\mathbf{Y}}_1 - \mathcal{M}_1(\mathcal{X})\right\|_{\mathrm{F}}^2. \tag{31}$$

Thus, to bound $l_1^{(0)}$, we only need to bound $\|\tilde{\mathbf{Y}}_1 - \mathcal{M}_1(\mathcal{X})\|_{\mathrm{F}}^2$, that is, the difference of $\tilde{\mathbf{Y}}_1$ to its population counterpart $\mathbb{E}\mathcal{M}_1(\mathcal{Y}) = \mathcal{M}_1(\mathcal{X})$, which is exactly the goal of Proposition 1. Then Theorem 3 can be concluded by combining (31), Proposition 1, and Lemma 1.

9.2 | Proof ideas of Proposition 1

We assume $r_k \le r_{-k}$ for the convenience of illustration. One can show by tensor algebra that

$$\|\tilde{\mathcal{Y}} - \mathcal{X}\|_{\mathrm{F}}^2 \lesssim \sum_{k=1}^d\left\|\hat{\mathbf{U}}_{k\perp}\mathcal{M}_k(\mathcal{X})\right\|_{\mathrm{F}}^2 + \left\|\mathcal{Z}\times_1\hat{\mathbf{U}}_1^\top\times\cdots\times_d\hat{\mathbf{U}}_d^\top\right\|_{\mathrm{F}}, \tag{32}$$

where $\hat{\mathbf{U}}_{k\perp}$ is the perpendicular subspace of $\hat{\mathbf{U}}_k$. The second term in (32) can be bounded by sub-Gaussian concentration inequalities while the analysis on the first term ($\|\hat{\mathbf{U}}_{k\perp}\mathcal{M}_k(\mathcal{X})\|_{\mathrm{F}}^2$) is more involved. Recall that

$$\begin{aligned}\hat{\mathbf{U}}_k &= \mathrm{SVD}_{r_k}\left(\mathcal{M}_k(\mathcal{Y})\left(\tilde{\mathbf{U}}_{k+1}\otimes\cdots\otimes\tilde{\mathbf{U}}_d\otimes\tilde{\mathbf{U}}_1\otimes\cdots\otimes\tilde{\mathbf{U}}_{k-1}\right)^\top\right),\\ \tilde{\mathbf{U}}_k &= \mathrm{SVD}_{r_k}(\mathcal{M}_k(\mathcal{Y})).\end{aligned}$$

A classic scheme to analyse $\|\hat{\mathbf{U}}_{k\perp}\mathcal{M}_k(\mathcal{X})\|_{\mathrm{F}}$ is by establishing an upper bound on the principle angles between the sub-spaces spanned by the preliminary estimated singular vectors ($\tilde{\mathbf{U}}_k$) and the true singular vectors ($\mathbf{U}_k := \mathrm{SVD}_{r_k}(\mathcal{M}_k(\mathcal{X}))$) in the sine-theta distance: $\|\sin\Theta(\tilde{\mathbf{U}}_k,\mathbf{U}_k)\| = \|\tilde{\mathbf{U}}_{k\perp}^\top\mathbf{U}_k\|$. To obtain an upper bound on $\|\sin\Theta(\tilde{\mathbf{U}}_k,\mathbf{U}_k)\|$, a singular value gap condition, that is, some lower bound on $\lambda_{r_k}(\mathcal{M}_k(\mathcal{X})) - \lambda_{r_k+1}(\mathcal{M}_k(\mathcal{X}))$, is crucial as indicated by the classic literature on matrix perturbation theory (Davis & Kahan, 1970; Wedin, 1972). Since there is no singular value gap condition in the context of Proposition 1, it is difficult to prove the desired bounds via bounding $\|\sin\Theta(\tilde{\mathbf{U}}_k,\mathbf{U}_k)\|$.

Let $\mathbf{U}_k' = \mathrm{SVD}_{r_k'}(\mathcal{M}_k(\mathcal{X}))$ for some $r_k' \le r_k$. Our main idea for proving Proposition 1 is to decompose $\mathcal{X}$ into two parts:

$$\mathcal{X} = \mathcal{X}' + (\mathcal{X} - \mathcal{X}'),\quad \text{where } \mathcal{X}' := \mathcal{X}\times_1\mathbb{P}_{\mathbf{U}_1'}\times\cdots\times_d\mathbb{P}_{\mathbf{U}_d'}.$$

By appropriately choosing r'_k, $\mathcal{X}'$ can be taken as the 'strong signal' tensor with large singular values; while $\mathcal{X} - \mathcal{X}'$ is the 'weak signal' tensor with small singular values. With such a decomposition, we only focus on estimating $\mathcal{X}'$ while leaving $\mathcal{X} - \mathcal{X}'$ as a bias term. To this end, we first introduce a deterministic matrix perturbation bound for subspaces of different dimensions.

Lemma 2. *Suppose the first r and the rest $p_1 - r$ singular vectors of $\mathbf{Y} \in \mathbb{R}^{p_1 \times p_2}$ are $\hat{\mathbf{U}} \in \mathbb{O}_{p_1,r}$ and $\hat{\mathbf{U}}_\perp \in \mathbb{O}_{p_1,p_1-r}$, respectively. For some $1 < r' \le r$, let $\mathbf{W} \in \mathbb{O}_{p_1,r'}$ be any p_1-by-r' column orthonormal matrix and $\mathbf{W}_\perp \in \mathbb{O}_{p_1,p_1-r'}$ be the orthogonal complement of $\mathbf{W}$. Given that $\sigma_{r'}(\mathbf{W}^\top \mathbf{Y}) > \sigma_{r+1}(\mathbf{Y})$, we have*

$$\left\| \hat{\mathbf{U}}_\perp^\top \mathbf{W} \right\| \le \frac{\sigma_{r'}(\mathbf{W}^\top \mathbf{Y}) \|\mathbf{W}_\perp^\top \mathbf{Y} \mathbb{P}_{\mathbf{Y}^\top \mathbf{W}}\|}{\sigma_{r'}^2(\mathbf{W}^\top \mathbf{Y}) - \sigma_{r+1}^2(\mathbf{Y})}.$$

Here $\mathbb{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^\dagger \mathbf{A}^\top$ is the projection operator.

To apply it to our problem, we further obtain the following probabilistic bound under the additive heterogeneous sub-Gaussian noises.

Lemma 3. *Let $\mathbf{Y} = \mathbf{X} + \mathbf{Z} \in \mathbb{R}^{p_1 \times p_2}$ where $\mathbf{X}$ is a rank- r matrix. Let $\hat{\mathbf{U}} = \mathrm{SVD}_r(\mathbf{Y})$, $\mathbf{U}_{r'} = \mathrm{SVD}_{r'}(\mathbf{X})$ for some $1 \le r' \le r$. Suppose $\mathbf{Z}_{ij}$ are independent mean-zero sub-Gaussian entries with $\mathbb{E}\mathbf{Z}_{ij}^2 = 1$ and $\|\mathbf{Z}_{ij}\|_{\psi_2} \le C$. Then, with probability at least $1 - \exp(-cp_1 \wedge p_2)$,*

$$\left\| \hat{\mathbf{U}}_\perp^\top \mathbf{U}_{r'} \right\| \le C \left(\frac{\sqrt{p_1}}{\sigma_{r'}(\mathbf{X})} + \frac{\sqrt{p_1 p_2}}{\sigma_{r'}^2(\mathbf{X})} \right).$$

Here, $\|\mathbf{Z}_{ij}\|_{\psi_2}$ denotes the sub-Gaussian norm of $\mathbf{Z}_{ij}$.

Applying Lemma 3 to $\tilde{\mathbf{U}}_k$ and $\mathbf{U}'_k$, we can obtain a neat bound for $\|\tilde{\mathbf{U}}_{k\perp}^\top \mathbf{U}'_k\|$, which can be further used to analyse the estimation error of $\mathcal{X}'$ via the bound of $\|\hat{\mathbf{U}}_{k\perp} \mathcal{M}_k(\mathcal{X}')\|_{\mathrm{F}}$. The complete proof is postponed to Section B.2.

Remark 8. We highlight the difference between the analyses of clustering in the matrix case (order $d = 2$) and the high-order tensor case (order $d \ge 3$) and illustrate the necessity of developing these techniques here. Note that in Gaussian mixture or bi-clustering models, the estimator $\tilde{\mathbf{Y}}_1$ is essentially the rank-r_1 truncated-SVD estimator of $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2}$ and it be can directly obtained that

$$\begin{aligned} \|\tilde{\mathbf{Y}}_1 - \mathcal{X}\|_{\mathrm{F}}^2 &\le 2r \|\tilde{\mathbf{Y}}_1 - \mathcal{X}\|^2 \le 4r \left(\|\tilde{\mathbf{Y}}_1 - \mathcal{Y}\|^2 + \|\mathcal{X} - \mathcal{Y}\|^2 \right) \\ &\overset{(*)}{\le} 8r_1 \|\mathcal{X} - \mathcal{Y}\|^2 = 8r_1 \|\mathcal{Z}\|^2 \overset{(**)}{\le} 8(p_1 + p_2) r_1, \qquad \text{w.h.p.} \end{aligned}$$

Here, (*) comes from the optimality of SVD while (**) comes from the spectral norm of Gaussian random matrix. Note that the above argument is free of any singular value gap as it does not involve any analysis on singular sub-space estimators. In contrast, there is no such an analytic property for the estimator $\tilde{\mathbf{Y}}_1$ when $d \ge 3$ and we have to analyse the accuracy of the singular subspace estimators $\{\tilde{\mathbf{U}}_k\}_{k=1}^d$ in a more complicated way. This makes the theoretical analysis fundamentally more difficult without the singular value gap condition in the tensor setting.

9.3 | Proof sketch of Theorem 2

To prove Theorem 2, we establish a local contraction property on $l_k^{(t)}$. Inspired by the convergence analysis of iterative algorithm for single discrete structure (Gao & Zhang, 2022), we introduce an oracle clustering procedure $\tilde{\mathcal{S}}, \tilde{z}_k$ as follows:

$$\begin{aligned}\tilde{\mathcal{S}}_{i_1,\dots,i_d} &= \text{Average}\left(\left\{\mathcal{Y}_{j_1,\dots,j_d} : (z_k)_{j_k} = i_k, \forall k \in [d]\right\}\right), \qquad i_k \in [r_k],\\ (\tilde{z}_k)_j &= \underset{a\in[r_k]}{\arg\min} \left\|(\mathcal{M}_k(\mathcal{Y}))_{j:}\,\mathbf{V}_k - \left(\mathcal{M}_k(\tilde{\mathcal{S}})\right)_{a:}\right\|_2^2, \qquad j \in [p_k],\end{aligned} \tag{33}$$

where $\mathbf{V}_k := \mathbf{W}_1 \otimes \cdots \mathbf{W}_{k-1} \otimes \mathbf{W}_k \otimes \cdots \otimes \mathbf{W}_d$ and $\mathbf{W}_k := \mathbf{M}_k\left(\text{diag}\left(1_{p_k}^\top \mathbf{M}_k\right)\right)^{-1}$ is the weighted membership matrix.

Consider the mode-1 clustering, for the oracle procedure (33), by definition $j \in [p_1]$, $(\tilde{z}_1)_j \neq (z_1)_j$ if and only if there exists $b \in [r_1] \setminus (z_1)_j$, such that

$$\left\|(\mathcal{M}_1(\mathcal{Y}))_{j:}\,\mathbf{V}_1 - \left(\mathcal{M}_1(\tilde{\mathcal{S}})\right)_{(z_1)_j:}\right\|_2^2 > \left\|(\mathcal{M}_1(\mathcal{Y}))_{j:}\,\mathbf{V}_1 - \left(\mathcal{M}_1(\tilde{\mathcal{S}})\right)_{b:}\right\|_2^2. \tag{34}$$

Note that (34) is equivalent to

$$\begin{aligned}&\left\langle (\mathcal{M}_1(\mathcal{E}))_{j:}\,\mathbf{V}_1, \left(\mathcal{M}_1(\tilde{\mathcal{S}})\right)_{(z_1)_j:} - \left(\mathcal{M}_1(\tilde{\mathcal{S}})\right)_{b:}\right\rangle\\ &< \frac{1}{2}\left(-\left\|(\mathcal{M}_1(\mathcal{S}))_{(z_1)_j:} - \left(\mathcal{M}_1(\tilde{\mathcal{S}})\right)_{b:}\right\|^2 + \left\|(\mathcal{M}_1(\mathcal{S}))_{(z_1)_j:} - \left(\mathcal{M}_1(\tilde{\mathcal{S}})\right)_{(z_1)_j:}\right\|^2\right)\\ &\approx -\frac{1}{2}\left\|(\mathcal{M}_1(\mathcal{S}))_{(z_1)_j:} - (\mathcal{M}_1(\mathcal{S}))_{b:}\right\|^2.\end{aligned}$$

Therefore, we define the quantity

$$A_{1j} := \frac{\left\langle (\mathcal{M}_1(\mathcal{E}))_{j:}\,\mathbf{V}_1, \left(\mathcal{M}_1(\tilde{\mathcal{S}})\right)_{(z_1)_j:} - \left(\mathcal{M}_1(\tilde{\mathcal{S}})\right)_{b:}\right\rangle}{\left\|(\mathcal{M}_1(\mathcal{S}))_{(z_1)_j:} - (\mathcal{M}_1(\mathcal{S}))_{b:}\right\|^2},$$

and the following oracle loss with some small 'tolerace' constant $\delta \in (0,1)$:

$$\xi_1 = \frac{1}{p_1}\sum_{j=1}^{p_1}\sum_{b\in[r_1]/(z_1)_j}\left\|(\mathcal{M}_1(\mathcal{S}))_{(z_1)_j:} - (\mathcal{M}_1(\mathcal{S}))_{b:}\right\|_2^2 \cdot \mathbb{I}\left\{A_{1j} \le -\frac{1-\delta}{2}\right\}. \tag{35}$$

We can similarly define ξ_k for each $k \in [d]$. Intuitively speaking, ξ_k can be taken as a surrogate for the misclassification loss of the oracle clustering estimator $\tilde{z}_k$: we expect ξ_k is closely related to $l_k^{(t)}$ in (16) and may serve as the iteration limiting point. Different from the regular clustering, HLloyd algorithm involves iterations of all d memberships $z_k^{(t)}$; and $l_k^{(t+1)}$, the misclassification loss at step $(t+1)$, depends on $l_{k'}^{(t)}$ for all $k' \in [d]$. This makes the error contraction analysis for HLloyd much more involved. With a dedicated loss decomposition and error estimation, we can prove the following contraction inequality:

$$l_k^{(t+1)} \le \frac{3}{2}\xi_k + \frac{1}{2}\max_{k\in[d]} l_k^{(t)}. \tag{36}$$

By sub-Gaussian concentration, one can further prove that with high probability

$$\xi_k \lesssim \sigma^2 \exp\left(-\frac{c_2 p_{-k}}{r_{-k}} \cdot \frac{\Delta_k^2}{\sigma^2}\right). \tag{37}$$

Combining (36), (37), and Lemma 1, we establish the loss convergence (19) in Theorem 2.

FUNDING INFORMATION

Rungang Han, Yuetian Luo, and Anru R. Zhang were partly supported by NSF Grants CAREER-2203741, DMS-2023239, NIH Grant R01 GM131399, and funding from Wisconsin Alumni Research Foundation. M. Wang was partly supported by NSF CAREER DMS-2141865, DMS-1915978, DMS-2023239, EF-2133740, and funding from Wisconsin Alumni Research Foundation.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://tianchi.aliyun.com/dataset/dataDetail?dataId=649 and https://openflights.org/data.html#route.

ORCID

Anru R. Zhang https://orcid.org/0000-0002-8721-5252

REFERENCES

Abbe, E. (2017) Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1), 6446–6531.

Aloise, D., Deshpande, A., Hansen, P. & Popat, P. (2009) NP-hardness of Euclidean sum-of-squares clustering. *Machine learning*, 75(2), 245–248.

Amini, A.A. & Levina, E. (2018) On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1), 149–179.

Anandkumar, A., Ge, R., Hsu, D. & Kakade, S.M. (2014) A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1), 2239–2312.

Arthur, D. & Vassilvitskii, S. (2006) *k-means++: the advantages of careful seeding*. Technical report, Stanford.

Barak, B., Hopkins, S., Kelner, J., Kothari, P.K., Moitra, A. & Potechin, A. (2019) A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2), 687–735.

Barak, B. & Moitra, A. (2016) Noisy tensor completion via the sum-of-squares hierarchy. Proceedings of the *conference on learning theory*, pp. 417–445.

Brennan, M. & Bresler, G. (2020) *Reducibility and statistical-computational gaps from secret leakage*. proceedings of 33rd conference on learning theory, pp. 648–847.

Brennan, M., Bresler, G. & Huleihel, W. (2018) Reducibility and computational lower bounds for problems with planted sparse structure. *Conference On Learning Theory*, pp. 48–166.

Busygin, S., Prokopyev, O. & Pardalos, P.M. (2008) Biclustering in data mining. *Computers & Operations Research*, 35(9), 2964–2987.

Carroll, J.D. & Chang, J.-J. (1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika*, 35(3), 283–319.

Chen, Y. & Xu, J. (2016) Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *The Journal of Machine Learning Research*, 17(1), 882–938.

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

Chien, I.E., Lin, C.-Y. & Wang, I.-H. (2019) On the minimax misclassification ratio of hypergraph community detection. *IEEE Transactions on Information Theory*, 65(12), 8095–8118.

Davis, C. & Kahan, W.M. (1970) The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1), 1–46.

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

De Lathauwer, L., De Moor, B. & Vandewalle, J. (2000b) On the best rank-1 and rank-(r 1, r 2 , … , rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4), 1324–1342.

Diakonikolas, I., Kane, D. M. & Stewart, A. (2017). Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 73–84. IEEE.

Eckart, C. & Young, G. (1936) The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211–218.

Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S.S. & Xiao, Y. (2017) Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2), 8.

Gao, C., Lu, Y., Ma, Z. & Zhou, H.H. (2016) Optimal estimation and completion of matrices with biclustering structures. *The Journal of Machine Learning Research*, 17(1), 5602–5630.

Gao, C. & Zhang, A.Y. (2022) Iterative algorithm for discrete structure recovery. *The Annals of Statistics*, 50(2), 1066–1094.

Ghoshdastidar, D. & Dukkipati, A. (2015). Spectral clustering using multilinear SVD: analysis, approximations and applications. *Proceedings of the 29th AAAI conference on artificial intelligence*.

Han, R., Willett, R. & Zhang, A.R. (2022) An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1), 1–29.

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

Hopkins, S. B. & Steurer, D. (2017) Efficient bayesian estimation from few samples: community detection and related problems. *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, p. 379–390. IEEE.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M.I., Small, K. et al. (2016) Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics*, 48(9), 1094–1100.

Jain, A.K. (2010) Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8), 651–666.

Ke, Z. T., Shi, F. & Xia, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*.

Klopp, O., Tsybakov, A.B., Verzelen, N. et al. (2017) Oracle inequalities for network models and sparse Graphon estimation. *The Annals of Statistics*, 45(1), 316–354.

Kolda, T. & Bader, B. (2006) The Tophits model for higher-order web link analysis. *Workshop on link analysis, counterterrorism and security*, volume 7, pp. 26–29.

Kolda, T.G. (2001) Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23(1), 243–255.

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

Koniusz, P. & Cherian, A. (2016) Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 5395–5403.

Lee, C. & Wang, M. (2020) Tensor denoising and completion based on ordinal observations. *International Conference of Machine Learning*, to appear.

Lei, J., Chen, K. & Lynch, B. (2020) Consistent community detection in multi-layer network data. *Biometrika*, 107(1), 61–73.

Löffler, M., Zhang, A.Y. & Zhou, H.H. (2021) Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics*, 49(5), 2506–2530.

Lu, Y. & Zhou, H. H. (2016) Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.

Luo, Y., Raskutti, G., Yuan, M. & Zhang, A.R. (2021) A sharp blockwise tensor perturbation bound for orthogonal iteration. *Journal of Machine Learning Research*, 22(179), 1–48.

Luo, Y. & Zhang, A. R. (2020) Open problem: average-case hardness of hypergraphic planted clique detection. *Proceedings of 33rd conference on learning theory, PMLR*, Volume 125, pp. 3852–3856.

Luo, Y. & Zhang, A.R. (2022) Tensor clustering with planted structures: statistical optimality and computational limits. *The Annals of Statistics*, 50(1), 584–613.

Ma, Z. & Wu, Y. (2015) Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3), 1089–1116.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M. et al. (2015) The human transcriptome across tissues and individuals. *Science*, 348(6235), 660–665.

Milligan, G.W. & Cooper, M.C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4), 441–458.

Mirsky, L. (1960) Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1), 50–59.

Nickel, M., Tresp, V. & Kriegel, H.-P. (2011) A three-way model for collective learning on multi-relational data. In: *Proceedings of the 28th International Conference on Machine Learning*, Vol. 11, pp. 809–816.

Pananjady, A. & Samworth, R.J. (2022) Isotonic regression with unknown permutations: statistics, computation and adaptation. *The Annals of Statistics*, 50(1), 324–350.

Richard, E. & Montanari, A. (2014) A statistical model for tensor PCA. *In Advances in Neural Information Processing Systems*, 2897–2905.

Shan, L., Lin, L., Sun, C. & Wang, X. (2016) Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization. *Electronic Commerce Research and Applications*, 16, 30–42.

Sun, W.W., Lu, J., Liu, H. & Cheng, G. (2015) Provable sparse tensor decomposition. *Journal of Royal Statistical Association, Series B*, 79, 899–916.

Von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics and Computing*, 17(4), 395–416.

Wang, M., Fischer, J., Song, Y.S. et al. (2019) Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *The Annals of Applied Statistics*, 13(2), 1103–1127.

Wang, M. & Li, L. (2020) Learning from binary multiway data: probabilistic tensor decomposition and its statistical optimality. *Journal of Machine Learning Research*, 21(154), 1–38.

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

Wang, T., Berthet, Q. & Samworth, R.J. (2016) Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5), 1896–1930.

Wedin, P.-A. (1972) Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1), 99–111.

Wu, T., Benson, A.R. & Gleich, D.F. (2016) General tensor spectral co-clustering for higher-order data. *Advances in Neural Information Processing Systems*, 2559–2567.

Wu, Y. & Xu, J. (2021) Statistical problems with planted structures: information-theoretical and computational limits. *Information-Theoretic Methods in Data Science*, 383, 383–424.

Wu, Y. & Yang, P. (2020) Optimal estimation of Gaussian mixtures via denoised method of moments. *Annals of Statistics*, 48(4), 1981–2007.

Wu, Y. & Zhou, H. H. (2019) Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $o(\sqrt{n})$ iterations. *arXiv preprint arXiv:1908.10935*.

Xia, D., Yuan, M. & Zhang, C.-H. (2021) Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *The Annals of Statistics*, 49(1), 76–99.

Young, J.-G., St-Onge, G., Desrosiers, P. & Dubé, L.J. (2018) Universality of the stochastic block model. *Physical Review E*, 98(3), 032309.

Zha, H., He, X., Ding, C., Gu, M. & Simon, H.D. (2002) Spectral relaxation for K-means clustering. *Advances in Neural Information Processing Systems*, 1057–1064.

Zhang, A. & Han, R. (2019) Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*, 114(528), 1708–1725.

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

Zhang, C., Han, R., Zhang, A.R. & Voyles, P.M. (2020) Denoising atomic resolution 4D scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy*, 219, 113123.

Zhang, Y., Bi, X., Tang, N. & Qu, A. (2021) Dynamic tensor recommender systems. *Journal of machine learning research*, 22, 1–35.

Zhang, Z., Allen, G.I., Zhu, H. & Dunson, D. (2019) Tensor network factorizations: relationships between brain structural connectomes and traits. *Neuroimage*, 197, 330–343.

Zhou, H., Li, L. & Zhu, H. (2013) Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502), 540–552.

Zhou, Z. & Amini, A.A. (2019) Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20, 47.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Han, R., Luo, Y., Wang, M. & Zhang, A.R. (2022) Exact clustering in tensor block model: Statistical optimality and computational limit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5), 1666–1698. Available from: https://doi.org/10.1111/rssb.12547