Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

Jiaxin Hu and Miaoyan Wang
University of Wisconsin - Madison

Abstract

We consider the problem of multiway clustering in the presence of unknown degree heterogeneity. Such data problems arise commonly in applications such as recommendation system, neuroimaging, community detection, and hypergraph partitions in social networks. The allowance of degree heterogeneity provides great flexibility in clustering models, but the extra complexity poses significant challenges in both statistics and computation. Here, we develop a degree-corrected tensor block model with estimation accuracy guarantees. We present the phase transition of clustering performance based on the notion of angle separability, and we characterize three signal-to-noise regimes corresponding to different statistical-computational behaviors. In particular, we demonstrate that an intrinsic statistical-to-computational gap emerges only for tensors of order three or greater. Further, we develop an efficient polynomial-time algorithm that provably achieves exact clustering under mild signal conditions. The efficacy of our procedure is demonstrated through two data applications, one on human brain connectome project, and another on Peru Legislation network ~~datasets~~dataset.

Index Terms

tensor clustering, degree correction, statistical-computational efficiency, human brain connectome networks

I. Introduction

MULTIWAY arrays have been widely collected in various fields including social networks (Anandkumar et al., 2014), neuroscience (Wang et al., 2017), and computer science (Koniusz and Cherian, 2016). Tensors effectively represent the multiway data and serve as the foundation in higher-order data analysis. One data example is from multi-tissue multi-individual gene expression study (Hore et al., 2016; Wang et al., 2019), where the data tensor consists of expression measurements indexed by (gene, individual, tissue) triplets. Another example is *hypergraph* network (Ahn et al., 2019; Ghoshdastidar and Dukkipati, 2017; Ghoshdastidar et al., 2017; Ke et al., 2019) in social science. A K-uniform hypergraph can be naturally represented as an order-K tensor, where each entry indicates the presence of K-way hyperedge among nodes (a.k.a. entities). In both examples, identifying the similarity among tensor entities is important for scientific discovery.

~~This paper studies~~ We study the problem of multiway clustering based on a data tensor. The goal of multiway clustering is to identify a checkerboard structure from a noisy data tensor. Figure 1 illustrates the noisy tensor and the underlying checkerboard structures discovered by multiway clustering methods. ~~The checkerboard structure serves as a meta tool to many popular structures including the low-rankness (Young et al., 2018), latent space models (Wang and Li, 2020), and isotonic models (Pananjady and Samworth, 2020).~~ In the hypergraph example, the multiway clustering aims to identify the underlying block partition of nodes based on their higher-order connectivities; therefore, we also refer to the clustering as *higher-order clustering*. The most common model for higher-order clustering is called *tensor block model* (TBM) (Wang and Zeng, 2019), which extends the usual matrix stochastic block model (Abbe, 2017) to tensors. The matrix analysis tools, however, are sub-optimal for higher-order clustering. Developing tensor tools for solving block models has received increased interest recently (Chi et al., 2020; Han et al., 2020; Wang and Zeng, 2019).

Classical tensor block model suffers from drawbacks to model real world data in spite of the popularity. The key underlying assumption of block model is that all nodes in the same community are exchangeable; i.e., the nodes have no individual effects apart from the block effects. However, the exchangeability assumption is often non-realistic.

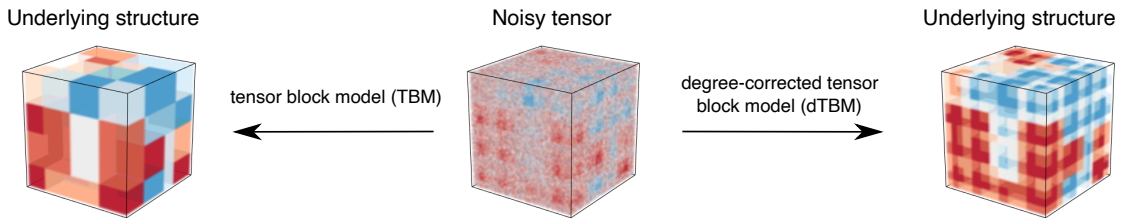

Fig. 1: Examples for order-3 tensor block model (TBM) with and without degree correction. Both TBM and dTBM have four communities on each mode, while dTBM allows a richer structure with degree heterogeneity.

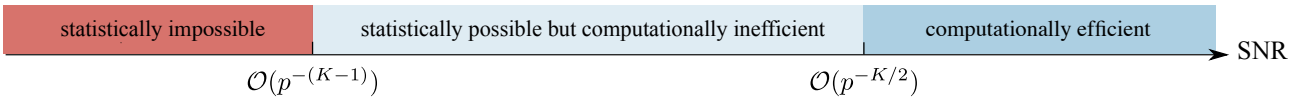

Fig. 2: SNR thresholds for statistical and computational limits in order-K dTBM with dimension $(p, ..., p)$ and $K \geq 2$. The SNR gap between statistical possibility and computational efficiency exists only for tensors with $K \geq 3$.

Each node may contribute to the data variation by its own multiplicative effect. Such degree heterogeneity appears commonly in social networks. ~~Ignoring the individual~~ Ignoring the degree heterogeneity may seriously mislead the clustering results. For example, regular block model fails to model the member affiliation in the Karate Club network (Bickel and Chen, 2009) without addressing degree heterogeneity.

The *degree-corrected tensor block model* (dTBM) has been proposed recently to account for the degree heterogeneity (Ke et al., 2019). The dTBM combines a higher-order checkerboard structure with degree parameter $\boldsymbol{\theta} = (\boldsymbol{\theta}(1), \ldots, \boldsymbol{\theta}(p))^T$ to allow heterogeneity among p nodes. Figure 1 compares the underlying structures of TBM and dTBM with the same number of communities. The dTBM allows varying values within the same community, thereby allowing a richer structure. To solve dTBM, we project clustering objects to a unit sphere and ~~consider the angle similarity; detailed algorithms~~ perform iterative clustering based on angle similarity. We refer to the algorithm as the *spherical* clustering; detailed procedures are in Section IV. ~~On one hand, the *spherical*~~ The spherical clustering avoids the estimation of nuisance degree heterogeneity. ~~On the other hand, the~~ The usage of angle similarity brings new challenges to the ~~derivations of~~ theoretical results, and ~~the polar coordinates based techniques are equipped in the proof.~~ we develop new polar-coordinate based techniques in the proofs.

Our contributions. The primary goal of this paper is to provide both statistical and computational guarantees for dTBM. Our main contributions are summarized below.

- We develop a general dTBM and establish the identifiability for the uniqueness of clustering using the notion of angle ~~seperability~~separability.
- We present the phase transition of clustering performance with respect to three different statistical and computational behaviors. We characterize, for the first time, the critical signal-to-noise (SNR) thresholds in dTBMs, revealing the intrinsic distinctions among (vector) one-dimensional clustering, (matrix) biclustering, and (tensor) higher-order clustering. Specific SNR thresholds and algorithm ~~behaviours~~ behaviors are depicted in Figure 2.
- We provide an angle-based algorithm that achieves exact clustering *in polynomial time* under mild conditions. Simulation and data studies demonstrate the outperformance of our algorithm compared with existing higher-order clustering algorithms.

The last two contributions, to our best knowledge, are new to the literature of dTBMs.

Related work. Our work is closely related to ~~,~~ but also distinct from several lines of existing research. Table I summarizes the most relevant models.

- *Block model for clustering.* Block models such as stochastic block model (SBM) and degree-corrected SBM ~~has~~ have been widely used for matrix clustering problems. The theoretical properties and algorithm performance

	Gao et al. (2018)	Han et al. (2020)	Ghoshdastidar et al. (2017)	Ke et al. (2019)	**Ours**
Allow tensors of arbitrary order	$\times$	$\surd$	$\surd$	$\surd$	$\surd$
Allow degree heterogeneity	$\surd$	$\times$	$\surd$	$\surd$	$\surd$
Singular-value gap-free clustering	$\surd$	$\surd$	$\times$	$\times$	$\surd$
Misclustering rate (for order K^*)	-	$\exp(-p^{K/2})$	p^{-1}	p^{-2}	$\exp(-p^{K/2})$

TABLE I: Comparison between previous methods with our method. *We list the result for order-K tensors with $K \geq 3$ and general number of communities $r = \mathcal{O}(1)$.

for matrix block models have been well-studied (Gao et al., 2018); see the review paper (Abbe, 2017) and the references therein. However, The tensor counterparts are relatively less understood. ~~Table I summarizes the most relevant models. Specifically,~~

- ~~the~~ *Tensor block model.* The tensor block model (TBM, ~~Han et al. (2020); Wang and Zeng (2019)) ,~~ Ghoshdastidar et al. (2017); ~~is~~ a higher-order extension of SBM, but it fails to allow degree heterogeneity. The Cartesian coordinates based analysis in Han et al. (2020) is ~~also~~ non-applicable to handle the extra ~~flexibility brought from~~ complexity brought by degree heterogeneity. In contrast, our model addresses the degree heterogeneity, and ~~the polar coordinates based tools are adapted~~ we develop polar-coordinate based tools for the theoretical analysis~~;~~ .
- ~~the~~ *Degree-corrected block model.* The hypergraph degree-corrected block model (hDCBM) proposed by Ke et al. (2019); Yuan et al. (ress) accounts for degree heterogeneity. ~~Whereas~~However, the hDCBM is designed only for binary observations, and the proposed spectral algorithm in Ke et al. (2019) achieves ~~a~~ sub-optimal ~~clustering accuracy~~ polynomial clustering rate in higher-order scenarios. In constrast, our model allows discrete and continuous entries, and achieves ~~exponentially~~ *exponentially* fast rate in clustering tasks. More importantly, to our best knowledge, we are the first to provide the statistical and computational limits analyses for the degree-corrected block model in tensor clustering. See Fig 2 for overview of our results.
- *Global-to-local algorithm strategy.* Our methods generalize the recent global-to-local strategy for matrix learning (Chi et al., 2019; Gao et al., 2018; Yun and Proutiere, 2016) to tensors (Ahn et al., 2018; Han et al., 2020; Kim et al., 2018). Despite the conceptual similarity, we address several fundamental challenges associated with this non-convex, non-continuous problem. We show the insufficiency of the conventional tensor HOSVD ~~(Kolda and Bader, 2009)~~(De Lathauwer et al., 2000), and we develop a weighted higher-order initialization that relaxes the ~~eigen-gap~~ singular-value gap separation condition. Furthermore, our local iteration leverages the angle-based clustering in order to avoid explicit estimation of degree heterogeneity. Our bounds reveal the interesting interplay between the computational and statistical errors. We show that our final estimate ~~provably~~ *provably* achieves the exact clustering within only polynomial-time complexity.

Notation. We use lower-case letters (e.g., a, b) for scalars, lower-case boldface letters (e.g., $\boldsymbol{a}, \boldsymbol{\theta}$) for vectors, upper-case boldface letters (e.g., $\boldsymbol{X}, \boldsymbol{Y}$) for matrices, and calligraphy letters (e.g., $\mathcal{X}, \mathcal{Y}$) for tensors of order three or greater. We use $\mathbf{1}_p$ to denote a vector of length p with all entries to be 1. We use $|\cdot|$ for the cardinality of a set and $\mathbb{1}\{\cdot\}$ for the indicator function. For an integer $p \in \mathbb{N}_+$, we use the shorthand $[p] = \{1, 2, ..., p\}$. For a length-p vector $\boldsymbol{a}$, we use $a(i) \in \mathbb{R}$ to denote the i-th entry of $\boldsymbol{a}$, and use $\boldsymbol{a}_I$ to denote the sub-vector by restricting the indices in the set $I \subset [p]$. We use $\|\boldsymbol{a}\| = \sqrt{\sum_i a^2(i)}$ to denote the ℓ_2-norm, $\|\boldsymbol{a}\|_1 = \sum_i |a_i|$ to denote the ℓ_1 norm of $\boldsymbol{a}$. For two vector $\boldsymbol{a}, \boldsymbol{b}$ of the same dimension, we denote the angle between $\boldsymbol{a}, \boldsymbol{b}$ by

$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\|\boldsymbol{a}\| \, \|\boldsymbol{b}\|},$$

where $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ is the inner product of two vectors and $\cos(\boldsymbol{a}, \boldsymbol{b}) \in [-1, 1]$. We make the convention that $\cos(\boldsymbol{a}, \boldsymbol{b}) = \cos(\boldsymbol{a}^T, \boldsymbol{b}^T)$.

~~For a matrix $\boldsymbol{Y}$, we use $\boldsymbol{Y}_{i:}$ to denote the i-th row of the matrix.~~ Let $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ be an order-K $(p_1, ..., p_K)$-dimensional tensor. We use $\mathcal{Y}(i_1, \ldots, i_K)$ to denote the $(i_1, \ldots, i_K)$-th entry of $\mathcal{Y}$. The multilinear multiplication of a tensor $\mathcal{S} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ by matrices $\boldsymbol{M}_k \in \mathbb{R}^{p_k \times r_k}$ results in an order-d $(p_1, \ldots, p_K)$-dimensional tensor $\mathcal{X}$, denoted

$$\mathcal{X} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times \cdots \times_K \boldsymbol{M}_K,$$

where the entries of $\mathcal{X}$ are defined by

$$\mathcal{X}(i_1,\ldots,i_K) = \sum_{(j_1,\ldots,j_K)} \mathcal{S}(j_1,\ldots,j_K)\boldsymbol{M}_1(i_1,j_1)\cdots\boldsymbol{M}_K(i_K,j_K).$$

For a matrix $\boldsymbol{Y}$, we use $\boldsymbol{Y}_{i,:}$ (respectively, $\boldsymbol{Y}_{:,i}$) to denote the i-th row (respectively, i-th column) of the matrix. Similarly, for an order-3 tensor, we use $\mathcal{Y}_{:,:,i}$ to denote the i-th matrix slide of the tensor. We use $\mathrm{Ave}(\cdot)$ to denote the operation of taking averages across elements and $\mathrm{Mat}_k(\cdot)$ to denote the unfolding operation that reshapes the tensor along mode k into a matrix. For a symmetric tensor $\mathcal{Y} \in \mathbb{R}^{p\times\cdots\times p}$, we omit the subscript and use $\mathrm{Mat}(\mathcal{Y}) \in \mathbb{R}^{p\times p^{K-1}}$ to denote the unfolding. For two sequences $\{a_p\},\{b_p\}$, we denote $a_p \lesssim b_p$ or $a_p = \mathcal{O}(b_p)$ if $\lim_{p\to\infty} a_p/b_p \le c$ for some constant $c \ge 0$, $a_p = o(b_p)$ if ~~$\lim_{p\to\infty} a_p/b_p \le c$~~ $\lim_{p\to\infty} a_p/b_p = 0$, and $a_p = \Omega(b_p)$ if ~~$cb_p \le a_p \le Cb_p$, for some constants $c, C \ge 0$~~both $b_p \lesssim a_p$ and $a_p \lesssim b_p$. Throughout the paper, we use the terms "community" and "clusters" exchangeably.

Organization. The rest of this paper is organized as follows. Section II introduces the degree-corrected tensor block model (dTBM) with three motivating examples and presents the identifiability of dTBM under the angle gap condition. We show the phase transition and the existence of statistical-computational gaps for the higher-order dTBM in Section III. In Section IV, we provide a polynomial-time two-stage algorithm with misclustering rate guarantees. Numerical studies including the ~~simulations to assess the theoretical results~~simulation, comparison with other methods, and ~~real data analysis on human brain connectome data and Peru legislation data are in Section V. Last, we conclude our paper~~ two real dataset analyses are in Sections V-VI. The main technical ideas we develop for addressing main theorems are provided in Section ~~??.~~ VII. Detailed proofs and extra theoretical results are provided in Appendix.

II. Model formulation and motivations

A. Degree-corrected tensor block model

Suppose we have an order-K data tensor $\mathcal{Y} \in \mathbb{R}^{p\times\cdots\times p}$. For ease of notation, we focus on symmetric tensors in this section; ~~our framework easily extends~~ the extension to general asymmetric tensors is provided in Section IV-C. Assume there exist $r \ge 2$ disjoint communities among the p nodes. We represent the community assignment by a function $z\colon [p] \mapsto [r]$, where $z(i) = a$ for i-th node that belongs to the a-th community. Then, $z^{-1}(a) = \{i \in [p]\colon z(i) = a\}$ denotes the set of nodes that belong to the a-th community, and $|z^{-1}(a)|$ denotes the number of nodes in the a-th community. Let $\boldsymbol{\theta} = (\theta(1),\ldots,\theta(p))^T$ denote the degree heterogeneity for p nodes. We consider the order-K dTBM (Ghoshdastidar et al., 2017; Ke et al., 2019),

$$\mathcal{Y}(i_1,\ldots,i_K) = \mathcal{S}(z(i_1),\ldots,z(i_K))\prod_{k=1}^{K}\theta_{i_k} + \mathcal{E}(i_1,\ldots,i_K), \tag{1}$$

where $\mathcal{S} \in \mathbb{R}^{r\times\cdots\times r}$ is an order-K tensor collecting the block means among communities, and $\mathcal{E} \in \mathbb{R}^{p\times\cdots\times p}$ is a noise tensor consisting of independent ~~mean-zero~~ zero-mean sub-Gaussian entries with variance bounded by σ^2. The unknown parameters are z, S, and $\boldsymbol{\theta}$. The dTBM can be equivalently written in a compact form of tensor-matrix product:

$$\mathbb{E}\mathbb{E}\mathcal{Y} = \mathcal{S}\times_1 \boldsymbol{\Theta}\boldsymbol{M}\times_2\cdots\times_K\boldsymbol{\Theta}\boldsymbol{M}, \tag{2}$$

where $\boldsymbol{\Theta} = \mathrm{diag}(\theta(1),...,\theta(p)) \in \mathbb{R}^{p\times p}$ is a diagonal matrix, $\boldsymbol{M} \in \{0,1\}^{p\times r}$ is the membership matrix associated with community assignment z such that $\boldsymbol{M}(i,j) = \mathbb{1}\{z(i) = j\}$. By definition, each row of $\boldsymbol{M}$ has one copy of 1's and 0's elsewhere. Note that the discrete nature of $\boldsymbol{M}$ renders our model (2) more challenging than Tucker decomposition. We call a tensor $\mathcal{Y}$ an r-block tensor with degree $\boldsymbol{\theta}$ if $\mathcal{Y}$ admits dTBM (2). ~~Here, we give two special cases of dTBM.~~ The goal of clustering is to estimate z from a single noisy tensor $\mathcal{Y}$. We are particularly interested in the high-dimensional regime where p grows whereas $r = \mathcal{O}(1)$.

B. Motivating ~~Examples.~~ examples

Here, we provide ~~three~~ four applications to illustrate the practical necessity of dTBM.

a) Tensor block model: Consider the model (2). Let $\theta(i) = 1$ for all $i \in [p]$. The model (2) reduces to the tensor block model, which is widely used in previous clustering algorithms (Chi et al., 2020; Han et al., 2020; Wang and Zeng, 2019). The theoretical results in TBM serve as benchmarks for dTBM.

b) Community detection in hypergraphs: Hypergraph network is a powerful tool to ~~collect~~ represent the complex entity relations with higher-order interactions (Ke et al., 2019). A typical undirected hypergraph is denoted as $H = (V, E)$, where $V = [p]$ is the set of nodes and E is the set of undirected hyperedges. Each hyperedge in E is a subset of V, and we call the hyperedge an order-K edge if the corresponding subset involves K nodes. We call H a K-uniform hypergraph if E only contains order-K edges.

~~Similar with the adjacency matrix, it~~ It is natural to represent the K-uniform hypergraph ~~by~~ using a binary order-K adjacency tensor. Let $\mathcal{Y} \in \{0,1\}^{p\times\cdots\times p}$ denote the adjacency tensor, where the entries encode the presence or absence of ~~hyperdeges~~ order-K edges among p nodes. Specifically,

$$\mathcal{Y}(i_1, ..., i_K) = \begin{cases} 1 & \text{if } (i_1, ..., i_K) \in E \\ 0 & \text{if } (i_1, ..., i_K) \notin E \end{cases},$$

for all $(i_1, \ldots, i_K) \in [p]^K$~~.~~, we have

$$\mathcal{Y}(i_1, ..., i_K) = \begin{cases} 1 & \text{if } (i_1, ..., i_K) \in E, \\ 0 & \text{if } (i_1, ..., i_K) \notin E. \end{cases}$$

Assume there exist r disjoint communities among p nodes~~. The~~, and the connection probabilities depend on the community assignments and node effects. Then, the equation (2) models ~~$\mathbb{E}\mathcal{Y}$~~ $\mathbb{E}\mathcal{Y}$ with unknown degree heterogeneity $\boldsymbol{\theta}$ and ~~subgaussianity~~ sub-Gaussianity parameter $\sigma^2 = 1/4$.

c) Multi-layer weighted network: Multi-layer weighted network data consists of multiple networks ~~with~~ over the same set of nodes. One representative example is the brain ~~structural~~ connectome data (Zhang et al., 2019). The multi-layer weighted network $\mathcal{Y}$ has dimension of $p \times p \times L$, where p denotes the number of brain regions ~~in~~ of interest, and L denotes the number of ~~layer~~layers (networks). Each of the L networks describes one aspect of the brain ~~connections, and the~~ connectivity, such as functional connectivity or structural connectivity. The resulting tensor $\mathcal{Y}$ ~~include a mixture slices with continuous, binary, and count entries~~consists of a mixture of slices with various data types.

Assume there exist r disjoint communities among p nodes and r_l disjoint communities among the L layers. The multi-layer network community detection is modeled by the generalized asymmetric dTBM model (2)

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{\Theta}\boldsymbol{M} \times_2 \boldsymbol{\Theta}\boldsymbol{M} \times_3 \boldsymbol{\Theta}_l\boldsymbol{M}_l,$$

where $(\boldsymbol{\theta} \in \mathbb{R}^p, \boldsymbol{M} \in \{0,1\}^{p\times r})$ and $(\boldsymbol{\theta}_l \in \mathbb{R}^L, \boldsymbol{M}_l \in \{0,1\}^{L\times r_l})$ are the degree heterogeneity and membership matrices corresponding to the community structure for p nodes and L layers, respectively.

d) Gaussian higher-order clustering: Datasets in various fields such as medical image, genetics, and computer science are formulated as Gaussian tensors. One typical example is the multi-tissue gene expression dataset, which records the different gene expression in different individuals and different tissues. The dataset, denoted as $\mathcal{Y} \in \mathbb{R}^{p\times n\times t}$, consists of the expression data for p genes of n individuals in t tissues.

Assume there exist r_1, r_2, r_3 disjoint clusters for p genes, n individuals, and t tissues, respectively. We apply the generalized asymmetric dTBM model (2)

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \mathbf{\Theta}_1 \boldsymbol{M}_1 \times_2 \mathbf{\Theta}_2 \boldsymbol{M}_2 \times_3 \mathbf{\Theta}_3 \boldsymbol{M}_3,$$

where $\{(\boldsymbol{\theta}_k, \boldsymbol{M}_k)\}_{k=1}^3$ ~~refers to~~ represents the heterogeneity and membership for genes, individuals, and tissues.

Remark 1 (Comparison with non-degree models)**.** Our dTBM uses fewer block parameters than TBM. ~~Let the subscripts "deg" and "non" denote quantities in the models with and without degrees, respectively. Then, every r_{non}~~In particular, every non-degree r_1-block tensor can be represented by a ~~degree-corrected r_{deg}~~*degree-corrected* r_2-block tensor with ~~$r_{\text{deg}} \le r_{\text{non}}$~~$r_2 \le r_1$. In particular, there exist tensors with ~~$r_{\text{non}} = p$ but $r_{\text{deg}} = 1$~~$r_1 = p$ but $r_2 = 1$, so the reduction in ~~r~~ model complexity can be dramatic from p to 1. This fact highlights the benefits of introducing degree heterogeneity in higher-order clustering tasks.

C. Identifiability under angle gap condition

The goal of clustering is to estimate the partition function z from model (2). For ease of notation, we focus on symmetric tensors; the extension to non-symmetric tensors are similar. We use $\mathcal{P}$ to denote the following parameter space for $(z, \mathcal{S}, \boldsymbol{\theta})$,

$$\mathcal{P} = \left\{ (z, \mathcal{S}, \boldsymbol{\theta}) \colon \ \boldsymbol{\theta} \in \mathbb{R}_+^p, \ \frac{c_1 p}{r} \le |z^{-1}(a)| \le \frac{c_2 p}{r}, c_3 \le \|\text{Mat}(\mathcal{S})_{a:}\| \le c_4, \left\|\boldsymbol{\theta}_{z^{-1}(a)}\right\|_1 = |z^{-1}(a)|, a \in [r] \right\}, \quad (3)$$

where $c_i > 0$'s are universal constants. We briefly describe the rationale of the constraints in (3). First, the entrywise positivity constraint on $\boldsymbol{\theta} \in \mathbb{R}_+^p$ is imposed to avoid sign ambiguity between entries in $\boldsymbol{\theta}_{z^{-1}(a)}$ and $\mathcal{S}$~~and thereof allow~~. This constraint allows the trigonometric cos to describe the angle similarity in the ~~following~~ Assumption 1 below and Sub-algorithm 2 in Section IV~~. This~~. Note that the positivity constraint can be achieved without sacrificing model flexibility, by using a slightly larger dimension of $\mathcal{S}$ in the factorization (2); see Example 1 below. Second, ~~the~~ recall that the quantity $|z^{-1}(a)|$ denotes the number of nodes in a-th community. The constants c_1, c_2 in the $|z^{-1}(a)|$ bound assume the roughly balanced size across r communities. Third, the constants c_3, c_4 in the magnitude of $\text{Mat}(\mathcal{S})_{a:}$ requires no purely zero slide in $\mathcal{S}$, so the core tensor $\mathcal{S}$ is not trivially reduced to a lower rank. Lastly, the ℓ_1 normalization $\|\boldsymbol{\theta}_{z^{-1}(a)}\|_1 = |z^{-1}(a)|$ is imposed to avoid the scalar ambiguity between $\boldsymbol{\theta}_{z^{-1}(a)}$ and $\mathcal{S}$. This constraint, again, incurs no restriction to model flexibility but makes our presentation cleaner.

Example 1 (Positivity of degree parameters)**.** Here we provide an example to show the positivity constraints on $\boldsymbol{\theta}$ incurs no loss on the model flexibility. ~~We consider a~~ Consider an order-3 dTBM with core tensor $\mathcal{S} = 1$ and degree $\boldsymbol{\theta} = (1, 1, -1, -1)^T$. We have the mean tensor

$$\mathcal{X} = \mathcal{S} \times_1 \mathbf{\Theta}\boldsymbol{M} \times_2 \mathbf{\Theta}\boldsymbol{M} \times_3 \mathbf{\Theta}\boldsymbol{M},$$

where $\mathbf{\Theta} = \text{diag}(\boldsymbol{\theta})$ and $\boldsymbol{M} = (1, 1, 1, 1)^T$. Note that $\mathcal{X} \in \mathbb{R}^{4\times4\times4}$ is a 1-block tensor with ~~mixed-signed~~ *mixed-signed* degree $\boldsymbol{\theta}$, and the mode-3 slices of $\mathcal{X}$ are

$$\mathcal{X}_{::1} = \mathcal{X}_{::2} = -\mathcal{X}_{::3} = -\mathcal{X}_{::4} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}.$$

Now, instead of original decomposition, we encode ~~$\mathcal{S}$~~ $\mathcal{X}$ as a 2-block tensor with ~~positive-signed degree~~*positive-signed* degree. Specifically, we write

$$\mathcal{X} = \mathcal{S}' \times_1 \mathbf{\Theta}'\boldsymbol{M}' \times_2 \mathbf{\Theta}'\boldsymbol{M}' \times_3 \mathbf{\Theta}'\boldsymbol{M}',$$

where $\mathbf{\Theta}' = \text{diag}(\boldsymbol{\theta}') = \text{diag}(1, 1, 1, 1)$, the core tensor $\mathcal{S}' \in \mathbb{R}^{2\times2\times2}$ ~~have~~ has mode-3 slices

$$\mathcal{S}'_{::1} = -\mathcal{S}'_{::2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \boldsymbol{M}' = \begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_2 \end{bmatrix},$$

~~and $\Theta' = \mathrm{diag}(\boldsymbol{\theta}') = \mathrm{diag}(1,1,1,1)$.~~ , and the membership matrix $\boldsymbol{M}' \in \{0,1\}^{2\times 4}$ defines the clustering $z' : [4] \to [2]$,

$$\mathcal{S}'_{::1} = -\mathcal{S}'_{::2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \boldsymbol{M}' = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

The triplet $(z', \mathcal{S}', \boldsymbol{\theta}')$ lies in our parameter space (3). In general, we can always reparameterize a block-r tensor with mixed-signed degree using a block-$2r$ tensor with positive-signed degree. Since we ~~assumes~~ assume $r = \mathcal{O}(1)$ throughout the paper, the splitting does not affect the error rates of our interest.

We ~~first provide the identifiablity~~ now provide the identifiability conditions for our model before estimation procedures. When $r = 1$, the decomposition (2) is always unique (up to cluster label permutation) in $\mathcal{P}$, because dTBM is equivalent to the rank-1 tensor family under this case. When $r \geq 2$, the Tucker rank of signal tensor ~~$\mathbb{E}\mathcal{Y}$~~ $\mathbb{E}\mathcal{Y}$ in (2) is bounded by, but not necessarily equal to, the number of blocks r (Wang and Zeng, 2019). Therefore, one can not apply the classical identifiability conditions for low-rank tensors to dTBM. Here, we introduce a key separation condition on the core tensor.

Assumption 1 (Angle gap)**.** Let $\boldsymbol{S} = \mathrm{Mat}(\mathcal{S})$. Assume the minimal gap between normalized rows of $\boldsymbol{S}$ is bounded away from zero~~,~~ ; i.e.,

$$\Delta_{\min} := \min_{a \neq b \in [r]} \left\| \frac{\boldsymbol{S}_{a:}}{\|\boldsymbol{S}_{a:}\|} - \frac{\boldsymbol{S}_{b:}}{\|\boldsymbol{S}_{b:}\|} \right\| > 0, \quad \text{for} \quad r \geq 2, \tag{4}$$

~~and set~~

We make the convention $\Delta_{\min} = 1$ for $r = 1$. Equivalently, (4) says that none of the two rows in $\boldsymbol{S}$ are parallel~~,~~ ; i.e., ~~when $r \geq 2$, $\max_{a\neq b\in[r]} \cos(\boldsymbol{S}_{a:}, \boldsymbol{S}_{b:}) = 1 - \Delta^2_{\min}/2 < 1$.~~ $\max_{a\neq b\in[r]} \cos(\boldsymbol{S}_{a:}, \boldsymbol{S}_{b:}) = 1 - \Delta^2_{\min}/2 < 1$. The quantity $\Delta_{\min}$ characterizes the non-redundancy among clusters measured by angle separation. The denominators involved in definition (4) ~~is~~ are well posed because of the lower bound on $\|\boldsymbol{S}_{a:}\|$ in (3). ~~The following theorem shows that the~~

Our first main result is the following theorem showing the sufficiency and necessarily of the angle gap separation ~~is sufficient and necessary for~~ condition for the parameter identifiability under dTBM.

Theorem 1 (Model identifiability)**.** Consider the dTBM with $r \geq 2$. The parameterization (2) is unique in $\mathcal{P}$ up to cluster label permutations, if and only if Assumption 1 holds.

The identifiability guarantee for the dTBM is more appealing than classical Tucker model. In the Tucker model, the factor matrix $\boldsymbol{M}$ is identifiable only up to orthogonal rotations. In contrast, our model does not suffer from rotational invariance. As we will show in Section IV, each column of the membership matrix $\boldsymbol{M}$ can be precisely recovered under our algorithm. This property benefits the interpretation of dTBM in practice.

III. Statistical-computational ~~gaps~~ limits for higher-order tensors

In this section, we study the statistical and computational limits of dTBM. We ~~reserve the term higher-order tensors for tensors of order $K \geq 3$. We~~ propose signal-to-noise ratio (SNR),

$$\mathrm{SNR} := \Delta^2_{\min}/\sigma^2 = p^\gamma, \tag{5}$$

with varying $\gamma \in \mathbb{R}$ that quantifies different regimes of interest. We call γ the *signal exponent*. Intuitively, a larger SNR, or equivalently a larger γ, benefits the clustering in the presence of noise. With quantification (5), we consider the following parameter space,

$$\mathcal{P}(\gamma) = \mathcal{P} \cap \{\mathcal{S} \text{ satisfies SNR condition (5) with } \gamma\}.$$

Note that 1-block dTBM does not belong to the space $\mathcal{P}(\gamma)$ when $\gamma < 0$ by Assumption 1. Our goal is to characterize the clustering accuracy with respect to γ~~when $r \geq 2$~~. Let $\hat{z}$ and z be the estimated and true clustering functions in

the family (3). Define the misclustering error by

$$\ell(\hat{z}, z) = \frac{1}{p} \min_{\pi \in \Pi} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq \pi \circ z(i)\},$$

where $\pi : [r] \mapsto [r]$ is a permutation of cluster labels, $\circ$ denotes the composition operation, and Π denotes the collection of all possible permutations. The infinitum over all permutations accounts for the ambiguity in cluster label permutation.

In Sections III-A and III-B, we provide the lower bounds of $\ell(\hat{z}, z)$ for general Gaussian dTBMs (1) without symmetric assumptions. For general (asymmetric) Gaussian dTBMs, we assume Gaussian noise $\mathcal{E}(i_1,\ldots,i_K) \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$, and we extend the parameter space (3) to allow K clustering functions $(z_k)_{k\in[K]}$, one for each mode. For notational simplicity, we still use z and $\mathcal{P}(\gamma)$ for this general (asymmetric) model. All lower bounds should be interpreted as the worst-case results across K modes.

A. Statistical critical values

~~Our first main result is to show~~ The statistical limit means the minimal SNR required for solving dTBMs with *unlimited computational cost.* Our following result shows the minimax lower bound of SNR for exact recovery in dTBM.

Theorem 2 (Statistical lower bound)**.** Consider general Gaussian dTBMs under the parameter space $\mathcal{P}(\gamma)$ with $K \geq 1$. Assume $r \lesssim p^{1/3}$. If the signal exponent satisfies $\gamma < -(K-1)$, then, every estimator $\hat{z}_{\text{stat}}$ obeys

$$\sup_{(z,\mathcal{S},\boldsymbol{\theta})\in\mathcal{P}(\gamma)} \mathbb{E}\left[p\ell(\hat{z}_{\text{stat}}, z)\right] \geq 1.$$

Theorem 2 demonstrates the impossibility of exact recovery of the assignment when $\gamma < -(K-1)$ in the high-dimensional regime $p \to \infty$ for fixed r. The proof is information-theoretical, and therefore the results apply to all statistical estimators, including but not limited to, maximum likelihood estimation (MLE) (Wang and Zeng, 2019) and trace maximization (Ghoshdastidar and Dukkipati, 2017). ~~Our derived~~ As we will show in Section IV, the SNR threshold $-(K-1)$ is also a minimax upper bound, because MLE achieves exact recovery when $\gamma > -(K-1)$. Hence, the boundary $\gamma_{\text{stat}} \coloneqq -(K-1)$ is the critical value for statistical performance of dTBM.

B. Computational critical values

In this section, we derive the computational limits of dTBMs. The computational limit means the minimal SNR required for exactly recovery with *polynomial-time* computational cost. An important ingredient to establish the computational limits is the *hypergraphic planted clique (HPC) conjecture* (Brennan and Bresler, 2020; Zhang and Xia, 2018). The HPC conjecture indicates the impossibility of fully recovering the planted cliques with polynomial-time algorithm when the clique size is less than the number of vertices in the hypergraph. The formal statement of HPC detection ~~and conjecture can be found~~ conjecture is provided in Definition 1 and Conjecture 1 DIFdelbegin ~~following~~follows.

Definition 1 (Hypergraphic planted clique (HPC) detection)**.** Consider an order-K hypergraph $H = (V, E)$ where $V = [p]$ collects vertices and E collects all the order-K ~~-way~~ edges. Let $\mathcal{H}_k(p, 1/2)$ denote the Erdős-Rényi K-hypergraph where the edge $(i_1, \ldots, i_K)$ belongs to E with probability $1/2$. Further, we let $\mathcal{H}_K(p, 1/2, \kappa)$ denote the hyhpergraph with planted cliques of size κ. Specifically, we generate a hypergraph from $\mathcal{H}_k(p, 1/2)$, pick κ vertices uniformly from $[p]$, denoted K, and then connect all the hyperedges with vertices in K. Note that the clique size κ can be a function of p, denoted κ_p.

The order-K HPC detection aims to identify whether there exists a planted clique hidden in ~~a~~ an Erdős-Rényi K-hypergraph. ~~We formulate HPC detection~~ The HPC detection is formulated as the following hypothesis testing problem

$$H_0 : \ H \sim \mathcal{H}_K(p, 1/2) \quad \text{versus} \quad H_1 : \ H \sim \mathcal{H}_K(p, 1/2, \kappa_p).$$

Conjecture 1 (HPC conjecture)**.** Consider the HPC detection problem in Definition 1. Suppose the sequence $\{\kappa_p\}$ such that $\limsup_{p\to\infty} \log \kappa_p / \log \sqrt{p} \le (1-\tau)$. Then, for ~~any sequence of polynomial-times~~ every sequence of polynomial-time test $\{\varphi_p\} : H \mapsto \{0,1\}$ we have

$$\liminf_{p\to\infty} \mathbb{P}_{H_0}\left(\varphi_p(H)=1\right) + \mathbb{P}_{H_1}\left(\varphi_p(H)=0\right) \ge \frac{1}{2}.$$

Under the HPC conjecture, we establish the SNR lower bound that is necessary for any *polynomial-time* estimator to achieve exact clustering.

Theorem 3 (Computational lower bound)**.** Consider general Gaussian dTBMs under the parameter space $\mathcal{P}(\gamma)$ with $K \ge 2$. Assume HPC conjecture holds. If the signal exponent $\gamma < -K/2$, then, every *polynomial-time estimator* $\hat{z}_{\text{comp}}$ obeys

$$\liminf_{p\to\infty} \sup_{(z,\mathcal{S},\boldsymbol{\theta})\in\mathcal{P}(\gamma)} \mathbb{E}\left[p\ell(\hat{z}_{\text{comp}}, z)\right] \ge 1.$$

Theorem 3 indicates the impossibility of exact recovery by polynomial-time algorithms when $\gamma < -K/2$. Therefore, $\gamma_{\text{comp}} := -K/2$ is the critical value for computational performance of dTBM. In Section IV, we will show the condition $\gamma > -K/2$ suffices for our proposed polynomial-time estimator. Thus, $\gamma_{\text{comp}} := -K/2$ is the critical value for computational performance of dTBM.

Remark 2 (Statistical-computational gaps)**.** Now, we have established the phase transition of exact clustering under order-K dTBM by combing Theorems 2 and 3. Figure 2 summarizes our results of critical SNRs when $K \ge 2$. In the weak SNR region $\gamma < -(K-1)$, no statistical estimator succeeds in degree-corrected higher-order clustering. In the strong SNR region $\gamma > -K/2$, our proposed algorithm precisely recovers the clustering in polynomial time. In the moderate SNR regime, $-(K-1) \le \gamma \le -K/2$, the degree-corrected clustering problem is statistically easy but computationally hard. Particularly, dTBM reduces to matrix degree-corrected model when $K = 2$, and the statistical and computational bounds show the same critical value. When $K = 1$, dTBM reduces to the degree-corrected sub-Gaussian mixture model (GMM) with model

$$\boldsymbol{Y} = \boldsymbol{\Theta}\boldsymbol{M}\boldsymbol{S} + \boldsymbol{E},$$

where $\boldsymbol{Y} \in \mathbb{R}^{p\times d}$ collects n data points in $\mathbb{R}^d$, $\boldsymbol{S} \in \mathbb{R}^{r\times d}$ collects the d-dimensional centroids for r clusters, and $\boldsymbol{\Theta} \in \mathbb{R}^{p\times p}, \boldsymbol{M} \in \{0,1\}^{p\times r}, \boldsymbol{E} \in \mathbb{R}^{p\times d}$ have the same meaning as in dTBM. Lu and Zhou (2016) implies ~~polynomial-times~~ that polynomial-time algorithms are able to achieve the statistical minimax lower bound in GMM. Therefore, we conclude that the statistical-to-computational gap emerges only for higher-order tensors with $K \ge 3$. The result reveals the intrinsic distinctions among (vector) one-dimensional clustering, (matrix) biclustering, and (tensor) higher-order clustering.

Remark 3 (Comparison with non-degree models)**.** We compare our results to non-degree tensor models. The allowance of degree heterogeneity $\boldsymbol{\theta}$ makes the model more flexible, but it incurs extra statistical and computational complexity. Fortunately, we find that the extra complexity does not render the estimation of z qualitatively harder; see the comparison of our phase transition with non-degree TBM (Han et al., 2020).

IV. Polynomial-time algorithm under mild SNR

~~We present a two-stage clustering algorithm~~ In this section, we present an efficient polynomial-time clustering algorithm under mild SNR. The procedure takes a global-to-local approach. See Figure 3 for illustration. The global step finds the basin of attraction with polynomial miclustering error, whereas the local iterations improve the initial clustering to exact recovery. Both steps are critical to obtain a satisfactory algorithm output. ~~In this section~~In what follows, we first use the symmetric tensor as a working example to describe the algorithm procedures to gain insight. Our theoretical analysis focuses on the noisy tensor with i.i.d. sub-Gaussian noise such as Gaussian and uniform observations. The extensions for asymmetric tensor and Bernoulli observation and other practical issues are in ~~subsection~~Section IV-C.

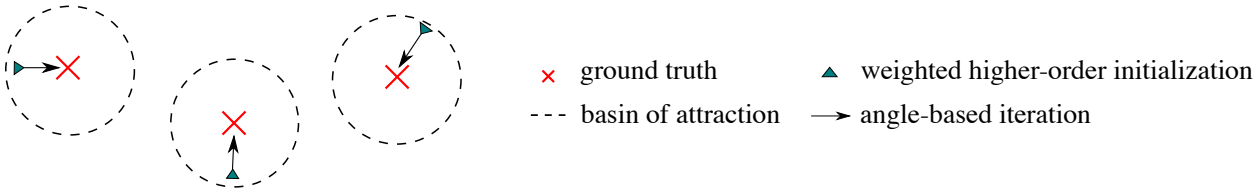

Fig. 3: Illustration of our global-to-local algorithm.

A. *Weighted higher-order initialization*

We start with weighted higher-order clustering algorithm as initialization. ~~We take the~~ We take an order-3 ~~tensor in illustration for simplicity.~~ symmetric tensor as illustration for insight. Consider noiseless case with $\mathcal{X} = \mathbb{E}\mathcal{Y}$ and $\boldsymbol{X} = \mathrm{Mat}(\mathcal{X})$. By model (2), for all $i \in [p]$, we have

$$\theta(i)^{-1}\boldsymbol{X}_{i:} = [\mathrm{Mat}(\mathcal{S} \times_2 \boldsymbol{\Theta}\boldsymbol{M} \times_3 \boldsymbol{\Theta}\boldsymbol{M})]_{z(i):}.$$

This implies that, all node i belonging to a-th community (i.e., $z(i) = a$) share the same normalized mean vector $\theta(i)^{-1}\boldsymbol{X}_{i:}$, and vice versa. Intuitively, one can apply k-means clustering to the vectors $\{\theta(i)^{-1}\boldsymbol{X}_{i:}\}_{i\in[p]}$, which leads to main idea of our Sub-algorithm 1.

Specifically, our initialization consists of denoising step and clustering step. The denoising step (lines 1-2 in Sub-algorithm 1) estimates $\mathcal{X}$ from $\mathcal{Y}$ by a double projection spectral method. The first projection performs HOSVD (De Lathauwer et al., 2000) via $\boldsymbol{U}_{\text{pre}} = \mathrm{SVD}_r(\mathrm{Mat}(\mathcal{Y}))$, where $\mathrm{SVD}_r(\cdot)$ returns the top-r left singular vectors. The second projection performs HOSVD on the projected $\mathcal{Y}$ onto the ~~multlilinear~~ multilinear Kronecker space $\boldsymbol{U}_{\text{pre}} \otimes \boldsymbol{U}_{\text{pre}}$; i.e.,

$$\hat{\boldsymbol{U}} = \mathrm{SVD}_r\left(\mathrm{Mat}\left(\mathcal{Y} \times_1 \boldsymbol{U}_{\text{pre}}\boldsymbol{U}_{\text{pre}}^T \times_2 \boldsymbol{U}_{\text{pre}}\boldsymbol{U}_{\text{pre}}^T\right)\right).$$

The final denoised tensor $\hat{\mathcal{X}}$ is defined by

$$\hat{\mathcal{X}} = \mathcal{Y} \times_1 \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T \times_2 \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T \times_3 \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T.$$

The double projection improves usual matrix spectral methods in order to alleviate the noise ~~tensor~~effects for $K \geq 3$ (Han et al., 2020).

The clustering step (lines 3-5 in Sub-algorithm 1) performs the weighted k-means clustering. We write $\hat{\boldsymbol{X}} = \mathrm{Mat}(\hat{\mathcal{X}})$, and normalize the rows into $\hat{\boldsymbol{X}}^s_{i:} = \|\hat{\boldsymbol{X}}_{i:}\|^{-1}\hat{\boldsymbol{X}}_{i:}$ as a surrogate of $\theta(i)^{-1}\boldsymbol{X}_{i:}$. Then, a weighted k-means clustering is performed on the normalized rows with weights equal to $\|\hat{\boldsymbol{X}}_{i:}\|^2$. The choice of weights is to bound the k-means objective function by the Frobenius-norm accuracy of $\hat{\mathcal{X}}$. Unlike existing clustering algorithm (Ke et al., 2019), we apply the clustering on the unfolded tensor $\hat{\boldsymbol{X}}$ rather than on the factors $\hat{\boldsymbol{U}}$. This strategy relaxes the ~~eigen-gap separation~~ singular-value gap condition (Gao et al., 2018; Han et al., 2020). We assign degenerate rows with purely zero entries to an arbitrarily random cluster; these nodes are negligible in high-dimensions because of the lower bound on $\|\mathrm{Mat}(\mathcal{S})_{a:}\|$ in (3). The final result gives the initial ~~clustering~~ cluster assignment $\hat{z}^{(0)}$. Full procedures are provided in Sub-algorithm 1.

~~We now establish the misclustering error rate of initialization. We call $\boldsymbol{\theta}$ is balanced, if the relative extent of heterogeneity is comparable across clusters in that~~

$$\min_{a\in[r]}\|\boldsymbol{\theta}_{z^{-1}(a)}\| = (1+o(1))\max_{a\in[r]}\|\boldsymbol{\theta}_{z^{-1}(a)}\|_1$$

~~Note that, the assumption does not preclude degree heterogeneity. Indeed, within each of the clusters, the highest degree can be $\theta(i) = \Omega(p)$, whereas the lowest degree can be $\theta(i) = \mathcal{O}(1)$.~~

We now establish the misclustering error rate of initialization. We call $\boldsymbol{\theta}$ is balanced, if the relative extent of

Algorithm: Multiway spherical clustering for degree-corrected tensor block model

Sub-algorithm 1: Weighted higher-order initialization

Input: Observation $\mathcal{Y} \in \mathbb{R}^{p\times\cdots\times p}$, number of ~~cluster~~ clusters r, relaxation factor $\eta > 1$ in k-means clustering.

1: Compute factor matrix $\boldsymbol{U}_{\text{pre}} = \text{SVD}_r(\text{Mat}(\mathcal{Y}))$ and the $(K-1)$-mode projection ~~$\mathcal{X}_{\text{pre}} = \mathcal{Y} \times_1 \boldsymbol{U}_{\text{pre}}^T \times_2 \cdots \times_{K-1} \boldsymbol{U}_{\text{pre}}^T$.~~ $\mathcal{X}_{\text{pre}} = \mathcal{Y} \times_1 \boldsymbol{U}_{\text{pre}}\boldsymbol{U}_{\text{pre}}^T \times_2 \cdots \times_{K-1} \boldsymbol{U}_{\text{pre}}\boldsymbol{U}_{\text{pre}}^T$.

2: Compute factor matrix $\hat{\boldsymbol{U}} = \text{SVD}_r(\text{Mat}(\mathcal{X}_{\text{pre}}))$ and denoised tensor $\hat{\mathcal{X}} = \mathcal{Y} \times_1 \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T \times_2 \cdots \times_K \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T$.

3: Let $\hat{\boldsymbol{X}} = \text{Mat}(\hat{\mathcal{X}})$ and $S_0 = \{i \in [p] : \|\hat{\boldsymbol{X}}_{i:}\| = 0\}$. Set $\hat{z}(i)$ randomly in $[r]$ for $i \in S_0$.

4: For all $i \in S_0^c$, compute normalized rows $\hat{\boldsymbol{X}}^s_{i:} := \|\hat{\boldsymbol{X}}_{i:}\|^{-1}\hat{\boldsymbol{X}}_{i:}$.

5: Solve the clustering $\hat{z}\colon [p] \to [r]$ and centroids $(\hat{\boldsymbol{x}}_j)_{j\in[r_k]}$ using weighted k-means, such that

$$\sum_{i\in S_0^c} \|\hat{\boldsymbol{X}}_{i:}\|^2 \|\hat{\boldsymbol{X}}^s_{i:} - \hat{\boldsymbol{x}}_{\hat{z}(i)}\|^2 \le \eta \min_{\bar{\boldsymbol{x}}_j, j\in[r], \bar{z}(i), i\in S_0^c} \sum_{i\in S^c} \|\hat{\boldsymbol{X}}_{i:}\|^2 \|\hat{\boldsymbol{X}}^s_{i:} - \bar{\boldsymbol{x}}_{\bar{z}(i)}\|^2.$$

Output: Initial clustering $z^{(0)} \leftarrow \hat{z}$.

Sub-algorithm 2: Angle-based iteration

Input: Observation $\mathcal{Y} \in \mathbb{R}^{p\times\cdots\times p}$, initialization $z^{(0)}\colon [p] \to [r]$ from Sub-algorithm 1, iteration number T.

6: **for** $t = 0$ to $T-1$ **do**

7: Update the block tensor $\mathcal{S}^{(t)}$ via $\mathcal{S}^{(t)}(i_1, ..., i_K) = \text{Ave}\{\mathcal{Y}(i_1, \ldots, i_K) : z^{(t)}(i_k) = j_k, k \in [K]\}$.

8: Calculate reduced tensor $\mathcal{Y}^{\text{d}} \in \mathbb{R}^{p\times r\times\cdots\times r}$ via

$$\mathcal{Y}^{\text{d}}(i, a_2, \ldots, a_K) = \text{Ave}\{\mathcal{Y}(i, i_2, \ldots, i_K) : z^{(t)}(i_k) = a_k, k \ne 1\}.$$

9: Let $\boldsymbol{Y}^{\text{d}} = \text{Mat}(\mathcal{Y}^{\text{d}})$ and $J_0 = \{i \in [p] : \|\boldsymbol{Y}^{\text{d}}_{i:}\| = 0\}$. Set $z^{(t+1)}(i)$ randomly in $[r]$ for $i \in J_0$.

10: Let $\boldsymbol{S}^{(t)} = \text{Mat}(\mathcal{S}^{(t)})$. For all $i \in J_0^c$ update the cluster assignment by

$$z(i)^{(t+1)} = \arg\max_{a\in[r]} \cos\left(\boldsymbol{Y}^{\text{d}}_{i:},\ \boldsymbol{S}^{(t)}_{a:}\right).$$

11: **end for**

Output: Estimated clustering $z^{(T)} \in [r]^p$.

heterogeneity is comparable across clusters in that

$$\min_{a\in[r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\| = (1 + o(1)) \max_{a\in[r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|. \tag{6}$$

Note that, the assumption (6) does not preclude degree heterogeneity. Indeed, within each of the clusters, the highest degree can be $\theta(i) = \Omega(p)$, whereas the lowest degree can be $\theta(i) = \mathcal{O}(1)$.

Theorem 4 (Error for weighted higher-order initialization)**.** Consider the general sub-Gaussian ~~dTBM~~ dTBM with i.i.d. noise under the parameter space $\mathcal{P}$ and Assumption 1. Assume $\boldsymbol{\theta}$ is balanced and $\min_{i\in[p]} \theta(i) \ge c$ for some constant $c > 0$. Let $z^{(0)}$ denote the output of Sub-algorithm 1. With probability going to 1, we have

$$\ell(z^{(0)}, z) \lesssim \frac{r^K p^{-K/2}}{\text{SNR}}. \tag{7}$$

Remark 4 (Comparison to previous results)**.** For fixed SNR, our initialization error rate with $K = 2$ agrees with the initialization error rate ~~$\mathcal{O}(p)$~~ $\mathcal{O}(p^{-1})$ in matrix models (Gao et al., 2018). Furthermore, in the special case of non-degree TBMs with $\theta_1 = \cdots = \theta_p = 1$, we achieve the same initial misclustering error $\mathcal{O}(p^{-K/2})$ as in non-degree models (Han et al., 2020). Theorem 4 implies the advantage of our algorithm in achieving both accuracy and model flexibility.

Remark 5 (Failure of conventional tensor HOSVD)**.** If we use conventional HOSVD for tensor denoising; that is, we use $\boldsymbol{U}_{\text{pre}}$ in place of $\hat{\boldsymbol{U}}$ in line 2, then the misclustering rate becomes ~~$\mathcal{O}(p)$~~ $\mathcal{O}(p^{-1})$ for all $K \ge 2$. This rate is substantially worse than our current rate (7).

Remark 6 (Singular-value gap-free clustering)**.** Note that our ~~initialization works on~~ clustering directly applies

to the estimated mean tensor $\hat{\mathcal{X}}$ ~~directly~~ rather than the leading tensor ~~decomposition factors of $\mathcal{X}$. On one hand, clustering on $\hat{\mathcal{X}}$ avoids the eigen-gap assumption in previous work~~ Ke et al. ~~(2019). On the other hand, vanilla spectral methods working on decomposition factors suffers~~ factors $\hat{\boldsymbol{U}}$. Applying clustering to the tensor factors suffers from the non-identifiability ~~of the eigenspace with orthogonal transformation~~ issue due to the infinitely many orthogonal rotations when the number of blocks $r \geq 3$ in the absence of singular-value gaps. Such ambiguity ~~comes from the possible multiplicity of eigenvalue and causes~~ causes the trouble for effective clustering (Abbe et al., 2020). In contrast, our ~~eigen free strategy working on~~ initialization algorithm applies the clustering to the overall mean tensor $\hat{\mathcal{X}}$~~avoids such~~. This strategy avoids the non-identifiability issue regardless of the number of blocks and singular-value gaps.

B. Angle-based iteration

Our Theorem 4 has shown the polynomially decaying error rate from our initialization. Now we improve the error rate to exponential decay using local iterations. We propose an angle-based local iteration to improve the outputs from Sub-algorithm 1. To gain the intuition, consider an one-dimensional degree-corrected clustering problem with data vectors $\boldsymbol{x}_i = \theta(i)\boldsymbol{s}_{z(i)} + \boldsymbol{\epsilon}_i, i \in [p]$, where $\boldsymbol{s}_i$'s are known cluster centroids, $\theta(i)$'s are unknown positive degrees, and $z\colon [p] \mapsto [r]$ is the ~~clustering~~ cluster assignment of interest. The angle-based k-means algorithm estimates the assignment z by minimizing the angle between data vectors and centroids; i.e.,

$$z(i) = \arg\max_{a\in[r]} \cos(\boldsymbol{x}_i,\ \boldsymbol{s}_a), \quad \text{for all } i \in [p]. \tag{8}$$

The classical Euclidean-distance based clustering (Han et al., 2020) fails to recover z in the presence of degree heterogeneity, even under noiseless case. In contrast, the proposed angle-based k-means achieves accurate recovery without explicit estimation of $\boldsymbol{\theta}$.

Our Sub-algorithm 2 shares the same spirit as in angle-based k-means. We still take the order-3 tensor for illustration. Specifically, Sub-algorithm 2 updates estimated core tensor and cluster assignment in each iteration. We use superscript $.^{(t)}$ to denote the estimate from t-th iteration, where $t = 1, \ldots$. For core tensor, we consider the following update strategy

$$\mathcal{S}^{(t)}(a_1, a_2, a_3) = \text{Ave}\{\mathcal{Y}(i_1, i_2, i_3)\colon z^{(t)}(i_k) = a_k, k \in [3]\}.$$

Intuitively, $\mathcal{S}^{(t)}$ becomes closer to the true core $\mathcal{S}$ as $z^{(t)}$ is more precise. For cluster assignment, we first aggregate the slices of $\mathcal{Y}$ and obtain a reduced tensor $\mathcal{Y}^{\text{d}} \in \mathbb{R}^{p\times r\times r}$ with given $z^{(t)}$, where

$$\mathcal{Y}^{\text{d}}(i, a_2, a_3) = \text{Ave}\{\mathcal{Y}(i, i_2, i_3)\colon z^{(t)}(i_k) = a_k, k \neq 1\}.$$

~~The row $\text{Mat}(\mathcal{Y}^{\text{d}})_{i:}$ and $\text{Mat}(\mathcal{S}^{(t)})_{a:}$ corresponds~~ We use $\boldsymbol{Y}^d$ and $\boldsymbol{S}^{(t)}$ to denote the $\text{Mat}(\mathcal{Y}^{\text{d}})$ and $\text{Mat}(\mathcal{S}^{(t)})$. The rows $\boldsymbol{Y}^d_{i:}$ and $\boldsymbol{S}^{(t)}_{a:}$ correspond to the $\boldsymbol{x}_i$ and $\boldsymbol{s}_a$ ~~s~~ in the one-dimensional clustering (8). Then, we obtain the updated assignment ~~as~~ by

$$z(i)^{(t+1)} = \arg\max_{a\in[r]} \text{sincos}\left(\boldsymbol{Y}^{\text{d}}_{i:}, \boldsymbol{S}^{(t)}_{a:}\right), \quad \text{for all } i \in [p],$$

~~where $z^{(t+1)}(i)$ is randomly assigned for degenerate rows.~~ provided that $\boldsymbol{S}^{(t)}_{a:}$ is a non-zero vector. Otherwise, if $\boldsymbol{S}^{(t)}_{a:}$ is a zero vector, then we make the convention to assign $z^{(t+1)}(i)$ randomly in $[p]$. Full procedures for our angle-based iteration are described in Sub-algorithm 2.

We ~~then~~ now establish the misclustering error rate of iterations under the stability assumption.

Definition 2 (Locally linear stability)**.** Define the ε-neighborhood of z by $\mathcal{N}(z, \epsilon) = \{\bar{z}\colon \ell(\bar{z}, z) \leq \epsilon\}$. Let $\bar{z}\colon [p] \to [r]$ be a clustering function. We define two ~~cluster-size vectors for~~ vectors associated with $\bar{z}$,

$$\boldsymbol{p}(\bar{z}) = (|\bar{z}^{-1}(1)|, \ldots, |\bar{z}^{-1}(r)|)^T, \quad \boldsymbol{p}_{\boldsymbol{\theta}}(\bar{z}) = (\|\boldsymbol{\theta}_{\bar{z}^{-1}(1)}\|_1, \ldots, \|\boldsymbol{\theta}_{\bar{z}^{-1}(r)}\|_1)^T.$$

We call the degree is ε-locally linearly stable if and only if

$$\sin(\boldsymbol{p}(\bar{z}),\ \boldsymbol{p}_{\boldsymbol{\theta}}(\bar{z})) \lesssim \varepsilon\Delta_{\min}, \quad \text{for all } \bar{z} \in \mathcal{N}(z, \varepsilon). \tag{9}$$

Roughly speaking, the vector $\boldsymbol{p}(\bar{z})$ represents the raw cluster sizes, and $\boldsymbol{p}_\theta(\bar{z})$ represents the relative cluster sizes weighted by degrees. The local stability holds trivially for $\varepsilon = 0$ based on the construction of parameter space (3). ~~The locally linear stability avoids the concentration of entities with extremely large or small heterogeneity, when a good estimated assignment with a small misclustering error ϵ is given.~~ The condition (9) controls the impact of node degree to the $\boldsymbol{p}_\theta(\cdot)$ with respect to the misclassification rate ε and angle gap.

Theorem 5 (Error for angle-based iteration)**.** Consider the setup as in Theorem 4. Assume the local linear stability of degree holds in ~~all neighborhoods~~ the neighborhood $\mathcal{N}(z, \epsilon)$ for ~~any $\epsilon \le \log^{-1} p$~~ $\epsilon \gtrsim \log^{-1} p$. Suppose $r = \mathcal{O}(1)$ and $\mathrm{SNR} \gtrsim p^{-K/2} \log p$. Let $z^{(t)}$ denote the t-th iteration output in Sub-algorithm 2 with initialization $z^{(0)}$ from Sub-algorithm 1. With probability going to 1, there exists a contraction parameter $\rho \in (0, 1)$ such that

$$\ell(z, \hat{z}^{(t+1)}) \lesssim \underbrace{\mathrm{SNR}^{-1} \exp\left(-\frac{p^{K-1}\mathrm{SNR}}{r^{K-1}}\right)}_{\text{statistical error}} + \underbrace{\rho^t \ell(z, z^{(0)})}_{\text{computational error}}. \tag{10}$$

~~The~~ From the conclusion (10), we find that the iteration error is decomposed into two parts: statistical error and computational error. The statistical error is unavoidable with noisy data regardless t, whereas the computational error decays in an exponential rate as the number of iterations $t \to \infty$.

Theorem 5 implies that, with probability going to 1, our estimate $z^{(T)}$ achieves exact recovery within polynomial iterations; more precisely,

$$z^{(T)} = \pi \circ z, \quad \text{for all } T \gtrsim \log_{1/\rho} p,$$

for some permutation $\pi \in \Pi$. ~~We call our algorithm~~ Therefore, our combined algorithm is *computationally efficient* ~~with~~ as long as $\mathrm{SNR} \gtrsim p^{-K/2} \log p$. Note that~~the~~, ignoring the logarithmic term, the minimal SNR requirement, ~~$p^{-K/2} \log p$~~ $p^{-K/2}$, coincides with the computational lower bound in Theorem 3~~ignoring the logarithmic term~~. Therefore, our algorithm is optimal regarding the signal ~~demand~~ requirement and lies in the ~~most right "computationally efficient"~~ sharpest *computationally efficient* regime in Figure 2.

C. ~~*Extension*~~ *Extensions and practical* ~~*issue*~~*issues*

Extension for Bernoulli observations. The main difficulty to establish the statistical guarantee for Bernoulli observations lies in the initialization ~~Sub-Algorithm~~Sub-algorithm 1. Theorem 5 still holds for Bernoulli observations once the initialization accuracy satisfies the upper bound (7) in Theorem 4.

~~Specifically, the~~ We now provide a high-level explanation for the technical difficulty when applying Theorem 4 to Bernoulli observations. The derivation of Theorem 4 relies on the upper bound of the estimation error for the mean tensor in Lemma **??**~~,~~; i.e., with high probability

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim p^{K/2}, \tag{11}$$

where ~~$\mathcal{X} = \mathbb{E}[\mathcal{Y}]$~~ $\mathcal{X} = \mathbb{E}\mathcal{Y}$ and $\hat{\mathcal{X}}$ is defined in Step 2 of Sub-algorithm 1. Unfortunately, the inequality (11) ~~only holds~~ holds only for i.i.d. sub-Gaussian observations while Bernoulli observations are generally not identically distributed.

One possible remedy is to apply singular value decomposition to the ~~unfolded Bernoulli observation~~ *square unfolding* (Mu et al., 2014) of Bernoulli tensor $\mathcal{Y}$. Let the matrix $\mathrm{Mat}_{sq}(\mathcal{Y}) \in \{0,1\}^{\lfloor p^{K/2} \rfloor \times \lceil p^{K/2} \rceil}$ denote the ~~square unfolded binary tensor. We have the estimate~~ nearly square unfolded Bernoulli tensor. Define a new estimate

$$\hat{\mathcal{X}}' = \underset{\mathrm{rank}(\mathrm{Mat}_{sq}(\mathcal{X})) \le r^{\lceil K/2 \rceil}}{\arg\min} \|\mathrm{Mat}_{sq}(\mathcal{X}) - \mathrm{Mat}_{sq}(\mathcal{Y})\|_F^2. \tag{12}$$

The optimization (12) is simply a matrix SVD problem. Following Lemma 7 in Gao et al. (2018), with high probability, ~~we have~~ the new estimate satisfies

$$\|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2 \lesssim p^{\lceil K/2 \rceil}.$$

Replacing the estimate $\hat{\mathcal{X}}$ by $\hat{\mathcal{X}}'$ in Theorem 4, the high probability upper bound for Bernoulli initialization is

$$\ell(z^{(0)}, z) \lesssim \frac{r^K p^{-\lfloor K/2 \rfloor}}{\text{SNR}}. \tag{13}$$

The Bernoulli bound (13) is relatively looser than Gaussian bound (7), especially when K is small. ~~A tighter Bernoulli bound will be achieved once a low-rank binary tensor estimation scheme with better accuracy is provided in the future.~~ Nevertheless, our bound (13) is already tighter than the pervious work (Ke et al., 2019). The investigation of the gap between upper bound $p^{-\lfloor K/2 \rfloor}$ and the lower bound $p^{-K/2}$ for Bernoulli tensors will be left as future work.

Extension for general dTBMs. Our two-stage algorithm is able to be extended for the general (asymmetric) dTBMs. Specifically, in the ~~Sub-Algorithm~~Sub-algorithm 1, we make the following changes: (1) Replace the matrcization $\text{Mat}(\mathcal{Y})$ by $\text{Mat}_k(\mathcal{Y})$; (2) Repeat the Steps 1-5 with mode-specified number of ~~cluster~~ clusters r_k; (3) Obtain the collection initialization $\{z_k^{(0)}\}_{k=1}^K$. In the ~~Sub-Algorithm~~Sub-algorithm 2, we make the following changes: (1) Take the collection $\{z_k^{(0)}\}_{k=1}^K$ as input, and update the block tensor $\mathcal{S}^{(t)}$ with the collection $\{z_k^{(t)}\}_{k=1}^K$, $\mathcal{S}^{(t)}(i_1, \ldots, i_K) = \text{Ave}\{\mathcal{Y}(i_1, \ldots, i_K) : z_k^{(t)}(i_k) = j_k, k \in [K]\}$; (2) Calculate reduced tensor $\mathcal{Y}_k^{\text{d}}$ for each mode via

$$\mathcal{Y}_k^{\text{d}}(a_1, \ldots, a_{k-1}, i, a_{k+1}, \ldots, a_K) = \text{Ave}\{\mathcal{Y}(i_1, \ldots, i_{k-1}, i, i_{k+1}, \ldots, i_K) : z^{(t)}(i_j) = a_j, j \neq k\};$$

(3) Repeat Step 8-10 with $\text{Mat}_k(\cdot), \mathcal{Y}_k^{\text{d}}$ for each $k \in [K]$ and obtain the collection $\{z_k^{(T)}\}_{k=1}^K$.

Correspondingly, Theorems 4 and 5 still hold with $\ell(z_k^{(0)}, z_k)$ and $\ell(z_k^{(t+1)}, z_k)$ for all $k \in [K]$. The detailed model extension for general asymmetric dTBMS can be found in Appendix.

Computational ~~Complexity~~complexity. Our two-stage algorithm has a ~~polynomial computational cost~~ computational cost polynomial in tensor dimension p. Specifically, the complexity of ~~Sub-Algorithm~~Sub-algorithm 1 is $\mathcal{O}(Kp^{K+1} + Krp^K)$, where the first term is contributed by the ~~first step SVD~~ double projection and the calculation of $\hat{\mathcal{X}}$, and the second term comes from normalization and the k-means. The cost of each update in ~~Sub-Algorithm~~Sub-algorithm 2 is $\mathcal{O}(p^K + pr^K)$, where p^K comes from the calculation of $\mathcal{S}^{(t)}$ and $\mathcal{Y}^{\text{d}}$, and pr^K comes from the normalization of ~~$\mathcal{Y}^{\text{d}}, \mathcal{S}^{(t)}$~~ $\mathcal{Y}^{\text{d}}$, the calculation of $\mathcal{S}^{(t)}$, and the cluster assignment update in Step 10.

~~Parameter Selection~~Hyper-parameter selection. ~~Note that we assume the~~ In our theoretical analysis, we have assumed the true number of clusters r is given ~~for the algorithm.~~ to our algorithm. In practice, the number of clusters r is often unknown, and we now propose a method to choose r from data. We impose the Bayesian information criterion (BIC) and choose the ~~number of cluster~~ cluster number that minimizes BIC~~,~~; i.e., under the symmetric Gaussian dTBM (2),

$$\hat{r} = \arg\min_{r \in \mathbb{Z}_+} \left(-p^K \mathcal{L}_{\mathcal{Y}}(\hat{z}(r), \hat{\mathcal{S}}(r), \hat{\boldsymbol{\theta}}(r)) + p_e(r) K \log p \right),$$

$$\hat{r} = \arg\min_{r \in \mathbb{Z}_+} \left(p^K \log(\|\hat{\mathcal{X}} - \mathcal{Y}\|_F^2) + p_e(r) K \log p \right), \tag{14}$$

$$\text{where} \quad \hat{\mathcal{X}} = \hat{\mathcal{S}}(r) \times_1 \hat{\boldsymbol{\Theta}}(r) \hat{\boldsymbol{M}}(r) \times_2 \cdots \times_K \hat{\boldsymbol{\Theta}}(r) \hat{\boldsymbol{M}}(r),$$

where the triplet ~~$(\hat{z}(r), \hat{\mathcal{S}}(r), \hat{\boldsymbol{\theta}}(r))$~~ $(\hat{z}(r), \hat{\boldsymbol{S}}(r), \hat{\boldsymbol{\theta}}(r))$ are estimated parameters with ~~given number of cluster~~ cluster number r, ~~$\mathcal{L}_{\mathcal{Y}}$ denote the log-likelihood with observations $\mathcal{Y}$, and $p_e = r^K + p(\log r + 1) - r$~~ and $p_e(r) = r^K + p(\log r + 1) - r$ is the effective number of parameters. ~~Particularly, we obtain the~~ Note that we have added the argument (r) to related quantities as functions of r. In particular, the estimate $\hat{\boldsymbol{\theta}}(r)$ in (14) is obtained by first calculating the reduced tensor ~~$\mathcal{Y}^{\text{d}}$~~ $\hat{\mathcal{Y}}^{\text{d}}$ with $\hat{z}(r)$, and then normalizing the row norms ~~$\|\boldsymbol{Y}_{i:}^{\text{d}}\|$~~ $\|\hat{\boldsymbol{Y}}_{i:}^{\text{d}}\|$ to 1 in each cluster; i.e.,

$$\hat{\boldsymbol{\theta}}(r) = (\theta($$

where $\hat{\boldsymbol{Y}}^{\mathrm{d}}(r) = \mathrm{Mat}(\hat{\mathcal{Y}}^{\mathrm{d}}(r))$~~and~~

$$\hat{\mathcal{Y}}^{\mathrm{d}}(r)(i, a_2, \ldots, a_K) = \mathrm{Ave}\{\mathcal{Y}(i, i_2, \ldots, i_K) : \hat{z}(i_k, r) = a_k, k \neq 1\}.$$

~~With Gaussian observation $\mathcal{Y}$, we have~~

$$\mathcal{L}_{\mathcal{Y}} = -\log(\|\hat{\mathcal{X}} - \mathcal{Y}\|_F^2), \quad \hat{\mathcal{X}} = \mathcal{S}(r) \times_1 \hat{\boldsymbol{\Theta}}(r)\hat{\boldsymbol{M}}(r) \times_2 \cdots \times_K \hat{\boldsymbol{\Theta}}(r)\hat{\boldsymbol{M}}(r).$$

, $\hat{\mathcal{Y}}^{\mathrm{d}}(r)(i, a_2, \ldots, a_K) = \mathrm{Ave}\{\mathcal{Y}(i, i_2, \ldots, i_K) : \hat{z}(i_k, r) = a_k, k \neq 1\}$, and $\hat{z}(i, r)$ denotes the community label for the i-th node with given cluster number r. We evaluate the performance of the BIC criterion in Section V-A.

V. Numerical studies

We evaluate the performance of the weighted higher-order initialization and angle-based iteration in this section. We report average errors and standard deviations across 30 replications in each experiment. Clustering accuracy is assessed by clustering error rate (CER, i.e., one minus rand index). Note that CER between $(\hat{z}, z)$ is equivalent to misclustering error $\ell(\hat{z}, z)$ up to constant multiplications (Meilă, 2012), and a lower CER indicates a better performance.

We generate order-3 tensors with *assortative* (Gao et al., 2018) core tensors to control SNR; i.e., we set $\mathcal{S}_{aaa} = s_1$ for $a \in [r]$ and others be s_2, where $s_1 > s_2 > 0$. Let $\alpha = s_1/s_2$. We set α close to 1 such that $1 - \alpha = o(p)$. In particular, we have $\alpha = 1 + \Omega(p^{\gamma/2})$ with $\gamma < 0$ by Assumption 1 and definition (5). Hence, we easily adjust SNR via varying α. Note that the assortative setting is proposed for simulations, and our algorithm is applicable for general tensors in practice. The cluster assignment z is randomly generated with equal probability across r clusters for each mode. Without further explanation, we generate degree heterogeneity $\boldsymbol{\theta}$ from absolute normal distribution ~~as~~ by $\theta(i) = |X_i| + 1 - 1/\sqrt{2\pi}$ with $|X_i| \overset{\text{i.i.d.}}{\sim} N(0, 1), i \in [p]$ and normalize $\boldsymbol{\theta}$ to satisfy (3). Also, we set $\sigma^2 = 1$ for Gaussian data without further specification.

A. Verification of theoretical results

The first experiment verifies statistical-computational gap described in Section III. Consider the Gaussian model with $p = \{80, 100\}$, $r = 5$. We vary γ in $[-1.2, -0.4]$ and $[-2.1, -1.4]$ for matrix ($K = 2$) and tensor ($K = 3$) clustering, respectively. Note that finding MLE under dTBM is computationally intractable. We approximate MLE using an oracle estimator~~;~~, i.e., the output of Sub-algorithm 2 initialized from true assignment. Figure 4a shows that both our algorithm and oracle estimator start to decrease around the critical value $\gamma_{\text{stat}} = \gamma_{\text{comp}} = -1$ in matrix case. In contrast, Figure 4b shows a significant gap in the phase transitions between the algorithm estimator and oracle estimator in tensor case. The oracle error rapidly decreases to 0 when $\gamma_{\text{stat}} = -2$, whereas the algorithm estimator tends to achieve exact clustering when $\gamma_{\text{comp}} = -1.5$. Figure 4 confirms the existence of the statistical-computational gap in our Theorems 2 and 3.

The second experiment verifies the performance guarantees of two algorithms: (i) weighted higher-order initialization; (ii) combined algorithm of weighted higher-order initialization and angle-based iteration. We consider both the Gaussian and Bernoulli models with $p = \{80, 100\}$, $r = 5$, $\gamma \in [-2.1, -1.4]$. Figure 5 shows the substantial improvement of combined algorithm over initialization, especially under weak and intermediate signals. This phenomenon agrees with the error rates in Theorems 4 and 5 and confirms the necessity of the local iterations.

The third experiment evaluates the empirical performance of the BIC criterion to select unknown number of ~~cluster~~clusters. We generate the data from an order-3 Gaussian model with $p = \{50, 80\}$, $r = \{2, 4\}$~~and consider the noise~~, and noise level $\sigma^2 \in \{0.25, 1\}$~~with $\alpha = 400$~~. Table II ~~implies~~ shows that our BIC criterion ~~exactly choose~~ well chooses the true r under ~~all the settingswith small $r = 2$~~most settings. Note that the BIC ~~underestimates the large~~ slightly underestimates the true number of clusters ($r = 4$) with smaller ~~SNR ($\sigma^2 = 1$)~~dimension and higher noise ($p = 50, \sigma = 1$), and the accuracy immediately increases with larger dimension $p = 80$. The improvement

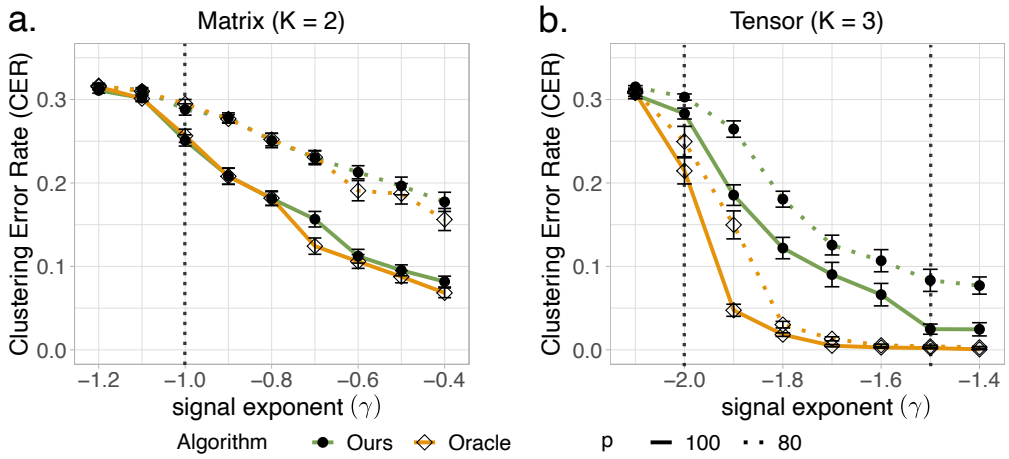

Fig. 4: SNR phase transitions for clustering in dTBM with $p = \{80, 100\}, r = 5$ under (a) matrix case with $\gamma \in [-1.2, -0.4]$ and (b) tensor case with $\gamma \in [-2.1, -1.4]$.

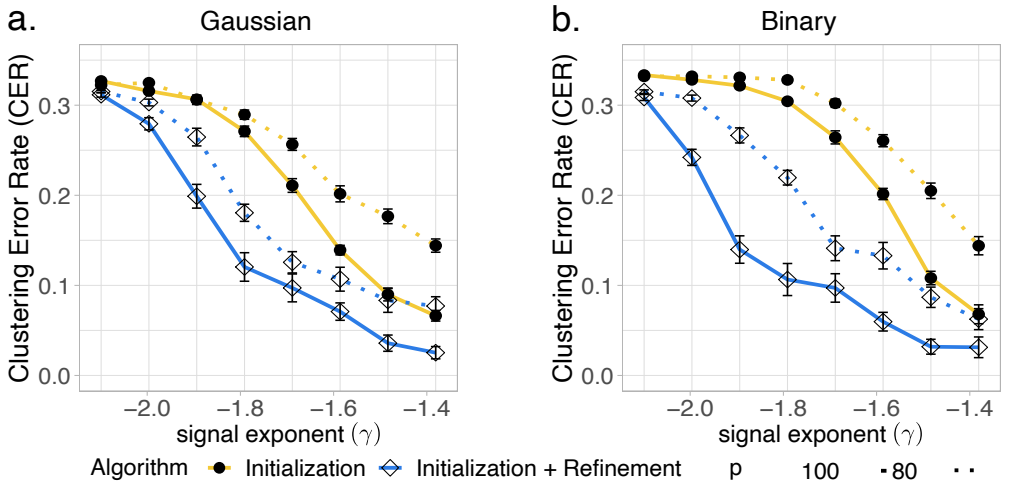

Fig. 5: CER versus signal exponent (γ) for initialization only and for combined algorithm. We set $p = \{80, 100\}, r = 5, \gamma \in [-2.1, -1.4]$ under (a) Gaussian models and (b) Bernoulli models.

follows ~~by~~ from the fact that a larger dimension p indicates a larger sample size in the tensor block model. Therefore, we conclude that BIC criterion is a reasonable way to tune the number of clusters.

B. Comparison with other methods

We compare our algorithm with following higher-order clustering methods:

Settings	$p = 50, \sigma^2 = 0.25$		$p = 50, \sigma^2 = 1$		$p = 80, \sigma^2 = 0.25$		$p = 80, \sigma^2 = 1$	
True number of clusters r	2	4	2	4	2	4	2	4
Estimated number of clusters $\hat{r}$	2(0)	3.9(0.25)	2(0)	3.1(0.52)	2(0)	4(0)	2(0)	3.9(0.31)

TABLE II: Estimated number of clusters given by BIC criterion under the ~~small~~ low noise $(\sigma^2 = 0.25)$ and ~~large~~ high noise $(\sigma^2 = 0.5)$ ~~cases~~settings. Numbers in parentheses are standard deviations of $\hat{r}$ over 30 replications.

- **HOSVD**: HOSVD on data tensor and k-means on the rows of the factor matrix;
- **HOSVD+**: HOSVD on data tensor and k-means on the ℓ_2-normalized rows of the factor matrix;
- **HLloyd** (Han et al., 2020): High-order ~~Lloyd algorithm and high-order spectral clustering (Han et al., 2020)~~clustering algorithm developed for non-degree tensor block models;
- **SCORE** (Ke et al., 2019): Tensor-SCORE for clustering ~~(Ke et al., 2019);~~ developed for binary tensors.

Among the four alternative algorithms, the **SCORE** is the closest method to ours. We set the tuning parameters of **SCORE** as in previous literature (Ke et al., 2019). The methods **SCORE** and **HOSVD+** are designed for degree models, whereas **HOSVD** and **HLloyd** are designed for non-degree models. We conduct two experiments to assess the impacts of (i) signal strength and (ii) degree heterogeneity, based on Gaussian and Bernoulli models with $p = 100, r = 5$. We refer to our algorithm as **dTBM** in the comparison.

We investigate the effects of signal to clustering performance by varying $\gamma \in [-1.5, -1.1]$. Figure 6 shows the consistent outperformance of our method **dTBM** among all algorithms. The sub-optimality of **SCORE** and **HOSVD+** indicates the necessity of local iterations on the clustering. Furthermore, Figure 6 shows the inadequacy of non-degree algorithms in the presence of mild degree heterogeneity. The experiment demonstrates the benefits of addressing heterogeneity in higher-order clustering tasks.

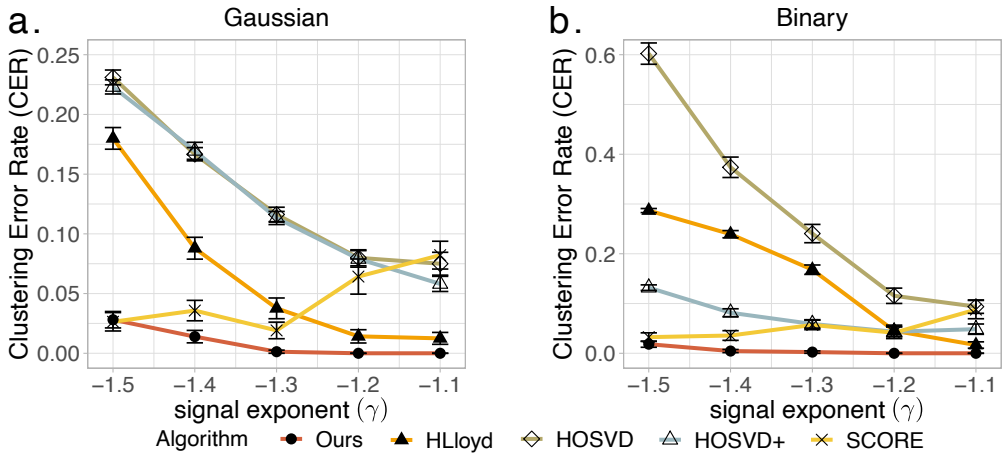

Fig. 6: CER versus signal exponent (denoted γ) for different methods. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under (a) Gaussian and (b) Bernoulli models.

The only exception in Figure 6 is the slightly better performance of **HLloyd** over **HOSVD+** under Gaussian model. However, we find the advantage of **HLloyd** disappears with higher degree heterogeneity. We perform extra simulations to verify the impact of degree effects. We use the same setting as in the first experiment in the Section V-B, except that we now generate the degree heterogeneity $\boldsymbol{\theta}$ from Pareto distribution ~~with shape parameter a~~ prior to normalization. ~~We~~ The density function of Pareto distribution is $f(x|a,b) = ab^a x^{-(a+1)} \mathbb{1}\{x \geq b\}$, where a is called *shape* parameter. We vary $a \in \{2, 6\}$ and choose b such that $\mathbb{E}X = a(a-1)^{-1}b = 1$ for X following Pareto(a, b). Note that a smaller a leads to a larger variance in $\boldsymbol{\theta}$ and hence a larger degree heterogeneity. We consider the Gaussian model under low $(a = 6)$ and high $(a = 2)$ degree heterogeneity. Figure 8 shows that the errors for non-degree algorithms (**HLloyd**, **HOSVD**) ~~increases~~ increase with degree heterogeneity. In addition, the advantage of **HLloyd** over **HOSVD+** disappears with higher degree heterogeneity.

The last experiment investigates the effects of degree heterogeneity to clustering performance. We fix the signal exponent $\gamma = -1.2$ and vary the extent of degree heterogeneity. In this experiment, we generate $\boldsymbol{\theta}$ from Pareto distribution prior to normalization. ~~The density function of Pareto distribution is $f(x|a,b) = ab^a x^{-(a+1)} \mathbb{1}\{x \geq b\}$, where a is called *shape* parameter. We vary~~ We vary the shape parameter $a \in [3, 6]$ ~~and choose b such that $\mathbb{E}[X] = a(a-1)^{-1}b = 1$ for X following Pareto (a, b). Note that a smaller a leads to a larger variance in $\boldsymbol{\theta}$ and hence a larger degree heterogeneity~~in the Pareto distribution to investigate a range of degree heterogeneities. Figure 7 demonstrates the stability of degree-corrected algorithms (**dTBM**, **SCORE**, **HOSVD+**) over the entire range of degree

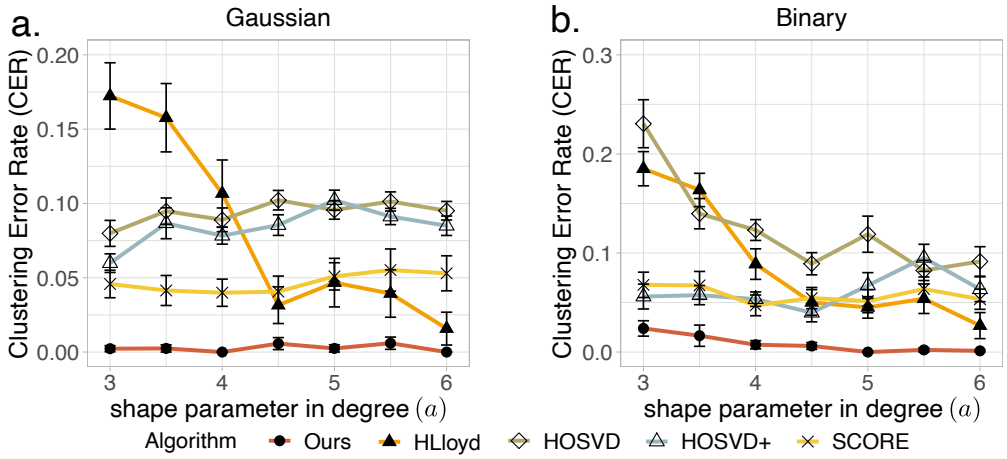

Fig. 7: CER versus shape parameter in degree (denoted $a \in [3, 6]$) for different methods. We set $p = 100, r = 5, \gamma = -1.2$ under (a) Gaussian and (b) Bernoulli models.

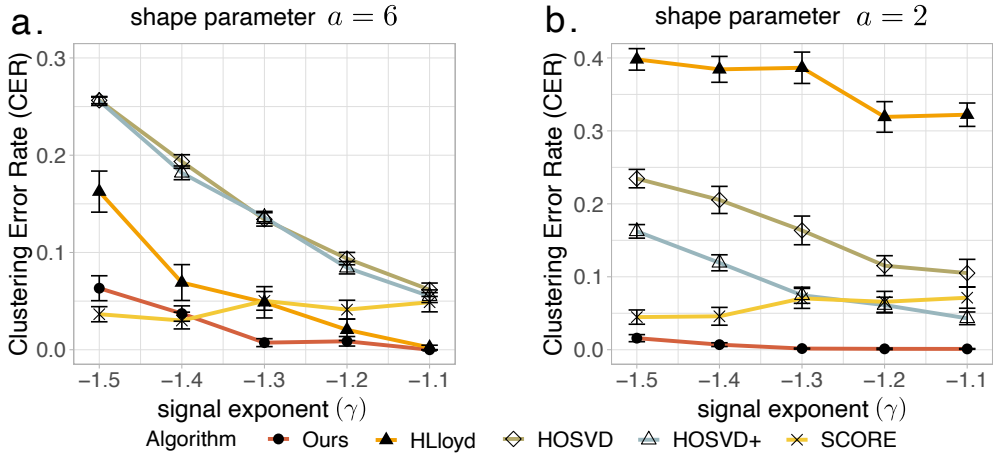

Fig. 8: CER comparison versus signal exponent (denoted γ) under (a) low (shape parameter $a = 6$) (b) high (shape parameter $a = 2$) degree heterogeneity. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under Gaussian model.

heterogeneity under consideration. In contrast, non-degree algorithms (**HLloyd**, **HOSVD**) show poor performance with large heterogeneity, especially in Bernoulli cases. This experiment, again, highlights the benefit of addressing degree heterogeneity in higher-order clustering.

VI. Real data applications

A. Human brain connectome data analysis

The Human Connectome Project (HCP) aims to construct the structural and functional neural connections in human brains (Van Essen et al., 2013). We preprocess the original dataset following Desikan et al. (2006) and partition the brain into 68 regions. The cleaned ~~data~~ dataset includes brain networks for 136 individuals. Each brain network is represented by a 68-by-68 binary symmetric matrix, where the ~~entries~~ entry with value 1 ~~refers to~~ indicates the presence of connection ~~among 68 nodes while~~ between node pairs, while the value 0 ~~refers to~~ indicates the absence. We use $\mathcal{Y} \in \{0, 1\}^{68\times68\times136}$ to denote the binary ~~observation~~tensor. Individual attributes such as gender and sex are recorded.

We apply our generalized ~~Algorithm~~ algorithm to the HCP data with the ~~number~~ numbers of clusters on three modes $r_1 = r_2 = 4$ and $r_3 = 3$. The selection of r_1 and r_2 follows the human brain anatomy and the symmetry in the brain network, and the r_3 is ~~chosen to be small~~ specified following previous analysis (Hu et al., 2021). ~~The~~ Because of the symmetry in the data, the estimated brain node clustering results are the same on the first and second ~~mode are the same~~modes. Figure 9 ~~indicates~~ shows that brain connection exhibits a strong spatial separation structure. Specifically, the first cluster, named *L.Hemis*, involves all the nodes in the left hemisphere. The nodes in the right hemisphere are further separated into three clusters led by the middle-part tissues in Temporal and Parietal lobes (*R.Temporal*), the back-part tissues in Occipital lobe (*R.Occipital*), and the front-part tissues in Frontal and Parietal lobes (*R.Supra*). This clustering result is ~~consistent with the common sense that the~~ reasonable since the left and right hemispheres often play different roles in human ~~brain~~brains.

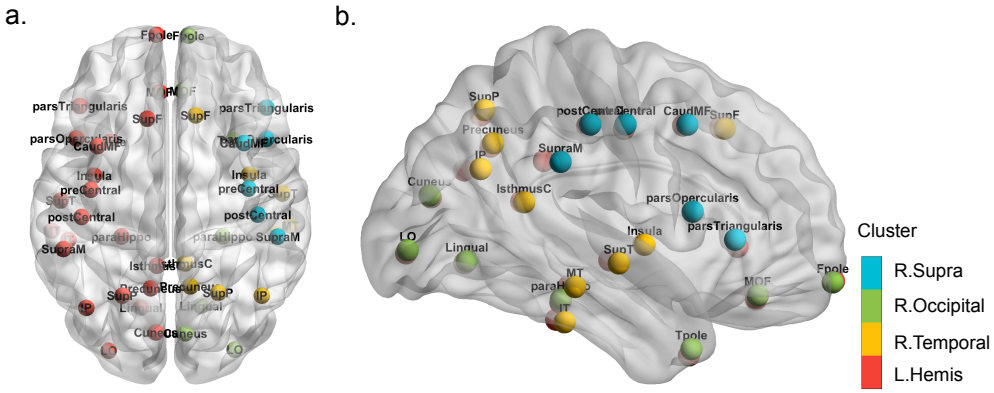

Fig. 9: Illustration of brain node clustering results for HCP data with (a) top and (b) side views.

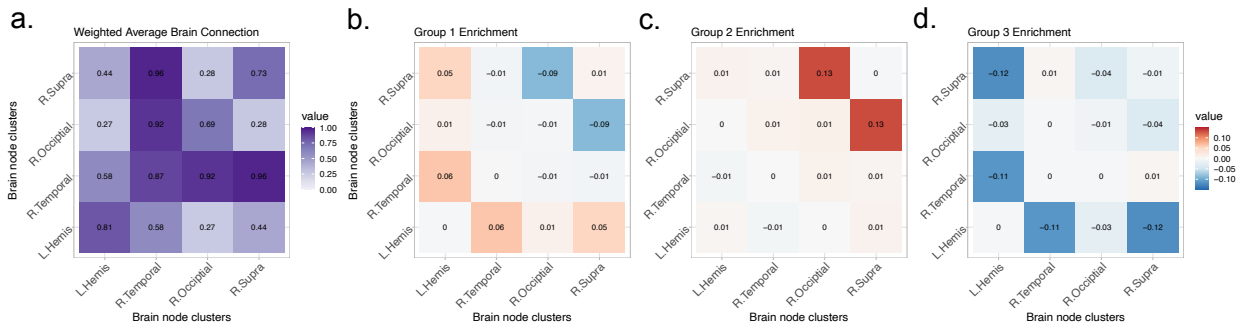

Fig. 10: Mode 3 slices of estimated core tensor $\hat{\mathcal{S}}$. (a) Average estimated slice weighted by the group size; (b)-(d) Group-specified enrichment, i.e., the ~~subtraction~~ difference between each slice of $\hat{\mathcal{S}}$ and the averaged slice.

Figure 10 illustrates the estimated core tensor $\hat{\mathcal{S}}$ with estimated clustering, and Figure 11 visualizes the average ~~observed~~ brain connections and the connection enrichment ~~with average observed~~ in contrast to average networks in each group. In general, we find that the inner-hemisphere connection has stronger connection compared to inter-hemisphere connections (Figure 10a). Also, the back and front parts (*R.Occipital*, *R.Supra*) are shown to have more interactions with temporal tissues than inner-cluster connections. In addition, the group 1 with 54% females ~~implies~~ shows an enrichment on the inter-hemisphere connections (Figure 10b), while group 4 with only 36% females exhibits a reduction (Figure 10d). This result agrees with previous findings in Hu et al. (2021). The enrichment on the back-front connection is also recognized in group 3 (Figure 10c). The interpretive ~~pattern in~~ patterns in our results demonstrate the usefulness of our clustering methods in the human brain connectome data application.

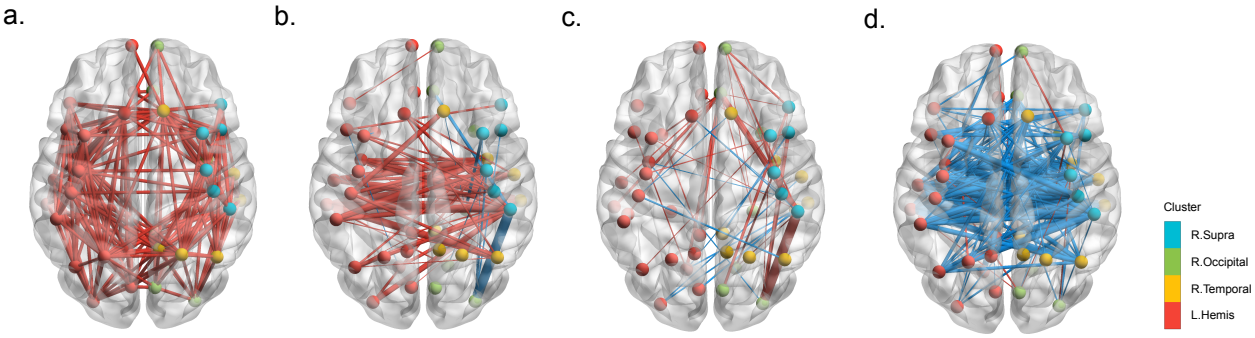

Fig. 11: Observed brain connections in the population and each group of individuals. (a) Average brain network; (b)-(d) Group-specified brain ~~networks enrichment~~ network enrichments in Groups 1-3. Red edges ~~refer to~~ represent the positive enrichment and blue edges ~~refer~~ represent the negative ~~reduction~~enrichment.

B. Peru Legislation data analysis

We also apply our method to the legislation networks in the Congress of the Republic of Peru (Lee et al., 2017). Because of the frequent political power shifts in the Peruvian Congress during 2006-2011, we choose to focus on the data for the first half of 2006-2007 year. The dataset records the co-sponsorship of 116 legislators from top 5 parties and 802 bill proposals. We reconstruct legislation network as an order-3 binary tensor $\mathcal{Y} \in \{0,1\}^{116\times116\times116}$, where $\mathcal{Y}_{ijk} = 1$ if the legislators (i, j, k) have sponsored the same bill, and $\mathcal{Y}_{ijk} = 0$ otherwise. The true party affiliations of legislators are provided and serve as the ground truth. We apply various higher-order clustering methods to $\mathcal{Y}$ with $r = 5$. Table III shows that our **dTBM** achieves the best performance compared to others. The second best method is the two-stage algorithm **HLloyd**, followed by the spectral methods **SCORE** and **HOSVD+**. This result is consistent with our simulations under strong signal and moderate degree heterogeneity. The comparison suggests that our method **dTBM** is more appealing in real-world applications.

Method	**dTBM**	**HOSVD**	**HOSVD+**	**HLloyd**	**SCORE**
CER	**0.116**	0.22	0.213	0.149	0.199

TABLE III: Clustering errors (measured by CER) for various methods in the analysis of Peru Legislation dataset.

VII. ~~Conclusion~~

~~We have developed a general degree-corrected tensor block model with a two-step angle-based polynomial-times algorithm. We have, for the first time, characterized the statistical and computational behaviors of the degree-corrected tensor block model under different signal-to-noise ratio regimes. Simulations and Peru Legislation and Human brain connection data analysis confirm the potential of our method for practical applications.~~

VII. Proof Sketches

In this section, we provide the proof sketches for ~~Theorem 4~~and ~~Theorem~~ the main Theorems 4-5. Detail proofs and extra theoretical results are provided in Appendix.

A. Proof sketch of Theorem 4

The proof of Theorem 4 is ~~mainly~~ inspired by the proof idea of Gao et al. (2018, Lemma 1)~~, and extra difficulties due to~~. The extra difficulties are the angle gap ~~assumption and tensor property are addressed in~~ characterization and multilinear algebra property in tensors; we address both challenges in our proof. Specifically, we control the

misclustering error by the estimation error of $\hat{\mathcal{X}}$ calculated in Step ~~2.~~ 2 of Sub-algorithm 1. We prove the following inequality

$$\ell(z^{(0)}, z) \lesssim \frac{1}{p} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^K} \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim \frac{r^K p^{-K/2}}{\mathrm{SNR}}, \tag{15}$$

where ~~$\mathcal{X} = \mathbb{E}[\mathcal{Y}]$~~ $\mathcal{X} = \mathbb{E}\mathcal{Y}$ is the true mean. The first inequality in (15) holds with the assumption ~~$\min_{i \in [p]} \theta(i) \geq c$~~ $\min_{i \in [p]} \theta(i) \geq c > 0$ in Theorem 4. The second inequality relies on ~~the intermediate~~ an important conclusion that the angle gap of mean tensor $\mathcal{X}$ is lower bounded by ~~a function~~ that of core tensor $\mathcal{S}$, i.e., the minimal angle gap $\Delta_{\min}$ defined in Assumption 1. Let $\boldsymbol{a}^s := \boldsymbol{a}/\|\boldsymbol{a}\|$ denote the normalized vector ~~and make~~ with the convention that $\boldsymbol{a}^s = 0$ if $\boldsymbol{a} = 0$. We want to show that

$$\min_{z(i) \neq z(j)} \|[\boldsymbol{X}_{i:}]^s - [\boldsymbol{X}_{j:}]^s\| \gtrsim \Delta_{\min}, \tag{16}$$

where $\boldsymbol{X} = \mathrm{Mat}(\mathcal{X})$. The most challenging part in the proof of Theorem 4 lies in the derivation of inequality (16), in which the proof of Gao et al. (2018) is no longer applicable due to different angle gap assumption in our dTBM. ~~Extra~~ We develop the extra padding technique in Lemma **??** and balance assumption (6) ~~are equipped~~ to derive (16). Last, we finish the proof of Theorem 4 by showing the third inequality ~~with~~ of (15) using Han et al. (2020, Proposition 1).

B. Proof sketch of Theorem 5

The proof of Theorem 5 is ~~mainly~~ inspired by the proof idea of Han et al. (2020, Theorem 2)~~, and extra polar coordinate based techniques are imposed due to~~ . We develop extra polar-coordinate based techniques with angle gap characterization to address the nuisance degree heterogeneity~~and angle gap assumption. We conduct the contraction property of the~~ . We introduce an intermediate quantity called misclustering loss

$$L^{(t)} = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1}\left\{z^{(t)}(i) = b\right\} \left\| \left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s \right\|^2,$$

~~and~~ where the superscript $\cdot^s$ denotes the normalized vector; i.e., $\boldsymbol{a}^s := \boldsymbol{a}/\|\boldsymbol{a}\|$ if $\boldsymbol{a} \neq 0$ and $\boldsymbol{a}^s = 0$ if $\boldsymbol{a} = 0$ for any vector $\boldsymbol{a}$. We show that $L^{(t)}$ provides an upper bound for the misclassification error of interest via the inequality $\ell^{(t)} \lesssim \frac{L^{(t)}}{\Delta_{\min}^2}$. Therefore, it suffices to control $L^{(t)}$. Further, we introduce the oracle estimators for core tensor ~~and assignment~~ under the true cluster assignment via

$$\tilde{\mathcal{S}} = \mathcal{Y} \times_1 \boldsymbol{W}^T \times_2 \cdots \times_K \boldsymbol{W}^T,$$

where ~~$\boldsymbol{W} = \boldsymbol{M}\left(\mathrm{diag}(\mathbb{1}_p^T \boldsymbol{M})\right)^{-1}$~~ $\boldsymbol{W} = \boldsymbol{M}\left(\mathrm{diag}(\mathbf{1}_p^T \boldsymbol{M})\right)^{-1}$ is the weighted true membership matrix. Let $\boldsymbol{V} = \boldsymbol{W}^{\otimes(K-1)}$ denote the Kronecker product of ~~$K-1$~~ $(K-1)$ copies of $\boldsymbol{W}$ matrices, and ~~similarly define~~ we define the t-th iteration quantities $\boldsymbol{W}^{(t)}, \boldsymbol{V}^{(t)}$ ~~with $\boldsymbol{M}^{(t)}$ corresponding to~~ corresponding to $\boldsymbol{M}^{(t)}$ (or equivalently $z^{(t)}$). To evaluate $L^{(t+1)}$, we ~~consider the event~~ prove the bound

$$\mathbb{1}\left\{z^{(t+1)}(i) = b\right\} = \mathbb{1}\left\{\|[\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2 \leq \|[\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2\right\} \leq A_{ib} + B_{ib}, \tag{17}$$

where $\boldsymbol{Y} = \mathrm{Mat}(\mathcal{Y})$, $\boldsymbol{S} = \mathrm{Mat}(\mathcal{S})$, $\boldsymbol{S}^{(t)} = \mathrm{Mat}(\mathcal{S}^{(t)})$ and

$$A_{ib} = \mathbb{1}\left\{\left\langle \boldsymbol{E}_{i:}\boldsymbol{V}, \left[\tilde{\boldsymbol{S}}_{z(i):}\right]^s - \left[\tilde{\boldsymbol{S}}_{b:}\right]^s \right\rangle \lesssim -\left\| \left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s \right\|^2\right\},$$
$$B_{ib} = \mathbb{1}\left\{\left\| \left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s \right\|^2 \lesssim F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)}\right\}.$$

The terms $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$ are controlled by $z^{(t)}, \mathcal{S}^{(t)}$~~, and detail definitions are~~ ; see the detailed definitions in (**??**), (**??**), (**??**). Note that the event A_{ib} only involves the oracle estimator independent of t, while all the terms related to the t-th iteration ~~estimator~~ are in B_{ib}. Thus, the inequality (17) decomposes the misclustering loss in the $(t+1)$-th iteration ~~by~~ into the oracle loss and the loss in t-th iteration. This decomposition leads to the separation of statistical error and computational error in the final upper bound of Theorem 5.

Specifically, we prove the contraction inequality

$$L^{(t+1)} \lesssim \xi + \rho L^{(t)}, \quad \xi = \frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]} A_{ib}\left\|\left[\boldsymbol{S}_{z(i)\cdot}\right]^s - \left[\boldsymbol{S}_{b\cdot}\right]^s\right\|^2, \tag{18}$$

where $\rho \in (0,1)$ is the contraction parameter, and we call ξ the oracle loss. Controlling the probability of event B_{ib} and obtaining the $\rho L^{(t)}$ term in the right hand side of (18) ~~is the most tricky and challenging part~~ are the most challenging parts in the proof of Theorem 5. Note that the true and estimated core tensors are involved ~~as~~ via their normalized rows such as $\boldsymbol{S}^s_{a:}, \tilde{\boldsymbol{S}}^s_{a:}, [\boldsymbol{S}^{(t)}_{a:}]^s$. The Cartesian coordinate based analysis in Han et al. (2020) is no longer applicable in our case. Instead, we use the ~~polar coordinate~~ polar-coordinate based analysis and the geometry property of trigonometric functions to derive the high probability upper bounds for $F^{(t)}_{ib}, G^{(t)}_{ib}, H^{(t)}_{ib}$.

Further, by sub-Gaussian concentration, we prove the high probability upper bound for oracle loss

$$\xi \lesssim \exp\left(-\frac{p^{K-1}\mathrm{SNR}}{r^{K-1}}\right). \tag{19}$$

Combining the decomposition (18) and the oracle bound (19), we finish the proof of Theorem 5.

ACKNOWLEDGMENTS

This research is supported in part by NSF grants DMS-1915978, DMS-2023239, EF-2133740, and funding from the Wisconsin Alumni Research foundation. We thank Zheng Tracy Ke, Rungang Han, Yuetian Luo for helpful discussions and for sharing software packages.

REFERENCES

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

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, 48(3):1452.

Ahn, K., Lee, K., and Suh, C. (2018). Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974.

Ahn, K., Lee, K., and Suh, C. (2019). Community recovery in hypergraphs. *IEEE Transactions on Information Theory*, 65(10):6561–6579.

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

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.

Brennan, M. and Bresler, G. (2020). Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR.

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

Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.

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

Desikan, R. S., Sgonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.

Ghoshdastidar, D. and Dukkipati, A. (2017). Uniform hypergraph partitioning: Provable tensor methods and sampling techniques. *Journal of Machine Learning Research*, 18(1):1638–1678.

Ghoshdastidar, D. et al. (2017). Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315.

Han, R., Luo, Y., Wang, M., and Zhang, A. R. (2020). Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*.

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

Hu, J., Lee, C., and Wang, M. (2021). Generalized tensor decomposition with features on multiple modes. *Journal of Computational and Graphical Statistics*, pages 1–15.

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

Kim, C., Bandeira, A. S., and Goemans, M. X. (2018). Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*.

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

Koniusz, P. and Cherian, A. (2016). Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5395–5403.

Lee, S. H., Magallanes, J. M., and Porter, M. A. (2017). Time-dependent community structure in legislation cosponsorship networks in the congress of the republic of peru. *Journal of Complex Networks*, 5(1):127–144.

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

Meilă, M. (2012). Local equivalences of distances between clusteringsa geometric perspective. *Machine Learning*, 86(3):369–389.

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

Pananjady, A. and Samworth, R. J. (2020). Isotonic regression with unknown permutations: Statistics, computation, and adaptation. *arXiv preprint arXiv:2009.02609*.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., and WU-Minn HCP Consortium (2013). The WU-Minn human connectome project: An overview. *Neuroimage*, 80:62–79.

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

Wang, M., Fischer, J., and Song, Y. S. (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. and 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. and Zeng, Y. (2019). Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, volume 32, pages 713–723.

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

Yuan, M., Liu, R., Feng, Y., and Shang, Z. (In Press). Testing community structures for hypergraphs. *The Annals of Statistics, arXiv preprint arXiv:1810.04617*.

Yun, S.-Y. and Proutiere, A. (2016). Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, volume 29, pages 965–973.

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

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

APPENDIX

We provide the proofs for all the theorems in our main paper. In each sub-section, we first show the proof of main theorem and ~~attach~~ then collect the useful lemmas in the end.

NOTATION

Before the proofs, we first introduce the ~~notations~~ notation used throughout the ~~following sections~~ appendix and the generalized dTBM without symmetric assumptions~~for the proofs of Theorem 1 and Theorem 2~~. The parameter space and minimal gap assumption ~~for~~ are also extended ~~the generalized dTBM. The conclusions of Theorem 1 and 2 in the main paper are obtained by simply setting $z_k = z, \boldsymbol{S}_k = \boldsymbol{S}, \boldsymbol{\theta}_k = \boldsymbol{\theta}, k \in [K]$~~ for the generalized dTBM.

~~Notations~~Preliminaries.

1) For ~~all~~ mode $k \in [K]$, denote the ~~tensor matricizations as~~ mode-k tensor matricizations by

$$\boldsymbol{Y}_k = \mathrm{Mat}_k(\mathcal{Y}), \quad \boldsymbol{S}_k = \mathrm{Mat}_k(\mathcal{S}), \quad \boldsymbol{E}_k = \mathrm{Mat}_k(\mathcal{E}), \quad \boldsymbol{X}_k = \mathrm{Mat}_k(\mathcal{X}).$$

2) For a vector $\boldsymbol{a}$, let $\boldsymbol{a}^s := \boldsymbol{a}/\|\boldsymbol{a}\|$ denote the normalized vector. We make the convention that ~~$\boldsymbol{a}^s = 0$ if $\boldsymbol{a} = 0$.~~ $\boldsymbol{a}^s = \boldsymbol{0}$ if $\boldsymbol{a} = \boldsymbol{0}$.

3) For a matrix ~~$\boldsymbol{A} \in \mathbb{R}^{n\times m} \in \mathbb{R}^{n^K \times m^K}$, let $\boldsymbol{A}^{\otimes K}$ denotes the kronecker~~ $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, let $\boldsymbol{A}^{\otimes K} := \boldsymbol{A} \otimes \cdots \otimes \boldsymbol{A} \in \mathbb{R}^{n^K \times m^K}$ denote the Kronecker product of K ~~matrices $\boldsymbol{A} \otimes \cdots \otimes \boldsymbol{A}$,~~ copies of matrices $\boldsymbol{A}$.

4) For a matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\|_\sigma$ denote the spectral norm of matrix $\boldsymbol{A}$, which is equal to the maximal singular value of $\boldsymbol{A}$; let $\lambda_k(\boldsymbol{A})$ denote the k-th largest singular value of $\boldsymbol{A}$; let $\|\boldsymbol{A}\|_F$ denote the Frobenius norm of matrix $\boldsymbol{A}$.

5) For two ~~terms~~ sequence a and b, let $a \asymp b$ if there exist two positive constants c, C such that $cb \le a \le Cb$.

~~Generalized~~ Model extension to generalized dTBM.

The general order-K $(p_1, \ldots, p_K)$-dimensional dTBM model with r_k communities and degree heterogeneity $\boldsymbol{\theta}_k = [\![\theta_k(i)]\!] \in \mathbb{R}_+^{p_k}$ is represented by

$$\mathcal{Y} = \mathcal{X} + \mathcal{E}, \quad \text{where} \quad \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{\Theta}_K \boldsymbol{M}_K, \tag{20}$$

where $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the data tensor, $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the mean tensor, $\mathcal{S} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is the core tensor, $\mathcal{E} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the noise tensor consisting of independent ~~mean-zero~~ zero-mean sub-Gaussian entries with variance bounded by σ^2, $\boldsymbol{\Theta}_k = \mathrm{diag}(\boldsymbol{\theta}_k)$, and $\boldsymbol{M}_k \in \{0,1\}^{p_k \times r_k}$ is the membership matrix corresponding to the assignment $z_k : [p_k] \mapsto [r_k]$, for all $k \in [K]$.

For ease of notation, we use $\{z_k\}$ to denote the collection $\{z_k\}_{k=1}^K$, and $\{\boldsymbol{\theta}_k\}$ to denote the collection $\{\boldsymbol{\theta}_k\}_{k=1}^K$. Correspondingly, we consider the parameter space for the triplet $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$,

$$\mathcal{P}(\{r_k\}) = \Big\{ (\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) : \\ \boldsymbol{\theta}_k \in \mathbb{R}_+^p, \frac{c_1 p_k}{r_k} |z_k^{-1}(a)| \le \frac{c_2 p_k}{r_k}, c_3 \le \|\boldsymbol{S}_{k,a:}\| \le c_4, \|\boldsymbol{\theta}_{k, z_k^{-1}(a)}\|_1 = |z_k^{-1}(a)|, a \in [r_k], k \in [K] \Big\}.$$

We call the ~~collection of~~ degree heterogeneity $\{\boldsymbol{\theta}_k\}$ is balanced if for all $k \in [K]$,

$$\min_{a\in[r]} \|\boldsymbol{\theta}_{k, z_k^{-1}(a)}\| = (1 + o(1)) \max_{a\in[r]} \|\boldsymbol{\theta}_{k, z_k^{-1}(a)}\|.$$

We also consider the generalized Assumption 1 on angle gap.

Assumption 2 (Generalized angle gap)**.** Recall $\boldsymbol{S}_k = \mathrm{Mat}_k(\mathcal{S})$. We assume the minimal gap between normalized rows of $\boldsymbol{S}_k$ is bounded away from zero ~~, for all $k \in [K]$;~~ i.e.,

$$\Delta_{\min} := \min_{k\in[K]} \min_{a\neq b\in[r_k]} \left\|\boldsymbol{S}^s_{k,a:} - \boldsymbol{S}^s_{k,b:}\right\| > 0.$$

~~for all $k \in [K]$.~~

Similarly, let $\mathrm{SNR} = \Delta^2_{\min}/\sigma^2$ with the generalized minimal gap $\Delta^2_{\min}$ defined in Assumption 2. We define the regime

$$\mathcal{P}(\gamma) = \mathcal{P}(\{r_k\}) \cap \{\mathcal{S} \text{ satisfies } \mathrm{SNR} = p^\gamma \text{ and } p_k \asymp p, k \in [K] \text{ satisfies } \mathrm{SNR} = p^\gamma \text{ and } p_k \asymp p, \text{for all } k \in [K]\}.$$

Proof of Theorem 1

Proof of Theorem 1. To study the identifiability, we consider the noiseless model with $\mathcal{E} = 0$. Assume there ~~exists~~ exist two parameterizations satisfying

$$\mathcal{X} = \mathcal{S} \times_1 \Theta_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \Theta_K \boldsymbol{M}'_K = \mathcal{S}' \times_1 \Theta'_1 \boldsymbol{M}'_1 \times_2 \cdots \times_K \Theta'_K \boldsymbol{M}'_K,$$

~~with~~ where $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\{r_k\})$ and $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\}) \in \mathcal{P}(\{r'_k\})$ are two sets of parameters. We prove the sufficient and necessary conditions separately.

($\Leftarrow$) For the necessity, it ~~is equivalent to show that there exists a triplet $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\})$ is not identical to $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation~~ suffices to construct two distinct parameters up to cluster label permutation, if the model (20) violates Assumption 2. Without loss of generality, we assume $\left\|\boldsymbol{S}^s_{1,1:} - \boldsymbol{S}^s_{1,2:}\right\| = 0$.

If $\boldsymbol{S}_{1,1:}$ is a zero vector, ~~consider~~ construct $\boldsymbol{\theta}'_1$ such that ~~$\boldsymbol{\theta}'_{z_1^{-1}(1)} \neq \boldsymbol{\theta}_{z_1^{-1}(1)}$~~ $\boldsymbol{\theta}'_{1,z_1^{-1}(1)} \neq \boldsymbol{\theta}_{1,z_1^{-1}(1)}$. Let $\{z'_k\} = \{z_k\}$, $\mathcal{S}' = \mathcal{S}$, and $\boldsymbol{\theta}'_k = \boldsymbol{\theta}_k$ for all $k = 2, \ldots, K$. Then the triplet $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\})$ is ~~not identical to $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation. We have a similar story when $\boldsymbol{S}_{1,2:}$ is a zero vector.~~

~~If neither $\boldsymbol{S}_{1,1:}$ nor $\boldsymbol{S}_{1,2:}$ is a zero vector, there exists a positive constant c such that $\boldsymbol{S}_{1,1:} = c\boldsymbol{S}_{1,2:}$. Thus,there exists a core tensor $\mathcal{S}_0 \in \mathbb{R}^{r_1-1\times\cdots\times r_K}$ such that~~

$$\mathcal{S} = \mathcal{S}_0 \times_1 \boldsymbol{C}\boldsymbol{R}, \quad \text{where} \quad \boldsymbol{C} = \mathrm{diag}(1, c, 1, ..., 1) \in \mathbb{R}^{r_1\times r_1}, \quad \boldsymbol{R} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & \mathbf{1}_{r_1-2} \end{pmatrix} \in \mathbb{R}^{r_1\times(r_1-1)}.$$

~~Let $\boldsymbol{D} = \mathrm{diag}(1 + c, 1, ..., 1) \in \mathbb{R}^{r_1-1\times r_1-1}$.Consider the parameterization~~

$$\boldsymbol{M}'_1 = \boldsymbol{M}_1\boldsymbol{R}, \quad \mathcal{S}' = \mathcal{S}_0 \times_1 \boldsymbol{D}, \quad \theta'_1(i) = \begin{cases} \frac{1}{1+c}\theta_1(i) & i \in z_1^{-1}(1) \\ \frac{c}{1+c}\theta_1(i) & i \in z_1^{-1}(2) \\ \theta_1(i) & \text{otherwise} \end{cases},$$

~~and $\boldsymbol{M}'_k = \boldsymbol{M}_k, \boldsymbol{\theta}'_k = \boldsymbol{\theta}_k$ for all $k = 2, \ldots, K$.Then we have constructed a triplet $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\})$ that is not identical to $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation.~~

~~For the sufficiency,it is equivalent to show that all possible triplets $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\})$ are identical to $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation if the model satisfies Assumption . We show the uniqueness of the parameters separately.~~

~~First, we show the uniqueness of $\boldsymbol{M}_k$ for all $k \in [K]$. Specifically, we show the uniqueness of the first mode membership matrix, i.e., $\boldsymbol{M}'_1 = \boldsymbol{M}_1\boldsymbol{P}_1$ where $\boldsymbol{P}_1$ is a permutation matrix. The uniqueness of $\boldsymbol{M}_k$ for $k = 2, \ldots, K$ can be showed in the same way, and we omit the repeated procedures.~~

~~On one hand, consider the pair of nodes (i, j) such that $z_1(i) = z_1(j)$. We have $\|\boldsymbol{X}^s_{1,z_1(i):} - \boldsymbol{X}^s_{1,z_1(j):}\| = 0$ and thus $\|(\boldsymbol{S}')^s_{1,z'_1(i):} - (\boldsymbol{S}')^s_{1,z'_1(j):}\| = 0$ by Lemma **??**. Then, by Assumption , we have $z'_1(i) = z'_1(j)$. On the~~

other hand, consider the pair of nodes (i,j) such that $z_1(i) \neq z_1(j)$. We have $\left\|\boldsymbol{X}^s_{1,i:} - \boldsymbol{X}^s_{1,j:}\right\| \neq 0$ and thus $\left\|(\boldsymbol{S}')^s_{1,z_1'(i):} - (\boldsymbol{S}')^s_{1,z_1'(j):}\right\| \neq 0$ by Lemma **??**. Hence, we have $z_1'(i) \neq z_1'(j)$. Therefore, we have proven that z_1' is equal to z_i up to label permutation.

Next, we show the uniqueness of $\boldsymbol{\theta}_k$ for all $k \in [K]$ given that $z_k = z_k'$. Similarly, we show the detailed proof of the uniqueness of $\boldsymbol{\theta}_1$, i.e.,$\boldsymbol{\theta}_1' = \boldsymbol{\theta}_1$, and omit the repeated procedures for $\boldsymbol{\theta}_k, k = 2, \ldots, K$distinct from $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation. Similar conclusion holds when $\boldsymbol{S}_{1,2:}$ is a zero vector. If neither $\boldsymbol{S}_{1,1:}$ nor $\boldsymbol{S}_{1,2:}$ is a zero vector, there exists a positive constant c such that $\boldsymbol{S}_{1,1:} = c\boldsymbol{S}_{1,2:}$. Thus, there exists a core tensor $\mathcal{S}_0 \in \mathbb{R}^{r_1 - 1 \times \cdots \times r_K}$ such that Let $\boldsymbol{D} = \operatorname{diag}(1 + c, 1,$.

Consider an arbitrary $j \in [p_1]$ such that $z_1(j) = a$. Then for all the node $i \in z_1^{-1}(a)$ in the same cluster of j, we have

$$\frac{\boldsymbol{X}_{1,z_1(i):}}{\boldsymbol{X}_{1,z_1(j):}} = \frac{\boldsymbol{X}'_{1,z_1(i):}}{\boldsymbol{X}'_{1,z_1(j):}}, \text{ which implies } \frac{\theta_1(j)}{\theta_1(i)} = \frac{\theta_1'(j)}{\theta_1'(i)}.$$

Let $\theta_1'(j) = c\theta_1(j)$ for some constant c. By the equation , we have $\theta_1'(i) = c\theta_1(i)$ for all $i \in z_1^{-1}(a)$. Note that $(\{z_k\}, \mathcal{S}', \{\boldsymbol{\theta}_k'\}) \in \mathcal{P}(\{r_k\})$. We have

$$\sum_{j \in z^{-1}(a)} \theta_1'(j) = c \sum_{j \in z^{-1}(a)} \theta_1(j) = 1,$$

which implies $c = 1$. Hence, we have proven $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1'$ given that $z_1 = z_1'$.

Last, we show the uniqueness of $\mathcal{S}$, i.e., $\mathcal{S}' = \mathcal{S} \times_1 \boldsymbol{P}_1^{-1} \times_2 \cdots \times_K \boldsymbol{P}_K^{-1}$, where $\boldsymbol{P}_k, k \in [K]$ are permutation matrices. Given $z_k' = z_k, \boldsymbol{\theta}_k' = \boldsymbol{\theta}_k$, we have $\boldsymbol{M}_k' = \boldsymbol{M}_k \boldsymbol{P}_k$ and $\boldsymbol{\Theta}_k' = \boldsymbol{\Theta}_k$ for all $k \in [K]$.

Let $\boldsymbol{D}_k = \left[(\boldsymbol{\Theta}_k' \boldsymbol{M}_k')^T (\boldsymbol{\Theta}_k' \boldsymbol{M}_k')\right]^{-1} (\boldsymbol{\Theta}_k' \boldsymbol{M}_k')^T, k \in [K]$. By the parameterization, we have

$$\begin{aligned}
\mathcal{S}' &= \mathcal{X} \times_1 \boldsymbol{D}_1 \times_2 \cdots \times_K \boldsymbol{D}_K \\
&= \mathcal{S} \times_1 \boldsymbol{D}_1 \boldsymbol{\Theta}_1 \boldsymbol{M}_1 \times_1 \cdots \times_K \boldsymbol{D}_K \boldsymbol{\Theta}_K \boldsymbol{M}_K \\
&= \mathcal{S} \times_1 \boldsymbol{P}_1^{-1} \times_2 \cdots \times_K \boldsymbol{P}_K^{-1}.
\end{aligned}$$

Therefore, we finish the proof of Theorem 1.

Useful Lemma for the Proof of Theorem 1

Consider the signal tensor $\mathcal{X}$ in the generalized dTBM with $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\{r_k\})$ and $r_k \geq 2$. Then, for any $k \in [K]$ and index pair $(i,j) \in [p_k]^2$, we have

$$\left\|\boldsymbol{S}^s_{k,z_k(i):} - \boldsymbol{S}^s_{k,z_k(j):}\right\| = 0 \quad \text{if and only if} \quad \left\|\boldsymbol{X}^s_{k,z_k(i):} - \boldsymbol{X}^s_{k,z_k(j):}\right\| = 0.$$

For simplicity, we show the detailed proof for $k = 1$ and drop the subscript k in $\boldsymbol{X}_k, \boldsymbol{S}_k$.The repeated proofs for $k = 2, \ldots, K$ are omitted.

By tensor matricization, we have

$$\boldsymbol{X}_{j:} = \theta_1(j) \boldsymbol{S}_{z_1(j):} \left[\boldsymbol{\Theta}_2 \boldsymbol{M}_2 \otimes \cdots \otimes \boldsymbol{\Theta}_K \boldsymbol{M}_K\right]^T.$$

Let $\tilde{\boldsymbol{M}} = \boldsymbol{\Theta}_2 \boldsymbol{M}_2 \otimes \cdots \otimes \boldsymbol{\Theta}_K \boldsymbol{M}_K$. Notice that for two vectors $\boldsymbol{a}, \boldsymbol{b}$ and two positive constants $c_1, c_2 > 0$, we have

$$\|\boldsymbol{a}^s - \boldsymbol{b}^s\| = \|(c_1 \boldsymbol{a})^s - (c_2 \boldsymbol{b})^s\|.$$

~~Thus it is sufficient to show the following statement that for any index pair $(i, j) \in [p_1]^2$~~,

$$\left\| \boldsymbol{S}^s_{z_1(i):} - \boldsymbol{S}^s_{z_1(j):} \right\| = 0 \quad \text{if and only if} \quad \left\| \left[\boldsymbol{S}_{z_1(i):} \tilde{\boldsymbol{M}}^T \right]^s - \left[\boldsymbol{S}_{z_1(j):} \tilde{\boldsymbol{M}}^T \right]^s \right\| = 0.$$