Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Generalized Tensor Decomposition With Features on Multiple Modes

Jiaxin Hu, Chanwoo Lee & Miaoyan Wang

To cite this article: Jiaxin Hu, Chanwoo Lee & Miaoyan Wang (2022) Generalized Tensor Decomposition With Features on Multiple Modes, Journal of Computational and Graphical Statistics, 31:1, 204-218, DOI: 10.1080/10618600.2021.1978471

To link to this article: https://doi.org/10.1080/10618600.2021.1978471

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 31, NO. 1, 204–218
https://doi.org/10.1080/10618600.2021.1978471

Generalized Tensor Decomposition With Features on Multiple Modes

Jiaxin Hu, Chanwoo Lee, and Miaoyan Wang

Department of Statistics, University of Wisconsin-Madison, Madison, WI

ABSTRACT

Higher-order tensors have received increased attention across science and engineering. While most tensor decomposition methods are developed for a single tensor observation, scientific studies often collect side information, in the form of node features and interactions thereof, together with the tensor data. Such data problems are common in neuroimaging, network analysis, and spatial-temporal modeling. Identifying the relationship between a high-dimensional tensor and side information is important yet challenging. Here, we develop a tensor decomposition method that incorporates multiple feature matrices as side information. Unlike unsupervised tensor decomposition, our supervised decomposition captures the effective dimension reduction of the data tensor confined to feature space of interest. An efficient alternating optimization algorithm with provable spectral initialization is further developed. Our proposal handles a broad range of data types, including continuous, count, and binary observations. We apply the method to diffusion tensor imaging data from human connectome project and multi-relational political network data. We identify the key global connectivity pattern and pinpoint the local regions that are associated with available features. The package and data used are available at *https://CRAN.R-project.org/package=tensorregress*. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2020
Revised June 2021

KEYWORDS
Alternating optimization; Exponential family distribution; Generalized multilinear model; Supervised dimension reduction; Tensor data analysis

1. Introduction

Multi-dimensional arrays, known as tensors, are often collected with side information on multiple modes in modern scientific and engineering studies. A popular example is in neuroimaging (Zhou, Li, and Zhu 2013). The brain connectivity networks are collected from a sample of individuals, accompanied by individual characteristics such as age, gender, and diseases status (see Figure 1(a)). Another example is in network analysis (Hoff 2005; Berthet and Baldin 2020). A typical social network consists of nodes that represent people and edges that represent friendships. Side information such as people's demographic information and friendship types are often available. In both examples, we are interested in identifying the variation in the data tensor (e.g., brain connectivities, social community patterns) that is affected by available features. These seemingly different scenarios pose a common yet challenging problem for tensor data modeling.

In addition to the aforementioned challenges, many tensor datasets consist of non-Gaussian measurements. Examples include the political interaction dataset (Nickel, Tresp, and Kriegel 2011) which measures action counts between countries under various relations, and the brain connectivity network dataset (Zhou, Li, and Zhu 2013; Wang and Li 2020) which is a collection of binary adjacency matrices. Classical tensor decomposition methods are based on minimizing the Frobenius norm of deviation, leading to suboptimal predictions for binary- or count-valued response variables. A number of supervised tensor methods have been proposed (Li and Zhang 2017; Sun and Li 2017; Lock and Li 2018; Raskutti, Yuan, and Chen 2019; Hao et al. 2021) to address the tensor regression problem in various forms, such as scalar-to-tensor regression and tensor-response regression. These methods often assume Gaussian distribution for the tensor entries, or impose random designs for the feature matrices, both of which are less suitable for applications of our interest. The gap between theory and practice means a great opportunity to model paradigms and better capture the complexity in tensor data.

We present a general model and associated method for decomposing a data tensor whose entries are from exponential family with side information. We formulate the learning task as a structured regression problem, with tensor observation serving as the response, and the multiple side information as features. Figure 1(b) illustrates our model in the special case of order-3 tensors. A low-rank structure is imposed to the conditional mean of tensor observation, where unlike classical decomposition, the tensor factors $\boldsymbol{X}_k\boldsymbol{M}_k \in \mathbb{R}^{d_k \times r_k}$ belong to the space spanned by features $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ for $k = 1, 2, 3$. The unknown matrices $\boldsymbol{M}_k \in \mathbb{R}^{p_k \times r_k}$ (referred to as "dimension reduction matrices") link the conditional mean to the feature spaces, thereby allowing the identification of variations in the tensor data attributable to the side information.

Our proposal blends the modeling power of generalized linear model (GLM) and the exploratory capability of tensor dimension reduction in order to take the best out of both worlds. We leverage GLM to allow heteroscedacity due to the mean-variance relationship in the non-Gaussian data. This flexibility is important in practice. Furthermore, our low-rank model on the (transformed) conditional mean tensor effectively mitigates

CONTACT Miaoyan Wang *miaoyan.wang@wisc.edu* Department of Statistics, University of Wisconsin-Madison, Madison, WI.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

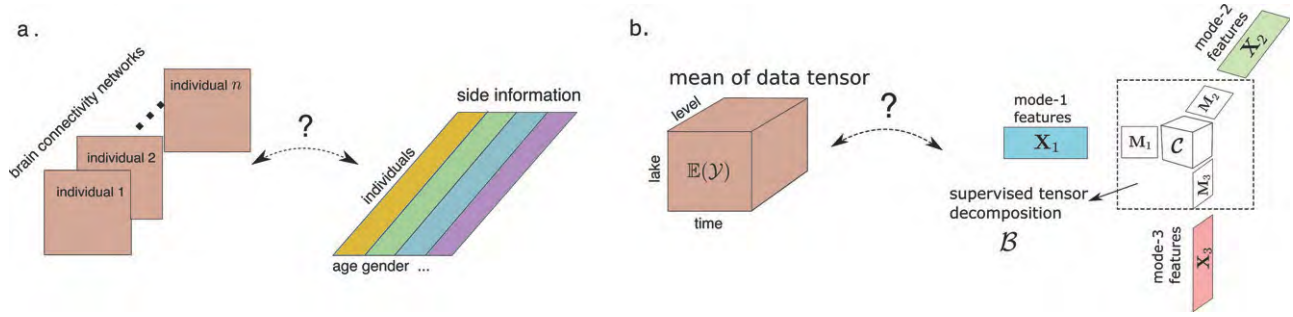

Figure 1. Examples of supervised tensor decomposition with side information. (a) Network population model. (b) Spatio-temporal growth model.

the curse of high dimensionality. In classical GLM, the sample size and feature dimension are well defined; however, in the tensor data analysis, we observe only one realization of an order-K tensor and up to K feature matrices. Both the number of tensor entries and feature dimension grow exponentially in K. Dimension reduction is therefore crucial for prediction and interpretability. We establish the statistical and algorithmic convergences of our estimator, and we quantify the gain in accuracy through simulations and case studies.

Our work is closely related to but also clearly distinctive from several lines of previous work. The first line is a class of *unsupervised* tensor decomposition such as classical Tucker and CP decomposition (De Lathauwer, De Moor, and Vandewalle 2000; Kolda and Bader 2009) and generalized decomposition for non-Gaussian data (Chi and Kolda 2012; Tarzanagh and Michailidis 2019; Hong, Kolda, and Duersch 2020; Li 2020). Regardless of the implementation, the unsupervised methods aim to find the best low-rank representation of a data tensor alone. In contrast, our model is a *supervised* tensor learning, which aims to identify the association between a data tensor and multiple features. The low-rank factorization is determined jointly by the tensor data and feature matrices in our model.

The second line of work studies the tensor-to-tensor regression. This category is further divided into three scenarios, depending on whether tensor is treated as predictors (Zhou, Li, and Zhu 2013; Raskutti, Yuan, and Chen 2019; Han, Willett, and Zhang 2020), as responses (Li and Zhang 2017; Sun and Li 2017; Zhang, Sun, and Li 2018; Lock and Li 2018; Luo et al. 2018), or both (Lock 2018; Gahrooei et al. 2020). As we show in Section 5, our supervised tensor decomposition falls into this general category, and we provide a *provable* solution in new settings that have broader practical significance. Earlier work in this vein (Lock 2018; Lock and Li 2018; Gahrooei et al. 2020; Li 2020) focuses on algorithm development, but not on the statistical accuracy. Li and Zhang (2017) introduces an envelope-based approach to identify sufficient dimension reduction (Adragni and Cook 2009), but its theory is restricted to Gaussian data with one-sided feature matrix only. Raskutti, Yuan, and Chen (2019) establishes the statistical accuracy for convex relaxed maximum likelihood estimator (MLE) of tensor regression. However, convex relaxation for tensor optimizations suffers from computational intractability and statistical sub-optimality. Recent work has demonstrated the success of non-convex approaches in various tensor problems (Sun and Li 2017; Zhang, Sun, and Li 2018; Raskutti, Yuan, and Chen 2019; Han, Willett, and Zhang 2020); we go step further by allowing multiple feature matrices with either fixed or random designs. In Sections 4.2, we show that incorporating multiple feature matrices substantially improves the statistical accuracy. We provide a detailed comparison in Section 5; see Table 2.

The third line of work uses side information for various tensor learning tasks, such as for completion (Song et al. 2019) and for recommendation system (Farias and Li 2019). These methods also study tensors with side information, but they take data-mining approaches to penalize predictions that are distant from side information. One important difference is that their goal is prediction but not parameter estimation. The effects of features and their interactions are not estimated in these data-driven approaches. In contrast, our goal is interpretable prediction, and we estimate the low-rank decomposition using a model-based approach. The model-based approach benefits the interpretability in prediction. In this regards, our method opens up new opportunities for tensor data analysis in a wider range of applications.

The remainder of the article is organized as follows. Section 2 introduces tensor preliminaries. Section 3 presents the main model and three motivating examples for supervised tensor decomposition. We describe the likelihood estimation and alternating optimization algorithm with theoretical guarantees in Section 4. Connection with related work is provided in Section 5. In Section 6, we present numerical experiments and assess the performance in comparison to alternative methods. In Section 7, we apply the method to diffusion tensor imaging data from human connectome project and multi-relational social network data. We conclude in Section 8 with discussions about our findings and avenues of future work. All proofs are deferred to the supplementary notes.

2. Preliminaries

We introduce the basic tensor properties used in the paper. We use lower-case letters (e.g., $a, b,$ and c) for scalars and vectors, upper-case boldface letters (e.g., $\boldsymbol{A}, \boldsymbol{B},$ and $\boldsymbol{C}$) for matrices, and calligraphy letters (e.g., $\mathcal{A}, \mathcal{B},$ and $\mathcal{C}$) for tensors of order three or greater. We use $\boldsymbol{I}$ to denote the identity matrix whose dimension may vary from line by line given the contexts. Let $\mathcal{Y} = [y_{i_1,\ldots,i_K}] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K $(d_1, \ldots, d_K)$-dimensional tensor, where K is the number of modes and also called the order. The multilinear multiplication of a tensor $\mathcal{Y} \in$

$\mathbb{R}^{d_1\times\cdots\times d_K}$ by matrices $\boldsymbol{X}_k=[x^{(k)}_{i_k,j_k}]\in\mathbb{R}^{p_k\times d_k}$ is defined as

$$\mathcal{Y}\times_1\boldsymbol{X}_1\times\cdots\times_K\boldsymbol{X}_K=[\sum_{i_1,\ldots,i_K}y_{i_1,\ldots,i_K}x^{(1)}_{j_1,i_1}\cdots x^{(K)}_{j_K,i_K}],$$

which results in an order-K $(p_1,\ldots,p_K)$-dimensional tensor. For ease of presentation, we use the shorthand $\mathcal{Y}\times\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K\}$ to denote the tensor-by-matrix product. For any two tensors $\mathcal{Y}=[y_{i_1,\ldots,i_K}],\mathcal{Y}'=[y'_{i_1,\ldots,i_K}]$ of identical order and dimensions, their inner product is defined as

$$\langle\mathcal{Y},\,\mathcal{Y}'\rangle=\sum_{i_1,\ldots,i_K}y_{i_1,\ldots,i_K}y'_{i_1,\ldots,i_K}.$$

The tensor Frobenius norm and maximum norm are defined as

$$||\mathcal{Y}||_F=\langle\mathcal{Y},\,\mathcal{Y}\rangle^{1/2},\quad\text{and}\quad||\mathcal{Y}||_\infty=\max_{i_1,\ldots,i_K}y_{i_1,\ldots,i_K}.$$

When a is a vector, we use $||a||_2=\langle a,a\rangle^{1/2}$ to denote the vector 2-norm. We use $[d]$ to denote the d-set $[d]=\{1,\ldots,d\}$, and use $\mathbb{O}(d,r)$ to denote the collection of all d-by-r matrices with orthonormal columns; that is, $\mathbb{O}(d,r)=\{\boldsymbol{P}\in\mathbb{R}^{d\times r}\colon\boldsymbol{P}^T\boldsymbol{P}=\boldsymbol{I}\}$.

A higher-order tensor can be reshaped into a lower-order object. We use $\mathrm{vec}(\cdot)$ to denote the operation that reshapes the tensor into a vector, and $\mathrm{Unfold}_k(\cdot)$ to denote the unfolding operation that reshapes the tensor along mode k into a matrix of size d_k-by-$\prod_{i\neq k}d_i$. We use $\mathrm{rank}(\mathcal{Y})=\boldsymbol{r}$ to denote the multilinear rank of an order-K tensor $\mathcal{Y}$, where $\boldsymbol{r}=(r_1,\ldots,r_K)$ is a length-K vector and r_k is the rank of matrix $\mathrm{Unfold}_k(\mathcal{Y})$ for $k\in[K]$. For ease of notation, we allow the basic arithmetic operators (e.g., $+,-,\geq$) and univariate functions $f\colon\mathbb{R}\to\mathbb{R}$ to be applied to tensors in an element-wise manner. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n\lesssim b_n$ or $a_n=\mathcal{O}(b_n)$ to denote the fact that $a_n\leq Cb_n$ for some constant $C>0$.

3. Motivation and Model

3.1. General Framework for Tensor Decomposition

We begin with a general framework for supervised tensor decomposition and then discuss its implication in three concrete examples. Let $\mathcal{Y}=[y_{i_1,\ldots,i_K}]\in\mathbb{R}^{d_1\times\cdots\times d_K}$ denote an order-K data tensor. Suppose the side information is available on each of the K modes. Let $\boldsymbol{X}_k=[x_{ij}]\in\mathbb{R}^{d_k\times p_k}$ denote the feature matrix on the mode $k\in[K]$, where x_{ij} denotes the j-th feature value for the i-th tensor entity, for $(i,j)\in[d_k]\times[p_k]$.

We propose a multilinear conditional mean model between the data tensor and feature matrices. Assume that, conditional on the features $\boldsymbol{X}_k$, the entries of tensor $\mathcal{Y}$ are independent realizations from an exponential family distribution. Further, the conditional mean tensor admits the rank-$\boldsymbol{r}$ model with $\boldsymbol{r}=(r_1,\ldots,r_K)$,

$$\begin{aligned}\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K)&=f\left(\mathcal{C}\times\{\boldsymbol{X}_1\boldsymbol{M}_1,\ldots,\boldsymbol{X}_K\boldsymbol{M}_K\}\right),\\&\text{with }\boldsymbol{M}_k^T\boldsymbol{M}_k=\boldsymbol{I}_{r_k},\ \boldsymbol{M}_k\in\mathbb{R}^{p_k\times r_k}\\&\text{for all }k=1,\ldots,K,\end{aligned}\tag{1}$$

where $\mathcal{C}\in\mathbb{R}^{r_1\times\cdots\times r_K}$ is an unknown full-rank core tensor, $\boldsymbol{M}_k\in\mathbb{R}^{p_k\times r_k}$ are unknown factor matrices for all $k\in[K]$, $f(\cdot)$ is a known link function whose form depending on the data type of $\mathcal{Y}$, and $\times$ denotes the tensor-by-matrix product. The choice of link function is based on the assumed distribution family of tensor entries. Common choices of link functions include identity link for Gaussian distribution, logistic link for Bernoulli distribution, and exponential link for Poisson distribution. In general, dispersion parameters can also be included in the model. Because our main focus is the tensor decomposition under the mean model, we suppress the dispersion parameter in this section for ease of presentation.

Figure 1(b) provides a schematic illustration of our model. The features $\boldsymbol{X}_k$ affect the distribution of tensor entries in $\mathcal{Y}$ through the reduced features $\boldsymbol{X}_k\boldsymbol{M}_k$, which are r_k linear combinations of features on mode k. We call $\boldsymbol{M}_k$ the "dimension reduction matrix" or "tensor factors." The core tensor $\mathcal{C}$ collects the interaction effects between reduced features across K modes. We call $\mathcal{B}=\mathcal{C}\times\{\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K\}$ the coefficient tensor, and $\Theta=\mathcal{B}\times\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K\}$ the linear predictor. By the definition of multilinear rank, the model (1) implies the linear predictor Θ and coefficient tensor $\mathcal{B}$ are of rank-$\boldsymbol{r}$. The conditional mean tensor $\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K)$ is however often high rank, due to the nonlinearity of the link function (Lee and Wang 2021).

Our goal is to estimate the low-rank tensor $\mathcal{B}$, or equivalently, the core tensor and factors $(\mathcal{C},\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K)$, from our model (1). We make several remarks about model identifiability. First, the identifiability of $\mathcal{B}$ requires the feature matrices $\boldsymbol{X}_k$ are of full column rank with $p_k\leq d_k$. We impose this rank nondeficiency assumption to $\boldsymbol{X}_k$; this is a mild condition common in literature (Li and Zhang 2017; Lock and Li 2018; Li 2020). In the presence of rank deficiency, we recommend to remove redundant features from $\boldsymbol{X}_k$ before applying our method. Second, the decomposition $\mathcal{B}=\mathcal{C}\times\{\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K\}$ are non-unique, as in standard tensor decomposition (Kolda and Bader 2009). For any invertible matrices $\boldsymbol{O}_k\in\mathbb{R}^{r_k\times r_k}$, $\mathcal{B}=\mathcal{C}\times\{\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K\}=\mathcal{C}'\times\{\boldsymbol{M}_1\boldsymbol{O}_1,\ldots,\boldsymbol{M}_K\boldsymbol{O}_K\}$ are two equivalent parameterizations with $\mathcal{C}'=\mathcal{C}\times\{\boldsymbol{O}_1^{-1},\ldots,\boldsymbol{O}_K^{-1}\}$. To resolve this ambiguity, we impose orthonormality to $\boldsymbol{M}_k\in\mathbb{O}(p_k,r_k)$ and assess the estimation error of $\boldsymbol{M}_k$ using angle distance. The angle distance is invariant to orthogonal rotations due to its geometric definition. See Section 4.2 for more details. The orthonormality of $\boldsymbol{M}_k$ is imposed purely for technical convenience. This normalization incurs no impacts in our statistical inference, but may help with numerical stability in empirical optimization (De Lathauwer, De Moor, and Vandewalle 2000; Kolda and Bader 2009). Finally, the problem size is quantified by p_k and d_k, where p_k specifies the number of features and d_k the number of samples at mode $k\in[K]$. Our theory treats the rank r_k as known and fixed, whereas both p_k and d_k are allowed to increase. The adaptation to unknown rank in practice will be addressed in Section 4.3.

3.2. Three Examples

We give three seemingly different examples that can all be formulated as our supervised tensor decomposition model (1).

Example 1 (Spatio-temporal growth model). The growth curve model (Gabriel 1998; Srivastava, von Rosen, and Von Rosen 2008) was originally proposed as an example of bilinear model for matrix data, and we adopt its higher-order extension here.

Let $\mathcal{Y} = [y_{ijk}] \in \mathbb{R}^{d\times m\times n}$ denote the pH measurements of d lakes at m levels of depth and for n time points. Suppose the sampled lakes belong to q types, with p lakes in each type. Let $\{\ell_j\}_{j\in[m]}$ denote the sampled depth levels and $\{t_k\}_{k\in[n]}$ the time points. Assume that the expected pH trend in depth is a polynomial of order at most r and that the expected trend in time is a polynomial of order s. Then, the conditional mean model for the spatio-temporal growth can be represented as follows:

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) = \mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1,\ \boldsymbol{X}_2\boldsymbol{M}_2,\ \boldsymbol{X}_3\boldsymbol{M}_3\}, \tag{2}$$

where $\boldsymbol{X}_1 = \text{blockdiag}\{\mathbf{1}_p, \ldots, \mathbf{1}_p\} \in \{0,1\}^{d\times q}$ is the design matrix for lake types, and

$$\boldsymbol{X}_2 = \begin{pmatrix} 1 & \ell_1 & \cdots & \ell_1^r \\ 1 & \ell_2 & \cdots & \ell_2^r \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \ell_m & \cdots & \ell_m^r \end{pmatrix}, \quad \boldsymbol{X}_3 = \begin{pmatrix} 1 & t_1 & \cdots & t_1^s \\ 1 & t_2 & \cdots & t_2^s \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \cdots & t_n^s \end{pmatrix}$$

are the design matrices for spatial and temporal effects, respectively, $\mathcal{C} \in \mathbb{R}^{r_1\times r_2\times r_3}$ is the unknown core tensor, and $\boldsymbol{M}_k$ are unknown dimension reduction matrices on each mode. The factors $\boldsymbol{X}_k\boldsymbol{M}_k$ are reduced features in the mean model (2). The spatial-temporal model is a special case of our supervised tensor decomposition model (1), with features available on each of the three modes.

Example 2 (Network population model). Network response model (Rabusseau and Kadri 2016) is recently developed for neuroimaging analysis. The goal is to study the relationship between brain network connectivity pattern and features of individuals. Suppose we have a sample of n observations, $\{(\boldsymbol{Y}_i, \boldsymbol{x}_i)\colon i = 1, \ldots, n\}$, where for each individual $i \in [n]$, $\boldsymbol{Y}_i \in \{0,1\}^{d\times d}$ is the undirected adjacency matrix whose entries indicate presences/absences of connectivities between d brain nodes, and $\boldsymbol{x}_i \in \mathbb{R}^p$ is the individual's feature such as age, gender, cognition score, etc. The network-response model has the conditional mean

$$\text{logit}(\mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{x}_i)) = \mathcal{B} \times_3 \boldsymbol{x}_i, \quad \text{for } i = 1, \ldots, n, \tag{3}$$

where $\mathcal{B} \in \mathbb{R}^{d\times d\times p}$ is a rank-(r_1, r_1, r_2) coefficient tensor, and $\mathcal{B}$ is assumed to be symmetric in the first two modes.

The model (3) is a special case of our supervised tensor decomposition, with feature matrix on the last mode of the tensor. Specifically, we stack the network observations $\{\boldsymbol{Y}_i\}$ together and obtain an order-3 response tensor $\mathcal{Y} \in \{0,1\}^{d\times d\times n}$. Define a feature matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$. Then, the model (3) has the equivalent representation of supervised tensor decomposition,

$$\text{logit}(\mathbb{E}(\mathcal{Y}|\boldsymbol{X})) = \mathcal{C} \times \{\boldsymbol{M},\ \boldsymbol{M},\ \boldsymbol{X}\boldsymbol{M}'\},$$

where $\mathcal{C} \in \mathbb{R}^{r_1\times r_1\times r_2}$ is the core tensor, $\boldsymbol{M} \in \mathbb{R}^{d\times r_1}$ is the dimension reduction matrix on the first two modes, and $\boldsymbol{M}' \in \mathbb{R}^{p\times r_2}$ is for the last mode.

Example 3 (Dyadic data with node attributes). Dyadic dataset consists of measurements on pairs of objects. Common examples include graphs and networks. Let $\mathcal{G} = (V, E)$ denote a graph, where $V = [d]$ is the node set of the graph, and $E \subset V \times V$ is the edge set. Suppose that we also observe feature vector $\boldsymbol{x}_i \in \mathbb{R}^p$ associated to each node $i \in V$. A probabilistic model on the graph $\mathcal{G} = (V, E)$ can be described by the following matrix regression. The edge connects the two vertices i and j independently of other pairs, and the probability of connection is modeled as

$$\text{logit}\left(\mathbb{P}\left((i,j) \in E\right)\right) = \boldsymbol{x}_i^T\boldsymbol{B}\boldsymbol{x}_j = \langle \boldsymbol{B},\ \boldsymbol{x}_i^T\boldsymbol{x}_j\rangle, \tag{4}$$

where $\boldsymbol{B} \in \mathbb{R}^{p\times p}$ is a symmetric rank-r matrix. The low-rankness in $\boldsymbol{B}$ has demonstrated its success in modeling transitivity, balance, and communities in networks (Hoff 2005). We show that our supervised tensor decompostion (1) also incorporates the graph model as a special case. Let $\mathcal{Y} = [y_{ij}]$ be a binary matrix where $y_{ij} = \mathbb{1}_{(i,j)\in E}$. Define $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$. Then, the graph model (4) can be expressed as

$$\text{logit}(\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})) = \boldsymbol{C} \times \{\boldsymbol{X}\boldsymbol{M},\ \boldsymbol{X}\boldsymbol{M}\},$$

where $\boldsymbol{C} \in \mathbb{R}^{r\times r}, \boldsymbol{M} \in \mathbb{R}^{p\times r}$ are from the singular value decomposition of $\boldsymbol{B} = \boldsymbol{M}\boldsymbol{C}\boldsymbol{M}^T$.

In the above three examples and many other studies, researchers are interested in uncovering the variation in the data tensor that can be explained by features. Our supervised tensor decomposition (1) allows arbitrary numbers of feature matrices. When certain mode k has no side information, we set $\boldsymbol{X}_k = \boldsymbol{I}$ in the model (1). In particular, our model (1) reduces to classical unsupervised tensor decomposition (De Lathauwer, De Moor, and Vandewalle 2000; Hong, Kolda, and Duersch 2020) when no side information is available; that is, $\boldsymbol{X}_k = \boldsymbol{I}$ for all $k \in [K]$.

4. Estimation

4.1. Rank-Constrained MLE

We develop a likelihood-based procedure to estimate $\mathcal{C}$ and $\boldsymbol{M}_k$ in (1). We adopt the exponential family as a flexible framework for different data types. In a classical generalized linear model with a scalar response y and feature $\boldsymbol{x}$, the density is expressed as

$$p(y|\boldsymbol{x}, \boldsymbol{\beta}) = c(y, \phi)\exp\left(\frac{y\theta - b(\theta)}{\phi}\right) \text{ with } \theta = \boldsymbol{\beta}^T\boldsymbol{x},$$

where $b(\cdot)$ is a known function, θ is the linear predictor, $\phi > 0$ is the dispersion parameter, and $c(\cdot)$ is a known normalizing function. The choice of link functions depends on the data types and on the observation domain of y, denoted $\mathbb{Y}$. For example, the observation domain is $\mathbb{Y} = \mathbb{R}$ for continuous data, $\mathbb{Y} = \mathbb{N}$ for count data, and $\mathbb{Y} = \{0,1\}$ for binary data. The canonical link function f is chosen to be $f(\cdot) = b'(\cdot)$, the first-order derivative of $b(\cdot)$. Table 1 summarizes the canonical link functions for common types of distributions.

In our context, we model the entries in data tensor $\mathcal{Y}$, conditional on linear predictor Θ, as independent draws from an

Table 1. Canonical links for common distributions.

Data type	Gaussian	Poisson	Bernoulli
Domain $\mathbb{Y}$	$\mathbb{R}$	$\mathbb{N}$	$\{0,1\}$
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1+\exp(\theta))$
link $f(\theta)$	θ	$\exp(\theta)$	$(1+\exp(-\theta))^{-1}$

exponential family. Ignoring constants that do not depend on Θ, the quasi log-likelihood of Equation (1) is equal to Bregman distance between $\mathcal{Y}$ and $b'(\Theta)$

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K) = \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1,\ldots,i_K} b(\theta_{i_1,\ldots,i_K}),$$
$$\text{where} \quad \Theta = \mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1, \ldots, \boldsymbol{X}_K\boldsymbol{M}_K\}. \tag{5}$$

We propose the constrained maximum quasi-likelihood estimate (MLE),

$$(\hat{\mathcal{C}}_{\text{MLE}}, \hat{\boldsymbol{M}}_{1,\text{MLE}}, \ldots, \hat{\boldsymbol{M}}_{K,\text{MLE}}) = \arg\max_{(\mathcal{C},\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K)\in\mathcal{P}(\boldsymbol{r})} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K), \tag{6}$$

where the parameter space $\mathcal{P}(\boldsymbol{r})$ is defined by

$$\mathcal{P}(\boldsymbol{r}) = \left\{ (\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K) \;\middle|\; \boldsymbol{M}_k \in \mathbb{O}(p_k, r_k) \text{ for all } k \in [K], \; ||\Theta||_\infty \le \alpha \right\}, \tag{7}$$

with a large constant $\alpha > 0$. Recall that $\mathcal{B} = \mathcal{C} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{M}_K\}$ by definition. Correspondingly, we estimate the coefficient tensor $\mathcal{B}$ by

$$\hat{\mathcal{B}}_{\text{MLE}} = \hat{\mathcal{C}}_{\text{MLE}} \times \{\hat{\boldsymbol{M}}_{1,\text{MLE}}, \ldots, \hat{\boldsymbol{M}}_{K,\text{MLE}}\}.$$

The maximum norm constraint on the linear predictor Θ is a technical condition to ensures the existence (boundedness) of MLE. The condition precludes the ill-defined MLE when the optimizer of (6) diverges to $\pm\infty$; this phenomenon may happen in logistic regression when the Bernoulli responses $\{0, 1\}$ are perfectly separable by covariates (Wang and Li 2020). For Gaussian models, no maximum norm constraint is needed. In Section 4.2, we show that setting α to an extremely large constant does not compromise the statistical rate in quantities of interest. In practice, the unbounded search is often indistinguishable from the bounded search, since the boundary constraint $\|\Theta\|_\infty \le \alpha$ would likely never be active. Similar techniques are commonly used in high-dimensional non-Gaussian problems (Wang and Li 2020; Han, Willett, and Zhang 2020).

The optimization (6) is a non-convex problem with possibly local optimizers. We propose an alternating optimization algorithm to *approximately* solve Equation (6). The decision variables in the objective function (6) consist of $K + 1$ blocks of variables, one for the core tensor $\mathcal{C}$ and K for the factor matrices $\boldsymbol{M}_k$. We notice that, if any K out of the $K + 1$ blocks of variables are known, then the optimization reduces to a simple GLM with respect to the last block of variables. This observation leads to an iterative updating scheme for one block at a time while keeping others fixed. Given an initialization $(\hat{\mathcal{C}}^{(0)}, \hat{\boldsymbol{M}}_1^{(0)}, \ldots, \hat{\boldsymbol{M}}_K^{(0)})$ to be described in the next paragraph, the tth iterate from the algorithm is denoted $(\hat{\mathcal{C}}^{(t)}, \hat{\boldsymbol{M}}_1^{(t)}, \ldots, \hat{\boldsymbol{M}}_K^{(t)})$ for $t = 1, 2, 3, \ldots$ The iteration scheme is detailed in Algorithm 1.

We provide two initialization schemes, one with QR-adjusted spectral initialization (warm initialization), and the other with random initialization (cold initialization). The warm initialization is an extension of unsupervised spectral initialization (Zhang and Xia 2018) to supervised setting with multiple feature matrices. Specifically, we project normalized data tensor $\bar{\mathcal{Y}}$ to the normalized multilinear feature space and obtain an unconstrained coefficient tensor $\bar{\mathcal{B}}^{(0)}$. We perform a rank-$\boldsymbol{r}$ higher-order SVD (HOSVD) on $\bar{\mathcal{B}}$, which yields the rank-constrained $\hat{\mathcal{B}}^{(0)}$. The desired initialization is obtained by re-normalizing $\hat{\mathcal{B}}^{(0)}$ back to the original scales of features. The initialization scheme is described in Algorithm 2.

The warm initialization enjoys provable accuracy guarantees at a cost of extra technical assumptions (see Section 4.2). The cold initialization, on the other hand, shows robust in practice but its theoretical guarantee remains an open challenge (Luo and Zhang 2021). We incorporate both options in our software package to provide flexibility to practitioners.

Algorithm 1 Supervised Tensor Decomposition with Side Information

Input: Response tensor $\mathcal{Y} \in \mathbb{R}^{d_1\times\cdots\times d_K}$, feature matrices $\boldsymbol{X}_k \in \mathbb{R}^{d_k\times p_k}$ for $k = 1, \ldots, K$, target rank $\boldsymbol{r} = (r_1, \ldots, r_K)$, link function f, initialization $(\hat{\mathcal{C}}^{(0)}, \hat{\boldsymbol{M}}_1^{(0)}, \ldots, \hat{\boldsymbol{M}}_K^{(0)})$.

1: **for** $t = 1, 2, 3, \ldots$ **do**
2: **for** $k = 1$ to K **do**
3: Obtain the factor matrix $\hat{\boldsymbol{M}}_k^{(t)} \in \mathbb{R}^{p_k\times r_k}$ by a GLM with link function f.
4: Perform QR factorization $\hat{\boldsymbol{M}}_k^{(t)} = \boldsymbol{Q}_k\boldsymbol{R}_k$, where $\boldsymbol{Q}_k \in \mathbb{O}(p_k, r_k)$.
5: Update $\hat{\boldsymbol{M}}_k^{(t)} \leftarrow \boldsymbol{Q}_k$ and core tensor $\hat{\mathcal{C}}^{(t)} \leftarrow \hat{\mathcal{C}}^{(t)} \times_k \boldsymbol{R}_k$.
6: **end for**
7: Update the core tensor $\mathcal{C}$ by solving a GLM with $\text{vec}(\mathcal{Y})$ as response, $\otimes_{k=1}^K [\boldsymbol{X}_k\boldsymbol{M}_k]$ as features, and f as link function. Here $\otimes$ denotes the Kronecker product of matrices.
8: **end for**

Output: factor estimate $(\hat{\mathcal{C}}^{(t)}, \hat{\boldsymbol{M}}_1^{(t)}, \ldots, \hat{\boldsymbol{M}}_K^{(t)})$ from the t-th iterate, and coefficient tensor estimate $\hat{\mathcal{B}}^{(t)} = \hat{\mathcal{C}}^{(t)} \times \{\hat{\boldsymbol{M}}_1^{(t)}, \ldots, \hat{\boldsymbol{M}}_K^{(t)}\}$.

Algorithm 2 QR-adjusted spectral initialization

Input: Response tensor $\mathcal{Y} \in \mathbb{R}^{d_1\times\cdots\times d_K}$, feature matrices $\boldsymbol{X}_k \in \mathbb{R}^{d_k\times p_k}$, Tucker rank $\boldsymbol{r}$.

1: Normalize date tensor $\bar{\mathcal{Y}} \leftarrow \mathcal{Y}$ for Gaussian model, $\bar{\mathcal{Y}} \leftarrow 2\mathcal{Y} - 1$ for Bernoulli model, and $\bar{\mathcal{Y}} \leftarrow \log(\mathcal{Y} + 0.5)$ for Poisson model.
2: Normalize feature matrices via QR factorization $\boldsymbol{X}_k = \boldsymbol{Q}_k\boldsymbol{R}_k$ for all $k \in [K]$.
3: Obtain $\bar{\mathcal{B}} \leftarrow \bar{\mathcal{Y}} \times \{\boldsymbol{Q}_1^T, \ldots, \boldsymbol{Q}_K^T\}$ by projecting $\bar{\mathcal{Y}}$ to the multilinear feature space.
4: Obtain $\hat{\mathcal{B}}^{(0)} \leftarrow \text{HOSVD}(\bar{\mathcal{B}}, \boldsymbol{r})$.
5: Normalize representation $\{\hat{\mathcal{C}}^{(0)}, \hat{\boldsymbol{M}}_1^{(0)}, \ldots, \hat{\boldsymbol{M}}_K^{(0)}\}$ such that $\hat{\mathcal{C}}^{(0)} \times \{\hat{\boldsymbol{M}}_1^{(0)}, \ldots, \hat{\boldsymbol{M}}_K^{(1)}\} = \hat{\mathcal{B}}^{(0)} \times \{\boldsymbol{R}_1^{-1}, \ldots, \boldsymbol{R}_K^{-1}\}$ and $\hat{\boldsymbol{M}}_k^{(0)} \in \mathbb{O}(p, r)$ for all $k \in [K]$.

Output: Core tensor $\hat{\mathcal{C}}^{(0)}$ and factors $\hat{\boldsymbol{M}}_k^{(0)}$ for all $k \in [K]$.

4.2. Statistical Accuracy

This section presents the accuracy guarantees for both global and local optimizers of Equation (6). We first provide the sta-

tistical accuracy for the global MLE (6). Then, we provide the convergence rate for the local optimizer from Algorithm 1 with warm initialization. The rate reveals an interesting interplay between statistical and computational efficiency. We show that a polynomial number of iterations suffices to reach the desired accuracy under certain assumptions. The empirical performance for cold initialization is also investigated.

For cleaner exposition, we present the results for balanced setting in this section, that is, $p_1 = \cdots = p_K = p$, $r_1 = \cdots = r_K = r$, and $d_1 = \cdots = d_K = d$. The general setting follows exactly the same framework and incurs only notational complexity. We are particularly interested in the high-dimensional regime in which both d and p grows while $p \leq d$. The requirement $p \leq d$ is necessary to ensure rank nondeficiency of feature matrices $\boldsymbol{X}_k$. The classical MLE theory is not directly applicable, because the number of unknown parameters grows with the size of data tensor. We leverage the recent development in random tensor theory and high-dimensional statistics to establish the error bounds of the estimation.

Assumption 1. We make the following assumptions:

A1. There exist two positive constants $c_1, c_2 > 0$ such that $c_1 \leq \sigma_{\min}(\boldsymbol{X}_k) \leq \sigma_{\max}(\boldsymbol{X}_k) \leq c_2$ for all $k \in [K]$. Here $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the smallest and largest matrix singular values.

A1′. The feature matrices $\boldsymbol{X}_k$ are Gaussian designs with iid $N(0,1)$ entries.

A2. There exist two positive constants $L, U > 0$, such that $\min_{|\theta|\leq\alpha} b''(\theta) \geq \phi L$ and $\sup_{\theta\in\mathbb{R}} b''(\theta) \leq \phi U$. Here, α is the upper bound of the linear predictor in Equation (6), and $b''(\cdot)$ denotes the second-order derivative.

The assumptions are fairly mild. Assumptions A1 and A1' consider two separate scenarios about feature matrices. Assumption A1 is applicable when feature matrix is asymptotically nonsingular and has bounded spectral norm, whereas Assumption A1' imposes the commonly-used Gaussian design (Raskutti, Yuan, and Chen 2019). The Assumption 2 is essentially imposed to the response variance because of the identity $\mathrm{Var}(y|\theta) = \phi b''(\theta)$ (McCullagh and Nelder 1989). The lower bound ensures the non-degeneracy of the variance in the feasible domain of θ, whereas the upper bound ensures the finiteness of the variance in the entire family. In fact, except for Poisson responses, most members in the exponential family, for example, Gaussian, Bernoulli, and binomial responses, satisfy this condition.

4.2.1. Statistical Accuracy for Global Optimizers

We need some extra notation to state the results in full generality. Recall that the factor matrices $\boldsymbol{M}_k$ are identifiable only up to orthogonal rotations. Therefore, we choose to use angle distance to assess the estimation accuracy of $\boldsymbol{M}_k$. For any two column-orthonormal matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{O}(d, r)$ of same dimension, the angle distance is defined as

$$\sin\Theta(\boldsymbol{A},\boldsymbol{B}) = \max\left\{\frac{\langle \boldsymbol{x},\boldsymbol{y}\rangle}{||\boldsymbol{x}||_2||\boldsymbol{y}||_2} : \boldsymbol{x} \in \mathrm{Span}(\boldsymbol{A}),\ \boldsymbol{y} \in \mathrm{Span}(\boldsymbol{B}^{\perp})\right\},$$

where $\mathrm{Span}(\cdot)$ represents the column space of the matrix. We use the superscript "true" to denote the true parameters from generic decision variables in optimization. For instance, $\mathcal{B}_{\text{true}}$ denotes the true coefficient tensor, whereas $\mathcal{B}$ denotes a decision variable in (5).

Define the signal level λ as the minimal singular value of the unfolded matrices obtained from $\mathcal{B}_{\text{true}}$,

$$\lambda = \min_{k\in[K]} \sigma_r(\mathrm{Unfold}_k(\mathcal{B}_{\text{true}})).$$

Intuitively, λ quantifies the level of rank non-degeneracy for the true coefficient tensor $\mathcal{B}_{\text{true}}$.

Theorem 4.1 (Statistical rate for global optimizers). Consider generalized tensor models with multiple feature matrices. Under Assumptions A1 and A2 with scaled feature matrices $\bar{\boldsymbol{X}}_k = \sqrt{d}\boldsymbol{X}_k$, or Assumptions A1' and A2 with original feature matrices, we have

$$\max_{k\in[K]} \sin^2\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_{k,\text{MLE}}) \lesssim \frac{\phi(r^K + Kpr)}{\lambda^2 d^K},$$
$$||\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}_{\text{MLE}}||_F^2 \lesssim \frac{\phi(r^K + Kpr)}{d^K}, \qquad (8)$$

with probability at least $1 - \exp(-p)$.

Theorem 4.1 establishes the statistical convergence for the global MLE (6). The result in (8) implies that the estimation has a convergence rate $\mathcal{O}(Kp/d^K)$ as $(p, d) \to \infty$. This rate agrees with intuition, since in our setting, the number of parameters with K feature matrices is of order $\mathcal{O}(Kp)$, whereas the number of tensor entries $\mathcal{O}(d^K)$ corresponds to the total sample size. Because $p \leq d$, our rate is faster than $\mathcal{O}(d^{-(K-1)})$ obtained by tensor decomposition without features (Wang and Li 2020).

Inspection of our proof (Supplementary Notes) shows that the desired convergence rate holds not only for the MLE, but also for all local optimizers satisfying $\mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K) \geq \mathcal{L}_{\mathcal{Y}}(\mathcal{C}_{\text{true}}, \boldsymbol{M}_{1,\text{true}}, \ldots, \boldsymbol{M}_{K,\text{true}})$. The observation indicates the global optimality is not necessarily a serious concern in our context, as long as the convergent objective is large enough. In next section, we will provide the statistical accuracy for *local* optimizer with provable convergence guarantee, at a cost of extra signal requirement.

4.2.2. Empirical Accuracy for Local Optimizers

The optimization (6) is a non-convex problem due to the low-rank constraint in the feasible set $\mathcal{P}$. Under mild conditions, our warm initialization enjoys stable performance, and the subsequent iterations further improve the accuracy via linear convergence; i.e. sequence of iterates generated by Algorithm 1 converges to optimal solutions at a linear rate.

Proposition 4.1 (Polynomial-time angle estimation). Consider Gaussian tensor models with $b(\theta) = \theta^2/2$ in the objective function (5). Suppose the signal-to-noise ratio $\lambda^2/\phi \geq Cp^{K/2}d^{-K}$ for some sufficiently large universal constant $C > 0$. Under Assumption A1 with scaled feature matrices $\bar{\boldsymbol{X}}_k = \sqrt{d}\boldsymbol{X}_k$, or Assumption A1' with original feature matrices, the outputs from initialization Algorithm 2 and iteration Algorithm 1 satisfy the following two properties.

1. With probability at least $1 - \exp(-p)$.

$$\max_{k\in[K]} \sin^2\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(0)}) \leq \frac{1}{4}. \qquad (9)$$

2. Let $t = 1, 2, 3, \ldots$, denote the iteration. There exists a contraction parameter $\rho \in (0, 1)$, such that, with probability at least $1 - \exp(-p)$,

$$\max_{k\in[K]} \sin^2\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(t)}) \lesssim \underbrace{\frac{\phi p}{\lambda^2 d^K}}_{\text{statistical error}} + \underbrace{\rho^t \max_{k\in[K]} \sin\Theta^2(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(0)})}_{\text{algorithmic error}}. \quad (10)$$

Proposition 4.1 provides the estimation errors for algorithm outputs at initialization and at each of the subsequent iterations. The initialization bound (9) demonstrates the stability of warm initialization under a mild SNR requirement $\lambda^2/\phi \gtrsim p^{K/2}d^{-K}$. We can think of d as the sample size while p the number of parameters at mode K. This threshold is less stringent than $d^{K/2}$ required for unsupervised tensor decomposition features (Han, Willett, and Zhang 2020; Zhang and Xia 2018). The condition confirms that a higher sample size mitigates the required signal level. The iteration bound (10) consists of two terms: the first term is the statistical error, and the second is the algorithmic error. The algorithmic error decays exponentially with the number of iterations, whereas the statistical error remains the same as t grows. The statistical error is unavoidable and also appears in the global MLE; see Theorem 4.1.

As a direct consequence, we find the optimal iteration t after which the algorithmic error is negligible compared to statistical error.

Theorem 4.2 (Statistical rate for local optimizers). Consider the same condition as in Proposition 4.1 and the outputs by combining algorithms 1 and 2. There exists a constant $C > 0$, such that, after $t \gtrsim K \log_{1/\rho} p$ iterations, our algorithm outputs satisfies

$$\max_{k\in[K]} \sin^2\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(t)}) \lesssim \frac{\phi p}{\lambda^2 d^K},$$
$$||\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}^{(t)}||_F^2 \lesssim \frac{\phi(r^K + Kpr)}{d^K}.$$

In practice, the signal level λ is unknown, so the assumption in Theorem 4.2 is challenging to verify in practice. We supply the theory by providing an alternative scheme—random initialization—and investigate its empirical performance. Figure 2 shows the trajectories of objective function for order-3 tensors based on model (1), where $d \in \{25, 30\}$, $p = 0.4d$, $r \in \{3, 6\}$ at all three modes. We consider data tensors with Gaussian, Bernoulli, and Poisson entries. Under all combinations of the dimension d, rank r, and type of the entries, Algorithm 1 converges quickly in a few iterations upon random initialization, and the objective values at convergent points are close to or larger than the value at true parameters. In the experiment we conduct, we find little difference in the final estimation errors between the two initialization schemes. Random initialization appears good enough for Algorithm 1 to find a convergent point with desired statistical guarantees. In practice, we recommend to run both warm and cold initializations, and choose the one with better convergent objective values.

We conclude this section by revisiting the three examples mentioned in Section 3.

Example 1 (Spatio-temporal growth model). The estimated type-by-time-by-space coefficient tensor converges at the rate $\mathcal{O}\left((p + r + s)/(dmn)\right)$ with $(p, r, s) \le (d, m, n)$. The estimation achieves consistency as the dimension grows along either of the three modes.

Example 2 (Network population model). The estimated node-by-node-by-feature tensor converges at the rate $\mathcal{O}\left((2d + p)/(d^2 n)\right)$ with $p \le n$. The estimation achieves consistency as the number of individuals or the number of nodes grows.

Example 3 (Dyadic data with node attributes). The estimated feature-by-feature matrix converges at the rate $\mathcal{O}\left(p/d^2\right)$ with $p \le d$. Again, our estimation achieves consistency as the number of nodes grows.

4.3. Rank Selection and Computational Complexity

Our algorithm assumes the rank $\boldsymbol{r}$ is given. In practice, the rank is often unknown and must be determined from the data. We propose to use Bayesian information criterion (BIC) and choose the rank that minimizes BIC, where

$$\text{BIC}(\boldsymbol{r}) = -2\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_1, \ldots, \hat{\boldsymbol{M}}_K) + p_e(\boldsymbol{r}) \log(\textstyle\prod_k d_k). \quad (11)$$

Here, $p_e(\boldsymbol{r}) \overset{\text{def}}{=} \sum_k (p_k - r_k) r_k + \prod_k r_k$ is the effective number of parameters in the model. We choose $\hat{\boldsymbol{r}}$ that minimizes $\text{BIC}(\boldsymbol{r})$ via grid search. Our choice of BIC aims to balance between the goodness-of-fit for the data and the degree of freedom in the population model. We evaluate the empirical performance of BIC in Section 6.

The computational complexity of our Algorithm is $\mathcal{O}\left(d \sum_k p_k^3\right)$ for each iteration, where $d = \prod_k d_k$ is the total size of the data tensor. The update of K factor matrices is $\mathcal{O}(d \sum_k r_k^3 p_k^3)$ via standard GLM routines. Furthermore, we demonstrate that, under certain SNR conditions, a polynomial number of iterations suffices to reach the desired statistical accuracy. Therefore, the total computational cost is polynomial in p and d.

5. Connection to Other Tensor Regression Methods

We compare our supervised tensor decomposition (*STD*) with recent 12 tensor methods in the literature. Table 2 summarizes these methods with their properties from four aspects: i) model specification, ii) number of feature matrices allowed, (iii) capability of addressing non-Gaussian response, and (iv) capability of addressing non-independent noise. The four closet methods to our are *SupCP* (Lock and Li 2018), *Envelope* (Li and Zhang 2017), *mRRR* (Luo et al. 2018) and *GLSNet* (Zhang, Sun, and Li 2018); these methods all relate a data tensor to feature matrices with low-rank structure on the coefficients. As seen from the table, our method is the only one that allows multiple feature matrices among the five. *Envelope* and *SupCP* are developed for Gaussian data, and the Gaussianity facilities flexible extension to non-independent noise. In particular, *Envelope* allows noise correlation in Kronecker structured form, whereas *SupCP* allows noise correlation implicitly through decomposing the latent factors into fixed effects (related to features) and random

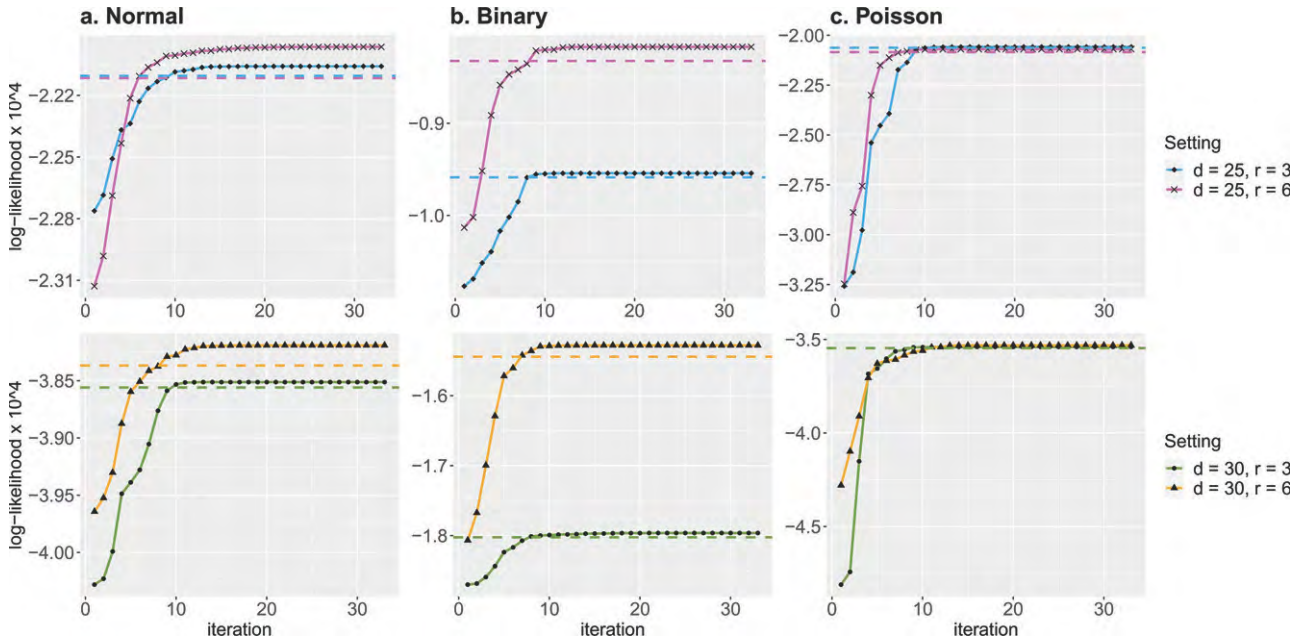

Figure 2. Trajectory of the objective function with various dimension d and rank r under (a) Gaussian, (b) Bernoulli, and (c) Poisson models. The dashed line represents the objective value at true parameters.

Table 2. Comparison of tensor regression/factorization methods.

Method	Model	No. of features	non-Gaussianity	Non-independence
STD (Ours)	$\mathbb{E}\mathcal{Y} = f(\mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\})$, $\mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\}$	3	$\surd$	$\times$
GCP, CP-ARP, CORALS	$\mathbb{E}\mathcal{Y} = f([\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3])$	0	$\surd$	$\times$
DCOT	$\mathbb{E}\mathcal{Y} = f((\mathcal{C}_1 + \mathcal{C}_2) \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\})$	0	$\surd$	$\times$
LRT, CRT	$y_n = \langle \mathcal{B}, \mathcal{X}_n \rangle + \epsilon_n$, various structure on $\mathcal{B}$	0	$\times$	$\times$
STAR	$y_n = \sum_m \langle \mathcal{B}_m, \mathcal{F}_m(\mathcal{X}_{ijk}) \rangle + \epsilon_n$, sparse-CP $\mathcal{B}_m$	0	$\times$	$\times$
SupCP	$\mathcal{Y} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3] + \mathcal{E}$, $\boldsymbol{A}_1 = \boldsymbol{X}\boldsymbol{B} + \mathcal{E}'$, $\mathcal{E} \perp \mathcal{E}'$	1	$\times$	$\surd$
mRRR	$\mathbb{E}\boldsymbol{Y} = f(\boldsymbol{X}\boldsymbol{B})$, low-rank $\boldsymbol{B}$	1	$\surd$	$\times$
Envelope	$\mathcal{Y} = \mathcal{B} \times_3 \boldsymbol{X} + \mathcal{E}$, $\mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{I}\}$, $\mathcal{E} \sim \mathcal{TN}(\Sigma_1, \Sigma_2, \boldsymbol{I})$	1	$\times$	$\surd$
GLSNet	$\mathbb{E}\mathcal{Y} = f(\mathbf{1} \otimes \Theta + \mathcal{B} \times_3 \boldsymbol{X})$, low-rank Θ, sparse $\mathcal{B}$	1	$\surd$	$\times$
STORE	$\mathcal{Y} = \mathcal{B} \times_3 \boldsymbol{X} + \mathcal{E}$, sparse-CP $\mathcal{B}$	1	$\times$	$\times$

NOTES: We focus on order-3 tensors for illustration. Calligraphic letters denote tensors, bold capital letters denote matrices, and little letters denote scalars. The dimension of tensors and matrices can be identified from the contexts.
– Data: tensor response $\mathcal{Y}$, feature matrices $\boldsymbol{X}, \boldsymbol{X}_k$, predictor tensor $\mathcal{X}_n$, scalar response y_n, sample index n, tensor mode $k = 1, 2, 3$.
– Parameter: Tuckor factors $\boldsymbol{M}_k$, CP factors $\boldsymbol{A}_k$, CP decomposition $[\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3]$, coefficient tensor and matrix $\mathcal{B}, \mathcal{B}_m, \Theta, \boldsymbol{B}$.
– Function: a known link function $f(\cdot)$, a known basis function $\mathcal{F}_m(\cdot)$.
– Noise: Gaussian tensor with iid entries $\mathcal{E}, \mathcal{E}'$, Gaussian tensor with Kronecker covariance $\mathcal{E} \sim \mathcal{TN}(\Sigma_1, \Sigma_2, \boldsymbol{I})$, meaning $\text{cov}(\text{vec}(\mathcal{E})) = \Sigma_1 \otimes \Sigma_2 \otimes \boldsymbol{I}$.
– GCP: Generalized canonical polyadic tensor decomposition (Hong, Kolda, and Duersch 2020);
– CP-APR: CP alternating Poisson regression (Chi and Kolda 2012);
– CORALS: Generalized co-clustering method (Li 2020);
– DCOT: Double core tensor decomposition (Tarzanagh and Michailidis 2019);
– SupCP: Supervised PARAFAC/CANDECOMP factorization (Lock and Li 2018);
– mRRR: Mixed-response reduced-rank regression (Luo et al. 2018);
– Envelope: Parsimonious tensor response regression (Li and Zhang 2017);
– GLSNet: Generalized connectivity matrix response regression (Zhang, Sun, and Li 2018);
– STORE: Sparse tensor response regression (Sun and Li 2017);
– LTR: Low-rank tensor regression (Han, Willett, and Zhang 2020);
– CRT: Convex regularized multiresponse tensor regression (Raskutti, Yuan, and Chen 2019);
– STAR: Sparse tensor additive regression (Hao et al. 2021).

effects (unrelated to features). On the other hand, the other three methods (*mRRR*, *GLSNet*, and *STD*) are developed for exponential family distribution with possibly non-additive noise. The generality makes the full modeling of noise correlation computationally challenging. We will compare the numerical performance of these methods in Section 6.

Our model also has a close connection to higher-order interaction model (Hao, Zhang, and Cheng 2020) and tensor-to-tensor regression (Lock 2018). Model (1) can be viewed as a regression model with across-mode interactions in the reduced feature space. We take an order-3 tensor under the Gaussian model for illustration. Let $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}$ denote the feature matrix on mode $k = 1, 2,$ and 3, respectively. Suppose that each mode has two-dimensional reduced features, denoted $\boldsymbol{M}_1\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$, $\boldsymbol{M}_2\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2]$, $\boldsymbol{M}_3\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2]$. Here $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{w}_1, \boldsymbol{w}_2$ are column vectors. Then the model (1) is equivalent to a regression model with across-mode interactions

$$\mathbb{E}(y_{ijk}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = c_{111}\boldsymbol{x}_{1i}\boldsymbol{z}_{1j}\boldsymbol{w}_{1k} + c_{121}\boldsymbol{x}_{i1}\boldsymbol{z}_{j2}\boldsymbol{w}_{k1} + \cdots + c_{221}\boldsymbol{x}_{2i}\boldsymbol{z}_{2j}\boldsymbol{w}_{1k} + c_{222}\boldsymbol{x}_{2i}\boldsymbol{z}_{2j}\boldsymbol{w}_{2k},$$

where $[c_{ijk}] \in \mathbb{R}^{2\times2\times3}$ are unknown interaction effects, $\boldsymbol{x}_{1i}$ denotes the i-th entry in the feature vector $\boldsymbol{x}_1$, and similar notations apply to other features. Note that lower-order interactions are naturally incorporated if we include an intercept column in the reduced feature matrices. The above example shows the connection of our supervised tensor decomposition to multivariate regressions.

6. Numerical Experiments

We evaluate the empirical performance of our supervised tensor decomposition (*STD*) through simulations. We consider order-3 tensors with a range of distribution types. Unless otherwise specified, the conditional mean tensor is generated form model (1), where the core tensor entries are iid drawn from Uniform[-1,1], the factor matrix $\boldsymbol{M}_k$ is uniformly sampled with respect to Haar measure from matrices with orthonormal columns. The feature matrix $\boldsymbol{X}_k$ is either an identity matrix (i.e., no feature available) or Gaussian random matrix with iid entries from $N(0,1)$. The linear predictor $\Theta = \mathcal{C} \times \{\boldsymbol{M}_1\boldsymbol{X}_1, \boldsymbol{M}_2\boldsymbol{X}_2, \boldsymbol{M}_3\boldsymbol{X}_3\}$ is scaled such that $||\Theta||_\infty = 1$. Conditional on the linear predictor $\Theta = [\theta_{ijk}]$, the entries in the tensor $\mathcal{Y} = [y_{ijk}]$ are drawn independently according to three probabilistic models:

(a) Gaussian model: continuous tensor entries $y_{ijk} \sim N\left(\alpha\theta_{ijk}, 1\right)$.
(b) Poisson model: count tensor entries $y_{ijk} \sim \text{Poisson}\left(e^{\alpha\theta_{ijk}}\right)$.
(c) Bernoulli model: binary tensor entries $y_{ijk} \sim \text{Bernoulli}\left(\frac{e^{\alpha\theta_{ijk}}}{1+e^{\alpha\theta_{ijk}}}\right)$.

Here $\alpha > 0$ controls the magnitude of the effect size, which is also the maximum norm of coefficient tensor as in (7). In each experiment, we report the summary statistics averaged across 30 simulation replications.

6.1. Finite-Sample Performance

The first experiment assesses the selection accuracy of our BIC criterion (11). We consider the balanced situation where $d_k = d$, $p_k = 0.4d_k$ for $k = 1, 2, 3$. We set $\alpha = 4$ and consider various combinations of dimension d and rank $\boldsymbol{r} = (r_1, r_2, r_3)$. For each combination, we minimize BIC using a grid search from $(r_1 - 3, r_2 - 3, r_3 - 3)$ to $(r_1 + 3, r_2 + 3, r_3 + 3)$. We remove invalid rank such as $r_{\max}^2 \geq \prod_{k=1}^3 r_k$ and use parallel search to reduce the computational cost. Table 3 reports the selected rank averaged over $n_{\text{sim}} = 30$ replicates. We find that in the high-rank setting with $d = 20$, the selected rank slightly underestimates the true rank, and the accuracy immediately improves when either the dimension increases to $d = 40$ or the rank reduces to $\boldsymbol{r} = (3, 3, 3)$. This agrees with our expectation, because in the tensor decomposition, the sample size is related to the number of tensor entries. A larger d implies a larger sample size, so the BIC selection becomes more accurate.

Table 3. Rank selection via BIC. The estimated ranks are averaged across 30 simulation.

True rank $\boldsymbol{r}$	$d = 20$ (Gaussian)	$d = 40$ (Gaussian)	$d = 20$ (Poisson)	$d = 40$ (Poisson)
(3, 3, 3)	(**3.0**, **3.0**, **3.0**)	(**3.0**, **3.0**, **3.0**)	(**3.0**, **3.0**, **3.0**)	(**3.0**, **3.0**, **3.0**)
(4, 4, 6)	(3.0, 3.0, **4.6**)	(**4.0**, **4.0**, **5.3**)	(3.0, 3.0, **5.3**)	(**4.0**, **4.0**, **5.6**)
(6, 8, 8)	(5.0, 5.0, 5.0)	(**6.0**, **8.0**, **8.0**)	(5.0, 5.0, 6.7)	(**6.0**, **8.0**, **8.0**)

NOTE: Bold number indicates the ground truth is within two standard deviations of the estimate.

The second experiment evaluates the accuracy when features are available on all modes. We set $\alpha = 10, d_k = d, p_k = 0.4d_k, r_k = r \in \{2, 4, 6\}$ and increase d from 30 to 60. Our theoretical analysis suggests that $\hat{\mathcal{B}}$ has a convergence rate $\mathcal{O}(d^{-2})$ in this setting. Figure 3 plots the mean squared error (MSE) $||\hat{\mathcal{B}} - \mathcal{B}_{\text{true}}||_F^2$ versus the effective sample size d^2 under three different distribution models. We find that the empirical MSE decreases roughly at the rate of $1/d^2$, which is consistent with our theoretical results. We also observe that, tensors with higher rank tend to yield higher estimation errors, as reflected by the upward shift of the curves as r increases. Indeed, a larger r implies a higher model complexity and thus greater difficulty in the estimation.

6.2. Comparison With Other Tensor Methods

We compare our supervised tensor decomposition with three other tensor methods:

- Supervised PARAFAC/CANDECOMP factorization (*SupCP*, (Lock and Li 2018)).
- Parsimonious tensor response regression (*Envelope*, (Li and Zhang 2017));
- Mixed-response reduced-rank regression (*mRRR*, (Luo et al. 2018));
- Generalized connectivity matrix response regression (*GLSNet*, (Zhang, Sun, and Li 2018));

These four methods are the closest methods to ours, in that they all relate a data tensor to feature matrices with low-rank structure on the coefficients. We consider Gaussian and Bernoulli tensors in experiments. For methods not applicable for Bernoulli data (*SupCP* and *Envelope*), we provide the algorithm $\{-1, 1\}$-valued tensors as inputs. Because *mRRR* allows matrix response only, we provide the algorithm the unfolded matrix of response tensor as inputs. We measure the accuracy using the response error defined as $1 - \text{Cor}(\hat{\mathcal{Y}}, f(\Theta_{\text{true}}))$, where $\hat{\mathcal{Y}}$ is the fitted tensor from each method, and $f(\Theta_{\text{true}})$ is the true conditional mean of the tensor. Note that the response error is a scale-insensitive metric; a smaller error implies a better fit of the model.

The comparison is assessed from three aspects: (i) benefit of incorporating features from multiple modes; (ii) prediction error with respect to sample size; (iii) robustness of model misspecification. We use similar simulation setups as in our first experiment in last section. We consider rank $\boldsymbol{r} = (3, 3, 3)$ (low) vs. $(4, 5, 6)$ (high), effect size $\alpha = 3$ (low) vs. 6 (high), dimension d ranging from 20 to 100 for modes with features, and $d = 20$ for modes without features. The method *Envelope* and *mRRR* require the tensor rank as inputs, respectively. For fairness, we provide both algorithms the true rank. The methods *SupCP* and *GLSNet* use different notions of model rank, and *GLSNet* takes sparsity as an input. We use a grid search to set

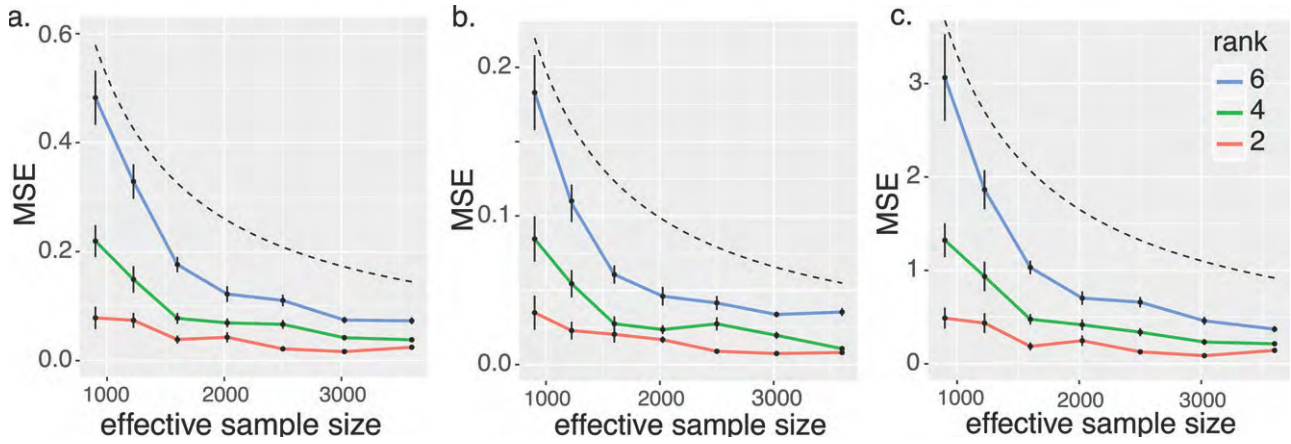

Figure 3. Estimation error against effective sample size. The three panels plot the MSE when the response tensors are generated from (i) Gaussian, (ii) Poisson, and (iii) Bernoulli models. The dashed curves correspond to $\mathcal{O}(1/d^2)$.

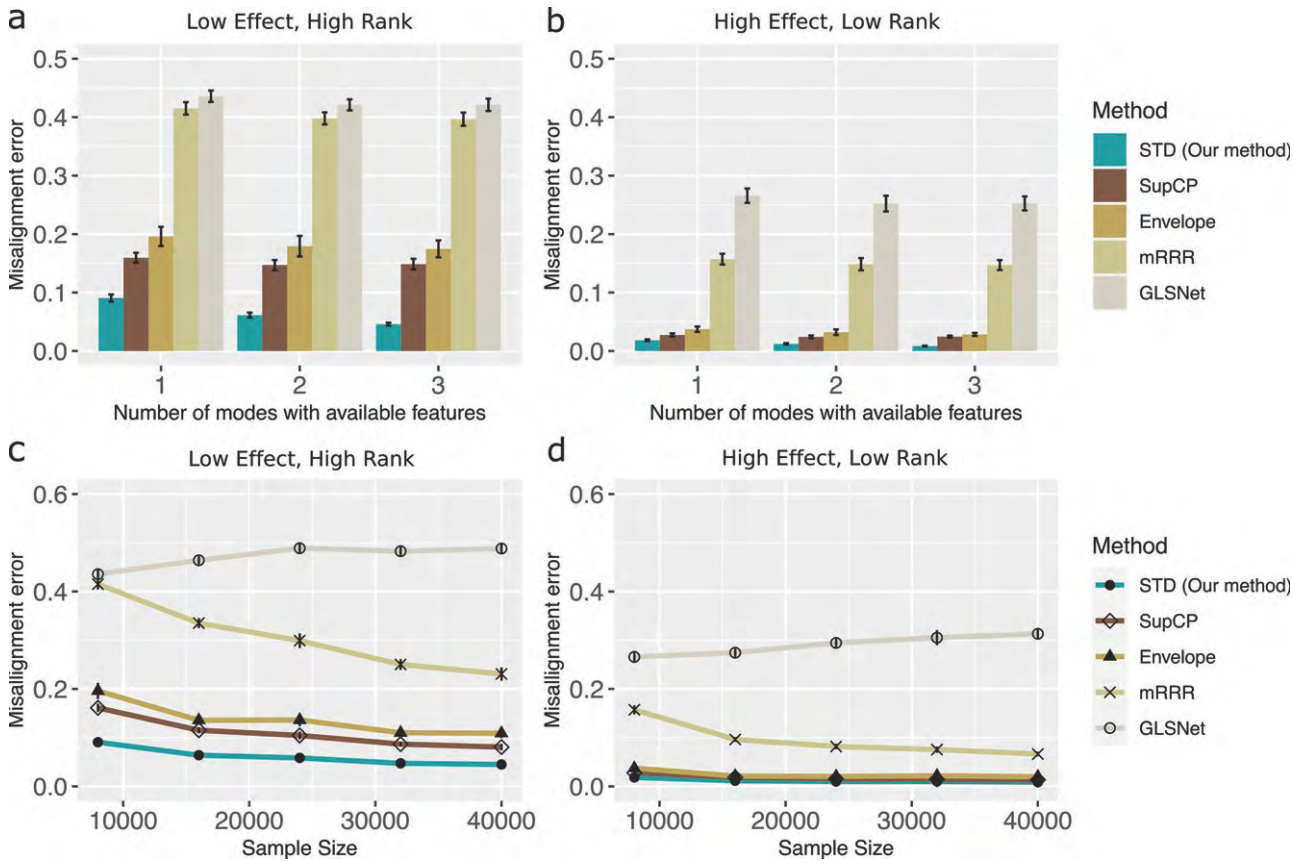

Figure 4. Comparison between tensor methods with Gaussian data. Panels (a) and (b) plot estimation error versus the number of modes with available features. Panels (c) and (d) plot ME versus the effective sample size d^2. We consider rank $\boldsymbol{r} = (3, 3, 3)$ (low), $\boldsymbol{r} = (4, 5, 6)$ (high), and effect size $\alpha = 3$ (low), $\alpha = 6$ (high).

the hyperparameters in *SupCP* and *GLSNet* that give the best performance.

Figures 4(a) and (b) show the impact of features to estimation error. We see that our *STD* outperforms others, especially in the low-effect high-rank setting. As the number of informative modes increases, the *STD* exhibits a reduction in error whereas others remain unchanged. The accuracy gain in Figure 4 demonstrates the benefit of incorporating informative features from multiple modes. In addition, we find that the relative performance among the competing methods reveals the benefits of low-rankness. The second best method is *SupCP* which imposes low-rankness on three modes; the next one is *Envelope* which imposes low-rankness on two modes; the less favorable one is *mRRR* which imposes low-rank structure on one mode only; the worst one is *GLSNet* which imposes sparsity but no low-rankness on the feature effects.

Figures 4(c) and (d) compare the prediction error with respect to effective sample size d^2. For fair comparison, we consider the setting with feature matrix on one mode only. We find that our *STD* method has similar performance as *Envelope*

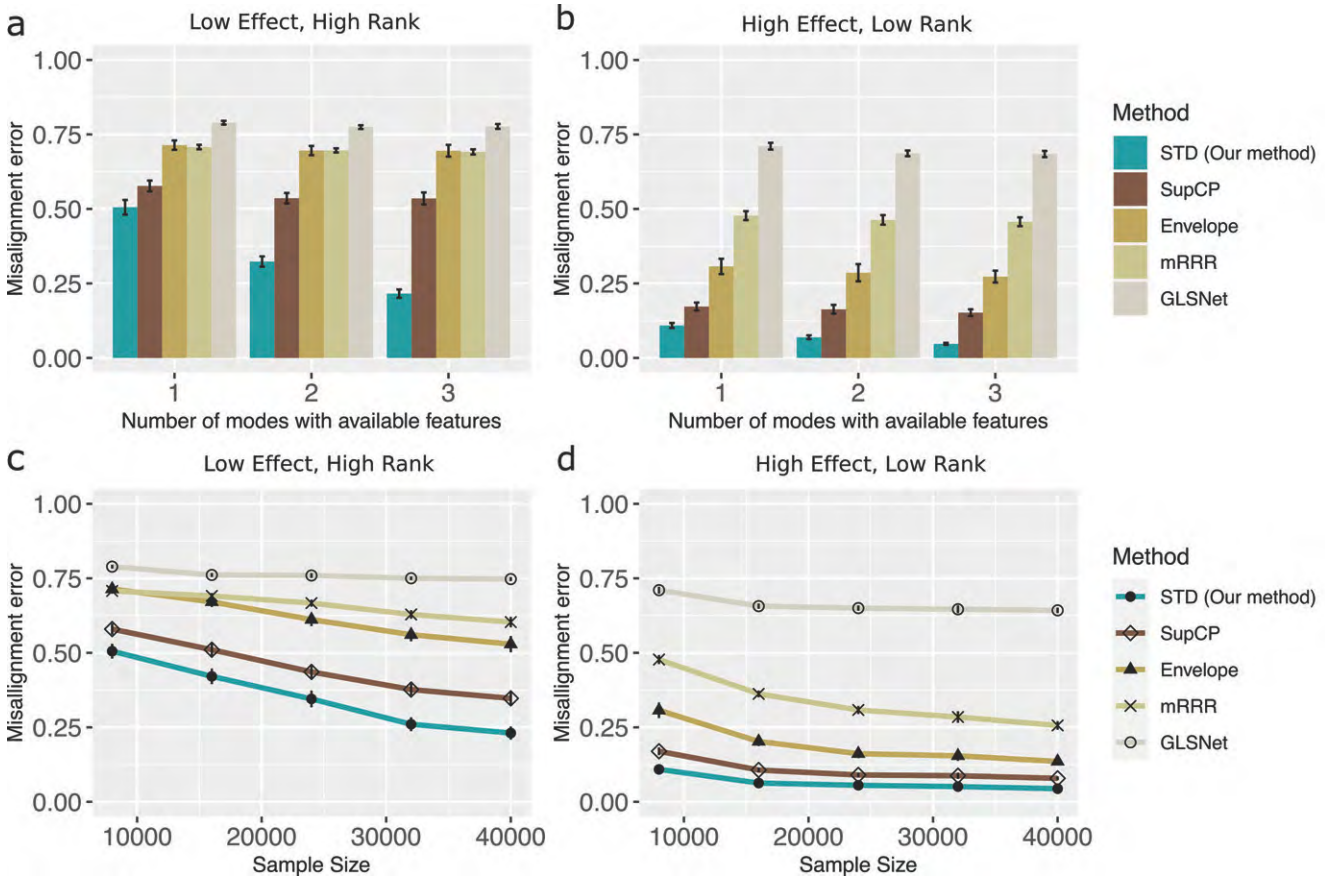

Figure 5. Comparison between tensor methods with Binary data. The panel legends are the same as in Figure 4.

and *SupCP* in the high-effect low-rank regime, whereas the improvement becomes more pronounced in the low-effect high-rank regime. The latter setting is notably harder, and our *STD* method shows advantage in addressing this challenge. Among other methods, *Envelope*, *SupCP*, and *mRRR* show decreasing errors as d increases, implying the benefits of low-rankness methods. In contrast, *GLSNet* suffers from nondecreasing error and indicates the poor fit of sparsity methods in addressing low-rank data.

We also evaluate the performance comparison with Bernoulli tensors. Figure 5 indicates the necessarity of generalized model in addressing non-Gaussian data. Indeed, methods that assume Gaussiannity (*Envelope* and *SucCP*) perform less favorably in Bernoulli setting (Figure 5(c)) compared to Gaussian setting (Figure 4(c)). Our method shows improved accuracy as the number of informative features increases (Figures 5(a) and (b)). In the absence of multiple features, our method still performs favorably compared to others (Figures 5(c) and (d)), for the same reasons we have argued in Gaussian data.

Finally, we assess the performance of our method *STD* under model misspecification. We consider two aspects: (i) non-independent noise, and (ii) sparse feature effects. Note that our method *STD* imposes neither of these two assumptions, so the experiment allows us to assess the robustness. We select competing methods from Table 2 that specifically addresses these two aspects. We use *Envelope* and *SupCP* as benchmark for noise correlation experiment, and *GLSNet* for sparsity experiment.

Figures 6(a) and (b) assesss the impact of noise correlation to the estimation accuracy. The data are simulated from *Envelope* model with envelope dimensions $r = (3, 3)$ (low) and $(4, 5)$ (high). The noise is generated from a zero-mean Gaussian tensor with Kronecker structured covariance; see Supplementary Notes for details. As expected, *Envelope* shows the best performance in the high correlation setting. Remarkably, we find that our method *STD* has comparable and sometimes better performance when noise correlation is moderate-to-low. In contrast, *SupCP* appears less suitable in this setting. Although *SupCP* allows noise correlation implicitly through latent random factors, the induced correlation may not belong to the Kronecker covariance structure in the simulation.

Figures 6(c) and (d) assess the impact of sparsity to estimation performance. We generate data from *GLSNet* model, except that we modify the coefficient tensor to be joint sparse and low-rank (the original *GLSNet* model assumes full-rankness on the coefficient tensor). The sparsity level (x-axis in Figures 6(c) and (d)) quantifies the proportion of zero entries in the coefficient tensor. Since neither our method *STD* nor *GLSNet* follow the simulated model, this setting allows a fair comparison. We find that our method outperforms *GLSNet* in the low-rank setting, whereas *GLSNet* shows a better performance in the high-rank

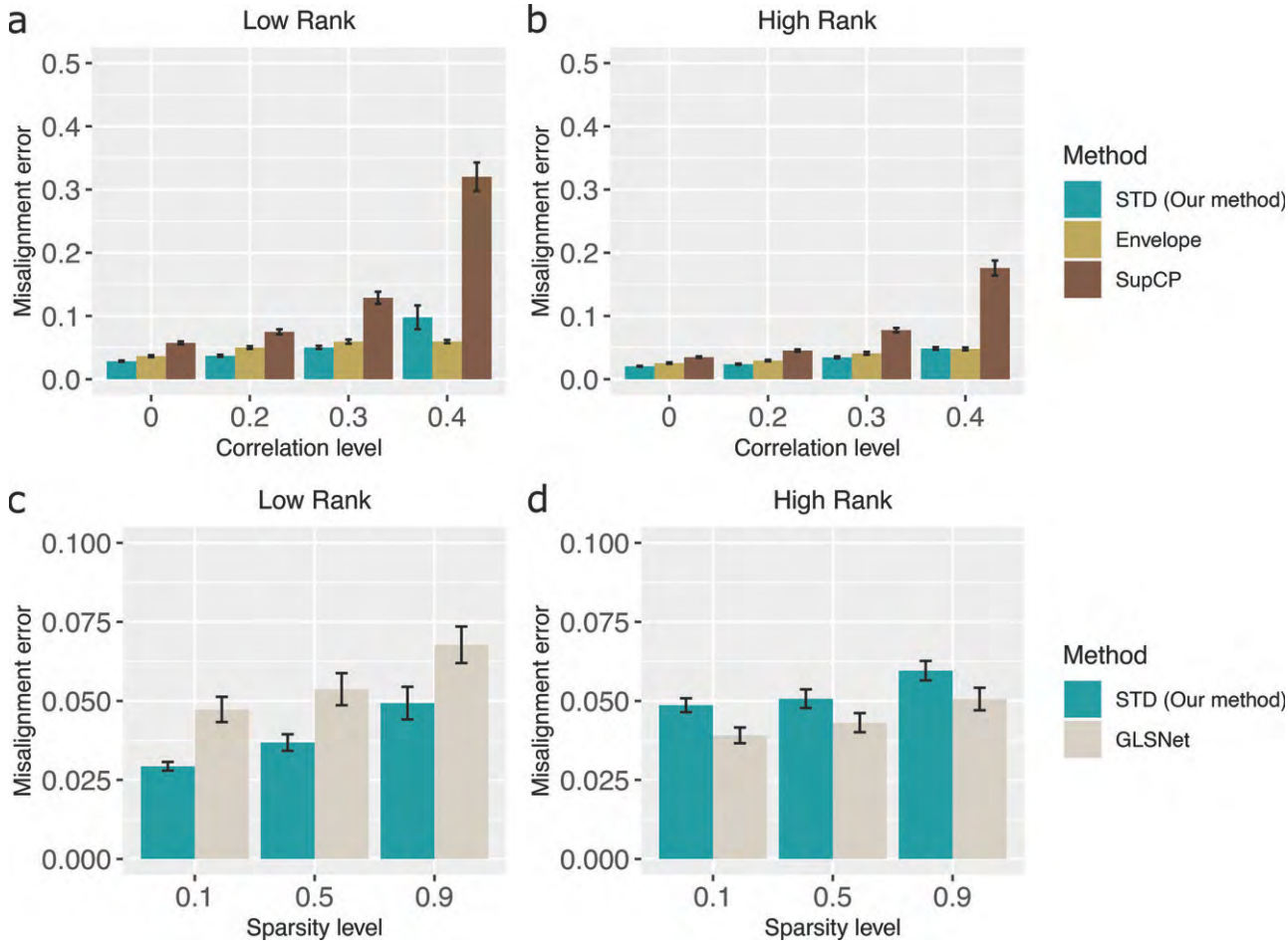

Figure 6. Comparison between tensor methods under model misspecification. Panels (a) and (b) assess the noise correlation, and panels (c)-(d) assess the sparsity.

setting. This observation suggests the robustness of our method to sparsity when the tensor of interest is simultaneously low-rank and sparse. When sparsity is the only salient structure, then methods specifically addressing sparsity would provide a better fit.

7. Data Applications

We apply our supervised tensor decomposition to two datasets. The first application studies the brain networks in response to individual attributes (i.e., feature on one mode), and the second application focuses on multi-relational network analysis with dyadic attributes (i.e., features on two modes).

7.1. Application to Human Brain Connection Data

The Human Connectome Project (HCP) aims to build a network map that characterizes the anatomical and functional connectivity within healthy human brains (Van Essen et al. 2013). We follow the preprocessing procedure as in (Desikan et al. 2006) and parcellate the brain into 68 regions of interest. The dataset consists of 136 brain structural networks, one for each individual. Each brain network is represented as a 68-by-68 binary matrix, where the entries encode the presence or absence of fiber connections between the 68 brain regions. We consider four individual features: gender (65 females vs. 71 males), age 22-25 ($n = 35$), age 26-30 ($n = 58$), and age 31+ ($n = 43$). The preprocessed dataset is released in our R package `tensorregress`. The goal is to identify the connection edges that are affected by individual features. A key challenge in brain network is that the edges are correlated; for example, the nodes in edges may be from a same brain region, and it is of importance to take into account the within-dyad dependence.

We perform the supervised tensor decomposition to the HCP data. The BIC selection suggests a rank $\boldsymbol{r} = (10, 10, 4)$ with quasi log-likelihood $\mathcal{L}_{\mathcal{Y}} = -174654.7$. We utilize the sum-to-zero contrasts in coding the feature effects, and depict only the top 3% edges whose connections are non-constant across the sample. Figure 7 shows the top edges with high effect size, overlaid on the Desikan atlas brain template (Desikan et al. 2006). We find that the global connection exhibits clear spatial separation, and that the nodes within each hemisphere are more densely connected with each other (Figure 7(a)). In particular, the superior-temproal (*SupT*), middle-temporal (*MT*) and Insula are the top three popular nodes in the network. Interestingly, female brains display higher inter-hemispheric connectivity, especially in the frontal, parental, and temporal lobes (Figure 7(b)). This is in agreement with a recent study showing that female brains are optimized for inter-hemispheric communication (Ingalhalikar et al. 2014). We find several edges with declined connection in

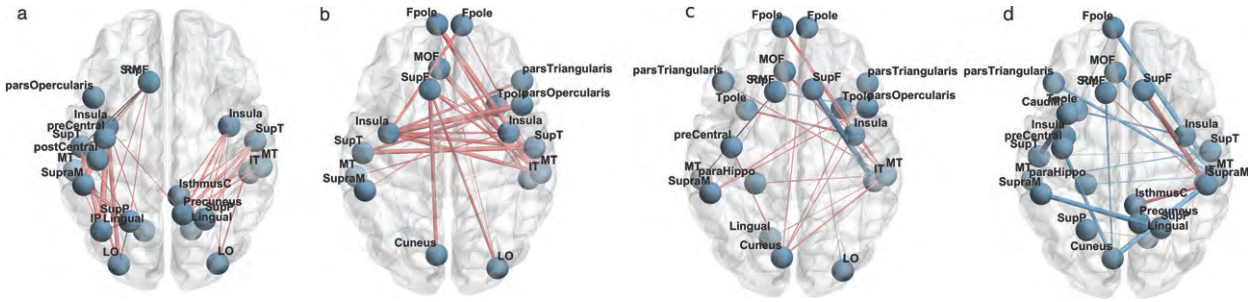

Figure 7. Top edges with large effects. (a) Global effect; (b) Female effect; (c) Aged 22–25; (d) Aged 31+. Red edges represent positive effects and blue edges represent negative effects. The edge-width is proportional to the magnitude of the effect size.

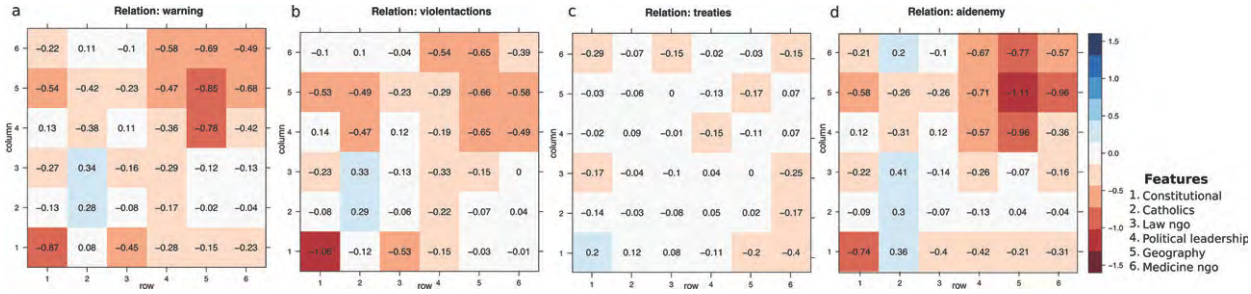

Figure 8. Estimated feature effects in the *Nations* data analysis. Panels (a)–(d) represent the estimated effects of country-level attributes toward the connection probability, for relations *warning*, *violentactions*, *treaties*, and *aidenemy*, respectively.

the group Age 31+. Those edges involve Frontal-pole (*Fploe*), superior-frontal (*SupF*) and Cuneus nodes. The Frontal-pole region is known for its importance in memory and cognition, and the detected decline with age further highlights its biological importance.

7.2. Application to Political Relation Data

The second application studies the multi-relational networks with node-level attributes. We consider *Nations* dataset (Nickel, Tresp, and Kriegel 2011) which records 56 relations among 14 countries between 1950 and 1965. The multi-relational networks can be organized into a $14 \times 14 \times 56$ binary tensor, with each entry indicating the presence or absence of an action, such as "sending tourist to," "export," "import," between countries. The 56 relations span the fields of politics, economics, military, religion, etc. In addition, country-level attributes are also available, and we focus on the following six features: *constitutional, catholics, law ngo, political leadership, geography*, and *medicine ngo*. The goal is to identify the variation in connections due to country-level attributes and their interactions.

We apply our tensor model to the *Nations* data. The multi-relational network $\mathcal{Y}$ is a binary data tensor, and the country attributes $\boldsymbol{X} \in \mathbb{R}^{14\times 6}$ are features on both the 1st and 2nd modes. We use BIC as guidance to select the rank of coefficient tensor $\mathcal{B}$. Since several rank configurations give similar BIC values, we present here the most interpretable results with $\boldsymbol{r} = (4, 4, 4)$. Detailed rank selection procedure is in Supplementary Notes. We perform the supervised tensor decomposition and obtain the dimension reduction matrices $\hat{\boldsymbol{M}}_k$ from the model. Then we apply K-mean clustering to dimension reduction matrix on each of the modes. Table S6 in Supplementary Notes shows the K-means clustering of the 56 relations based on the dimension reduction matrix on the 3rd mode. We find that the relations reflecting the similar aspects of actions are grouped together. In particular, Cluster I consists of military relations such as *violentactions*, *warnings* and *militaryactions*; Clusters II and III capture the economic relations such as *economicaid, booktranslations, tourism*; and Cluster IV represents the political alliance and territory relations.

To investigate the effects of dyadic attributes toward connections, we depict the estimated coefficients $\hat{\mathcal{B}} = [\hat{b}_{ijk}]$ for several relation types (Figure 8). The entry $\hat{b}_{ijk}$ estimates the contribution, at the logit scale, of feature pair (i, j) (ith feature for the "sender" country and jth feature for the "receiver" country) toward the connection of relation k. Several interesting findings emerge from the observation. We find that relations belonging to a same cluster tend to have similar feature effects. For example, the relations "warning" and "violentactions" are classified into Cluster I, and both exhibit similar effect patterns (Figures 8(a) and (b)). Moreover, the feature *constitutional* has a strong effect in the relation "violentactions" and "warning," whereas the effect is weaker in the relation "treaties." The result is plausible because the constitutional attributes affect political actions more often than economical actions. The entries in $\mathcal{B}$ are useful for revealing interaction effects in a context-specific way. From Figure 8, we find a strong interaction between *geography* and *political leadership* in the relation "warning", and a strong interaction between *geogrphy* and *medicine ngo* in the relation "aidenemy". The relation-specific effect pattern showcases the applicability of our method in revealing complex interactions.

8. Discussion and Future Work

We have developed a supervised tensor decomposition method with side information on multiple modes. One important

challenge of tensor data analysis is the complex interdependence among tensor entries and between multiple features. Our approach incorporates side information as feature matrices in the conditional mean tensor. The empirical results demonstrate the improved interpretability and accuracy over previous approaches. Applications to the brain connection and political relationship datasets yield conclusions with sensible interpretations, suggesting the practical utility of the proposed approach.

There are several possible extensions from the work. We have provided accuracy guarantees for parameter estimation in the supervised tensor model. Statistical inference based on tensor decomposition is an important future direction. Measures of uncertainty, such as confidence envelope for space estimation, would be useful. One possible approach would be performing parametric bootstrap (Efron and Tibshirani 1994) to assess the uncertainty in the estimation. For example, one can simulate tensors from the fitted low-rank model based on the estimates, and then assess the empirical distribution of the estimates. While being simple, bootstrap approach is often computationally expensive for large-scale data. Another possibility is to leverage recent development in debiased inference with distributional characterization (Chen et al. 2019). This approach has led to fruitful results for matrix data analysis. Uncertainly quantification involving tensors are generally harder, and establishing distribution theory for tensor estimation remains an open problem.

One assumption made by our method is that tensor entries are conditionally independent given the linear predictor Θ. This assumption can be extended by introducing a more general mixed-effect tensor model. For example, in the special case of Gaussian model, we can model the first two moments of data tensor using

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K) = \mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1,\cdots,\boldsymbol{X}_K\boldsymbol{M}_K\},$$
$$\mathrm{var}(\mathcal{Y}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K) = \Phi_1 \otimes \cdots \otimes \Phi_K,$$

where $\Phi_k \in \mathbb{R}^{d_k \times d_k}$ is the unknown covariance matrix on the mode $k \in [K]$. For general exponential family, an additional mean-variance relationship should also be considered. The joint estimation of mean model Θ and variance model Φ_k will lead to more efficient estimation in the presence of unmeasured confounding effects. However, the introduction of unknown covariance matrices Φ_k dramatically increases the number of parameters in the problem. Suitable regularization such as graphical lasso or specially-structured covariance (Li and Zhang 2017; Lock and Li 2018) should be considered. The extension of tensor modeling with heterogeneous mixed-effects will be an important future direction.

Although we have presented the data applications in the context of order-3 data tensors, the framework of the supervised tensor decomposition applies to a variety of multi-way datasets. One possible application is the integrative analysis of omics data, in which multiple types of omics measurements (gene expression, DNA methylation, microRNA) are collected in the same set of individuals (Lock et al. 2013; Wang, Fischer, and Song 2019). Other applications include time-series tensor data with multiple side information. Exploiting the benefits and properties of supervised tensor decomposition in specialized task will boost scientific discoveries.

Acknowledgments

We thank the editor, the associate editor, and two anonymous reviewers for their constructive feedback, which helped to improve the paper. We would also like to thank Zhuoyan Xu and Chen Zhang for help with the software and Figure 7. This research was supported by NSF grants DMS-1915978, DMS-2023239, and the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research foundation.

Supplementary Materials

Supplementary notes: technical proofs, additional simulation, and data analysis results.

Data and software: Our simulation code, R-package `tensorregress`, and datasets used in the article are available at *https://CRAN.R-project.org/package=tensorregress*

References

Adragni, K. P., and Cook, R. D. (2009), "Sufficient Dimension Reduction and Prediction in Regression," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367, 4385–4405. [205]

Berthet, Q., and Baldin, N. (2020), "Statistical and Computational Rates in Graph Logistic Regression," Proceedings of the 23th International Conference on Artificial Intelligence and Statistics, pp. 2719–2730. [204]

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019), "Inference and Uncertainty Quantification for Noisy Matrix Completion," *Proceedings of the National Academy of Sciences*, 116, pp. 22931–22937. [217]

Chi, E. C., and Kolda, T. G. (2012), "On Tensors, Sparsity, and Nonnegative Factorizations," *SIAM Journal on Matrix Analysis and Applications*, 33, 1272–1299. [205,211]

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000), "A Multilinear Singular Value Decomposition," *SIAM Journal on Matrix Analysis and Applications*, 21, 1253–1278. [205,206,207]

Desikan, R. S., Ségonne, 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, 968–980. [215]

Efron, B., and Tibshirani, R. J. (1994), *"An Introduction to the Bootstrap,"* CRC Monographs on Statistics and Applied Probability Series. New York: Chapman and Hall. [217]

Farias, V. F., and Li, A. A. (2019), "Learning Preferences with Side Information," *Management Science*, 65, 3131–3149. [205]

Gabriel, K. R. (1998), "Generalised Bilinear Regression," *Biometrika*, 85, 689–700. [206]

Gahrooei, M. R., Yan, H., Paynabar, K., and Shi, J. (2020), "Multiple Tensor-on-Tensor Regression: An Approach for Modeling Processes with Heterogeneous Sources of Data," *Technometrics*, 63, 1–23. [205]

Han, R., Willett, R., and Zhang, A. (2020), "An Optimal Statistical and Computational Framework for Generalized Tensor Estimation," The Annals of Statistics, in press. arXiv:2002.11255. [205,208,210,211]

Hao, B., Wang, B., Wang, P., Zhang, J., Yang, J., and Sun, W. W. (2021), "Sparse Tensor Additive Regression," *Journal of Machine Learning Research*, 22, 1–43. [204,211]

Hao, B., Zhang, A., and Cheng, G. (2020), "Sparse and Low-Rank Tensor Estimation Via Cubic Sketchings," *IEEE Transactions on Information Theory*, 66, 5927–5964. [211]

Hoff, P. D. (2005), "Bilinear Mixed-Effects Models for Dyadic Data," *Journal of the American Statistical Association*, 100, 286–295. [204,207]

Hong, D., Kolda, T. G., and Duersch, J. A. (2020), "Generalized Canonical Polyadic Tensor Decomposition," *SIAM Review*, 62, 133–163. [205,207,211]

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Hakonarson, H., Gur, R. E., Gur, R. C., and Verma, R. (2014), "Sex Differences in the Structural Connectome of the Human Brain," *Proceedings of the National Academy of Sciences*, 111, 823–828. [215]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [205,206]

Lee, C., and Wang, M. (2021), "Beyond the Signs: Nonparametric Tensor Completion Via Sign Series," arXiv:2102.00384. [206]

Li, G. (2020), "Generalized Co-Clustering Analysis Via Regularized Alternating Least Squares," *Computational Statistics & Data Analysis*, 150, 106989. [205,206,211]

Li, L., and Zhang, X. (2017), "Parsimonious Tensor Response Regression," *Journal of the American Statistical Association*, 112, 1131–1146. [204,205,206,210,211,212,217]

Lock, E. F. (2018), "Tensor-on-Tensor Regression," *Journal of Computational and Graphical Statistics*, 27, 638–647. [205,211]

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013), "Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types," *The Annals of Applied Statistics*, 7, 523–542. [217]

Lock, E. F., and Li, G. (2018), "Supervised Multiway Factorization," *Electronic Journal of Statistics*, 12, 1150–1180. [204,205,206,210,211,212,217]

Luo, C., Liang, J., Li, G., Wang, F., Zhang, C., Dey, D. K., and Chen, K. (2018), "Leveraging Mixed and Incomplete Outcomes Via Reduced-Rank Modeling," *Journal of Multivariate Analysis*, 167, 378–394. [205,210,211,212]

Luo, Y., and Zhang, A. R. (2021), "Low-Rank Tensor Estimation Via Riemannian Gauss-Newton: Statistical Optimality and Second-Order Convergence," arXiv:2104.12031. [208]

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models* (2nd ed.). CRC Monographs on Statistics and Applied Probability Series. Chapman and Hall. [209]

Nickel, M., Tresp, V., and Kriegel, H.-P. (2011), "A Three-Way Model for Collective Learning on Multi-Relational Data," *Proceedings of the 28th International Conference on Machine Learning*, pp. 809–816. [204,216]

Rabusseau, G., and Kadri, H. (2016), "Low-Rank Regression With Tensor Responses," *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 29, 1875–1883. [207]

Raskutti, G., Yuan, M., and Chen, H. (2019), "Convex Regularization for High-Dimensional Multiresponse Tensor Regression," *The Annals of Statistics*, 47, 1554–1584. [204,205,209,211]

Song, Q., Ge, H., Caverlee, J., and Hu, X. (2019), "Tensor Completion Algorithms in Big Data Analytics," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13, 1–48. [205]

Srivastava, M. S., von Rosen, T., and Von Rosen, D. (2008), "Models With a Kronecker Product Covariance Structure: Estimation and Testing," *Mathematical Methods of Statistics*, 17, 357–370. [206]

Sun, W. W., and Li, L. (2017), "STORE: Sparse Tensor Response Regression and Neuroimaging Analysis," *The Journal of Machine Learning Research*, 18, 4908–4944. [204,205,211]

Tarzanagh, D. A., and Michailidis, G. (2019), "Regularized and Smooth Double Core Tensor Factorization for Heterogeneous Data," arXiv: 1911.10454. [205,211]

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. [215]

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, 1103–1127. [217]

Wang, M., and Li, L. (2020), "Learning From Binary Multiway Data: Probabilistic Tensor Decomposition and its Statistical Optimality," *Journal of Machine Learning Research*, 21, 1–38. [204,208,209]

Zhang, A., and Xia, D. (2018), "Tensor SVD: Statistical and Computational Limits," *IEEE Transactions on Information Theory*, 64, 7311–7338. [208,210]

Zhang, J., Sun, W. W., and Li, L. (2018), "Network Response Regression for Modeling Population of Networks With Covariates," arXiv:1810.03192. [205,210,211,212]

Zhou, H., Li, L., and Zhu, H. (2013), "Tensor Regression With Applications in Neuroimaging Data Analysis," *Journal of the American Statistical Association*, 108, 540–552. [204,205]

Supplementary Notes to "Generalized tensor decomposition with features on multiple modes"

Jiaxin Hu, Chanwoo Lee, Miaoyan Wang

University of Wisconsin-Madison

The supplementary note consists of proofs (Section A), additional simulation results (Section B), and data applications (Section C).

A Proofs

A.1 Proof of Theorem 4.1

We denote several quantities:

$$\underline{\gamma} = \prod_{k\in[K]} \sigma_{\min}(\boldsymbol{X}_k), \quad \bar{\gamma} = \prod_{k\in[K]} \sigma_{\max}(\boldsymbol{X}_k), \quad \lambda = \min_{k\in[K]} \sigma_{r_k}(\text{Unfold}_k(\mathcal{B}_{\text{true}})), \tag{1}$$

where $\underline{\gamma}$ quantifies the rank non-deficiency of feature matrices, $\bar{\gamma}$ quantifies the magnitude of feature matrices, and λ is the smallest singular value of mode-k unfolded matrices $\text{Unfold}_k(\mathcal{B}_{\text{true}})$ for all possible $k \in [K]$. For notational convenience, we drop the subscript $\mathcal{Y}$ from the objective $\mathcal{L}_{\mathcal{Y}}(\cdot)$ and simply write as $\mathcal{L}(\cdot)$. We write $\mathcal{L}(\mathcal{B})$ in place of $\mathcal{L}(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K)$ when we want to emphasize the role of $\mathcal{B}$.

Proposition A.1 (sub-Gaussian residuals)**.** Define the residual tensor $\mathcal{E} = [\![\varepsilon_{i_1,\ldots,i_K}]\!] = \mathcal{Y} - b'(\Theta) \in \mathbb{R}^{d_1\times\cdots\times d_K}$. Under the Assumption A2, $\varepsilon_{i_1,\ldots,i_K}$ is a sub-Gaussian random variable with sub-Gaussian parameter bounded by ϕU, for all $(i_1, \ldots, i_K) \in [d_1]\times\cdots\times[d_K]$.

Proposition A.2 (Properties of tensor GLM)**.** Consider tensor GLMs under Assumption A2.

(a) (Strong convexity) For all $\mathcal{B}$ and all realized data tensor $\mathcal{Y}$,

$$\mathcal{L}(\mathcal{B}_{\text{true}}) \geq \mathcal{L}(\mathcal{B}) + \langle \nabla\mathcal{L}(\mathcal{B}_{\text{true}}), \mathcal{B}_{\text{true}} - \mathcal{B}\rangle + \frac{1}{2}\underline{\gamma}^2 L\|\mathcal{B}_{\text{true}} - \mathcal{B}\|_F^2,$$

where $\nabla L(\cdot)$ denotes the derivative of $\mathcal{L}$ with respect to $\mathcal{B}$.

(b) (Model complexity) Suppose $\mathcal{Y}$ follows generalized tensor model with parameter $\mathcal{B}_{\text{true}}$. Then, with probability at least $1 - \exp(-p)$,

$$\text{Err}_{\text{ideal}}(\boldsymbol{r}) := \sup_{\|\mathcal{B}\|_F=1, \mathcal{B}\in\mathcal{P}(\boldsymbol{r})} \langle \nabla\mathcal{L}(\mathcal{B}_{\text{true}}), \mathcal{B}\rangle \lesssim \bar{\gamma}\sqrt{\phi U(r^K + Kpr)}. \tag{2}$$

The proofs of Propositions A.1-A.2 are in Section A.3.

Proof of Theorem 4.1. First we prove the error bound for $\hat{\mathcal{B}}_{\text{MLE}}$. By the definition of $\hat{\mathcal{B}}_{\text{MLE}}$, $\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) - \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}_{\text{MLE}}) \leq 0$. By the strong convexity in Proposition A.2,

$$0 \geq \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) - \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}_{\text{MLE}}) \geq \langle \nabla\mathcal{L}(\mathcal{B}_{\text{true}}), \mathcal{B}_{\text{true}} - \hat{\mathcal{B}}_{\text{MLE}}\rangle + \frac{1}{2}\underline{\gamma}^2 L\|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}_{\text{MLE}}\|_F^2. \tag{3}$$

Rearranging (3) gives

$$\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F \leq \frac{2}{\underline{\gamma}^2 L}\left\langle \nabla\mathcal{L}(\mathcal{B}_{\text{true}}), \frac{\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}}{\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F}\right\rangle \leq \frac{2}{\underline{\gamma}^2 L}\text{Err}_{\text{ideal}}(2\boldsymbol{r}),$$

where the last inequality comes from the definition of $\text{Err}_{\text{ideal}}(2\boldsymbol{r})$ and the fact that $\text{rank}(\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}) \leq \text{rank}(\hat{\mathcal{B}}_{\text{MLE}}) + \text{rank}(\mathcal{B}_{\text{true}}) \leq 2\boldsymbol{r}$. By (2) in Proposition A.2, we have

$$\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F \lesssim \frac{\bar{\gamma}\sqrt{\phi U}}{\underline{\gamma}^2 L}\sqrt{r^K + Kpr}, \tag{4}$$

with probability at least $1 - \exp(-p)$.

Now, we specialize $\bar{\gamma}/\underline{\gamma}^2$ in the following two cases of assumptions on feature matrices.

[Case 1] Under Assumption A1 with scaled feature matrices, we have

$$\frac{\bar{\gamma}}{\underline{\gamma}^2} \leq \frac{c_2^K d^{K/2}}{c_1^{2K} d^K} \lesssim \sqrt{\frac{1}{d^K}}. \tag{5}$$

[Case 2] Under Assumption A1' with original feature matrices, the asymptotic behavior of

extreme singular values (Rudelson and Vershynin, 2010) are

$$\sigma_{\min}(\boldsymbol{X}_k) \asymp \sqrt{d} - \sqrt{p} \text{ and } \sigma_{\max}(\boldsymbol{X}_k) \asymp \sqrt{d} + \sqrt{p}, \quad \text{for all } k \in [K].$$

In this case, we obtain

$$\frac{\bar{\gamma}}{\underline{\gamma}^2} \asymp \frac{(\sqrt{d}+\sqrt{p})^K}{(\sqrt{d}-\sqrt{p})^{2K}} \lesssim \sqrt{\frac{1}{d^K}}. \tag{6}$$

Combining (4) with either (5) or (6), in both cases we obtain the same conclusion

$$\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F^2 \lesssim \frac{\phi(r^K + Kpr)}{d^K}. \tag{7}$$

Now we prove the bound for sinΘ distance. We unfold tensors $\mathcal{B}_{\text{true}}$ and $\hat{\mathcal{B}}_{\text{MLE}}$ along the mode k and obtain $\text{Unfold}_k(\mathcal{B}_{\text{true}})$ and $\text{Unfold}_k(\hat{\mathcal{B}}_{\text{MLE}})$. Notice that $\boldsymbol{M}_{k,\text{true}}$ and $\hat{\boldsymbol{M}}_{k,\text{MLE}}$ span the top-r left singular spaces of $\text{Unfold}_k(\mathcal{B}_{\text{true}})$ and $\text{Unfold}_k(\hat{\mathcal{B}}_{\text{MLE}})$, respectively. Applying Proposition A.2 to this setting gives

$$\sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_{k,\text{MLE}}) \leq \frac{\|\text{Unfold}_k(\hat{\mathcal{B}}_{\text{MLE}}) - \text{Unfold}_k(\mathcal{B}_{\text{true}})\|_F}{\sigma_{r_k}(\text{Unfold}_k(\mathcal{B}_{\text{true}}))} = \frac{\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F}{\lambda}. \tag{8}$$

The proof is complete by combining (7) and (8). □

A.2 Proofs of Proposition 4.1 and Theorem 4.2

Proof of Proposition 4.1. We express the Gaussian model as

$$\mathcal{Y} = \mathcal{B}_{\text{true}} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} + \mathcal{E},$$

where $\mathcal{E}$ is a noise tensor consisting of i.i.d. entries from $N(0, \sqrt{\phi})$. By QR decomposition on feature matrices, $\boldsymbol{X}_k = \boldsymbol{Q}_k \boldsymbol{R}_k$ for all $k \in [K]$, we have

$$\bar{\mathcal{Y}} = \mathcal{B}_{\text{true}} \times \{\boldsymbol{R}_1, \ldots \boldsymbol{R}_K\} + \bar{\mathcal{E}}, \tag{9}$$

where $\bar{\mathcal{Y}} = \mathcal{Y} \times \{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K\}$ and $\bar{\mathcal{E}} = \mathcal{E} \times \{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K\}$. Notice that entries of $\bar{\mathcal{E}} \in \mathbb{R}^{p \times \cdots \times p}$ are i.i.d drawn from $N(0, \sqrt{\phi})$ by the orthonormality of $\{\boldsymbol{Q}_k\}_{k=1}^K$. Reparameterize the signal

in (9) as

$$\begin{aligned}\mathcal{S}_{\text{true}} := \mathcal{B}_{\text{true}} \times \{\boldsymbol{R}_1, \dots \boldsymbol{R}_K\} &= \mathcal{C}_{\text{true}} \times \{\boldsymbol{R}_1 \boldsymbol{M}_{1,\text{true}}, \dots \boldsymbol{R}_K \boldsymbol{M}_{K,\text{true}}\} \\ &= \mathcal{C}'_{\text{true}} \times \{\boldsymbol{U}_{1,\text{true}}, \dots, \boldsymbol{U}_{K,\text{true}}\}, \end{aligned} \tag{10}$$

where $\boldsymbol{U}_{k.\text{true}} \in \mathbb{O}(p_k, r_k)$ are orthonormal matrices and $\mathcal{C}'_{\text{true}} \in \mathbb{R}^{r \times \cdots \times r}$ is a full rank core tensor. By definition of quantities in (1), we have

$$\lambda' := \min_{k \in [K]} \sigma_{\min}(\text{Unfold}_k(\mathcal{S}_{\text{true}})) \in [\lambda\underline{\gamma},\ \lambda\bar{\gamma}]. \tag{11}$$

Now our setup shares the same setting as in Zhang and Xia (2018, Theorem 1). We summarize the relationships between our algorithm outputs and the ones in Zhang and Xia (2018). For all $k \in [K]$,

1. $\boldsymbol{M}_{k,\text{true}} = \text{SVD}_{r_k}\left(\boldsymbol{R}_k^{-1}\boldsymbol{U}_{k,\text{true}}\right) :=$ the first r_k left singular vectors of $\boldsymbol{R}_k^{-1}\boldsymbol{U}_{k,\text{true}}$;
2. $\hat{\boldsymbol{M}}_k^{(t)} = \text{SVD}_{r_k}\left(\boldsymbol{R}_k^{-1}\hat{\boldsymbol{U}}_k^{(t)}\right)$ for all $t = 0, 1, 2, \dots$;

where $\hat{\boldsymbol{U}}_k^{(t)}$ denotes the t-th iteration output of Higher Order Orthogonal Iteration (HOOI) algorithm (Zhang and Xia, 2018) with inputs $\bar{\mathcal{Y}}$. The first relationship is from (10), and second relationship is from induction by t. Briefly, $t = 0$ holds because of the definition $\hat{\boldsymbol{M}}_k^{(0)}$ based on lines 4-5 of our initialization Algorithm 2. For $t \geq 1, \dots$, notice that $\hat{\boldsymbol{M}}_k^{(t)}$ is an optimizer of the objective

$$\|\bar{\mathcal{Y}} - \hat{\mathcal{C}}^{(t-1)} \times \{\boldsymbol{R}_1\hat{\boldsymbol{M}}_1^{(t)}, \dots, \boldsymbol{R}_{k-1}\hat{\boldsymbol{M}}_{k-1}^{(t)}, \boldsymbol{R}_k\boldsymbol{M}, \boldsymbol{R}_{k+1}\hat{\boldsymbol{M}}_{k+1}^{(t-1)}, \dots, \boldsymbol{R}_K\hat{\boldsymbol{M}}_K^{(t-1)}\}\|_F^2,$$

from the line 3 of Algorithm 1. By unfolding along the mode k, the optimizer $\boldsymbol{M}_k^{(t)}$ must satisfy

$$\begin{aligned}&\text{Unfold}_k\left(\bar{\mathcal{Y}} \times \left\{(\hat{\boldsymbol{M}}_1^{(t)})^T\boldsymbol{R}_1^{-1}, \dots, (\hat{\boldsymbol{M}}_{k-1}^{(t)})^T\boldsymbol{R}_{k-1}^{-1},\ \boldsymbol{I}_{p_k},\ (\hat{\boldsymbol{M}}_{k+1}^{(t-1)})^T\boldsymbol{R}_{k+1}^{-1}, \dots, (\hat{\boldsymbol{M}}_K^{(t-1)})^T\boldsymbol{R}_K^{-1}\right\}\right) \\ &\quad = \boldsymbol{R}_k\hat{\boldsymbol{M}}_k^{(t)}\text{Unfold}_k\left(\hat{\mathcal{C}}^{(t-1)}\right)\left(\boldsymbol{I}_{r_K} \otimes \cdots \otimes \boldsymbol{I}_{r_{k+1}} \otimes \boldsymbol{I}_{r_{k-1}} \otimes \boldsymbol{I}_{r_1}\right). \end{aligned} \tag{12}$$

Notice that the first r_k left singular vectors of the left side of (12) is $\hat{\boldsymbol{U}}_k^{(t)}$ in HOOI algorithm.

Therefore, we prove the the second relationship by induction.

Combination of Lemma A.4 and the relationships between our algorithm outputs and the ones in Zhang and Xia (2018) gives us

$$\left(\frac{\underline{\gamma}}{\bar{\gamma}}\right)^2 \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \leq \max_{k\in[K]} \sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(t)}) \leq \left(\frac{\bar{\gamma}}{\underline{\gamma}}\right)^2 \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}). \tag{13}$$

Now, we prove the property (a) in Proposition 4.1. Based on Lemma A.3(a), whenever $\lambda'/\sqrt{\phi} \geq C_{\text{gap}} p^{K/4}$, we have

$$\max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}) \leq c\left(\frac{p^{K/2}}{(\lambda\underline{\gamma})^2/\phi}\right), \tag{14}$$

with probability at least $1-\exp(-p)$. Notice that

$$\lambda' \overset{(11)}{\geq} \lambda\underline{\gamma} \gtrsim \lambda d^{K/2} \geq C_{\text{gap}}\sqrt{\phi} p^{K/4},$$

where the second inequality uses [Case 1] and [Case 2] in the proof of Theorem 4.1. The condition $\lambda/\sqrt{\phi} \geq C p^{K/4} d^{-K/2}$ guarantees a sufficiently large C_{gap} that satisfies $\lambda'/\sqrt{\phi} \geq C_{\text{gap}} p^{K/4}$. Thus combining (13) and (14) yields

$$\begin{aligned}\max_{k\in[K]} \sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(0)}) &\leq \left(\frac{\bar{\gamma}}{\underline{\gamma}}\right)^2 \left(\frac{\sqrt{\phi} p^{K/4}}{\lambda\underline{\gamma}}\right)^2 \\ &\leq \frac{1}{2},\end{aligned}$$

where the last inequality uses the fact that $\underline{\gamma} \asymp d^{K/2}$ and $\bar{\gamma}/\underline{\gamma}$ is bounded by a constant in [Case 1] and [Case 2], and the condition $\lambda/\sqrt{\phi} \geq C p^{K/4} d^{-K/2}$.

Now, we prove the property (b) in Proposition 4.1. Based on Lemma A.3(b), we have

$$\max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \lesssim \frac{\sqrt{p\phi}}{\lambda\underline{\gamma}} + \left(\frac{1}{2}\right)^t \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}),$$

with probability at least $1 - \exp(-p)$. Combining (13) with the above inequality yields

$$\begin{aligned}\max_{k\in[K]} \sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(t)}) &\lesssim \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \\ &\lesssim \frac{\sqrt{p\phi}}{\lambda\underline{\gamma}} + \left(\frac{1}{2}\right)^t \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}) \\ &\lesssim \frac{\sqrt{p\phi}}{\lambda\underline{\gamma}} + \left(\frac{1}{2}\right)^t \max_{k\in[K]} \sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(0)}).\end{aligned}$$

Finally, the proof is completed applying $\underline{\gamma} \asymp d^{K/2}$ from [Case 1] and [Case 2]. □

Proof of Theorem 4.2. Combining Proposition 4.1(b) and (14), we obtain

$$\max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \lesssim \frac{\sqrt{p\phi}}{\lambda\underline{\gamma}} + \left(\frac{1}{2}\right)^t \left(\frac{p^{K/2}}{(\lambda\underline{\gamma})^2/\phi}\right),$$

with probability at least $1 - \exp(-p)$. We set $t \gtrsim \log \frac{p^{(K-1)/2}}{\lambda\underline{\gamma}}$ to make the second term negligible. Therefore, the first part of proof is completed by noticing that

$$\frac{p^{(K-1)/2}}{\lambda\underline{\gamma}} \lesssim \log \frac{p^{(K-1)/2}}{\lambda d^{K/2}} \lesssim K \log p,$$

where the first inequality uses $\underline{\gamma} \asymp d^{K/2}$ from [Case 1] and [Case 2], and the last inequality is from the condition $\lambda/\sqrt{\phi} \geq Cp^{K/4}d^{-K/2}$.

For the estimation error with respect to Frobenius norm, direct application of Lemma A.3(c) with $t \gtrsim K \log p \gtrsim \log \frac{p^{(K-1)/2}}{\lambda\underline{\gamma}}$ yields

$$\|\hat{\mathcal{S}}^{(t)} - \mathcal{S}_{\text{true}}\|_F^2 \lesssim \phi(r^K + Kpr), \tag{15}$$

with probability at least $1 - \exp(-p)$. Notice that

$$\begin{aligned}\|\hat{\mathcal{S}}^{(t)} - \mathcal{S}_{\text{true}}\|_F^2 &= \|\left(\hat{\mathcal{B}}^{(t)} - \mathcal{B}_{\text{true}}\right) \times \{\boldsymbol{R}_1, \dots \boldsymbol{R}_K\}\|_F^2 \\ &\geq \underline{\gamma}^2 \|\hat{\mathcal{B}}^{(t)} - \mathcal{B}_{\text{true}}\|_F^2 \\ &\gtrsim d^K \|\hat{\mathcal{B}}^{(t)} - \mathcal{B}_{\text{true}}\|_F^2, \quad \text{from [Case 1] and [Case 2].}\end{aligned} \tag{16}$$

Combining (15) and (16) completes the proof. □

A.3 Auxiliary Lemmas

Proof of Proposition A.1. For ease of presentation, we drop the subscript $(i_1, \ldots, i_K)$ and simply write ε $(= y - b'(\theta))$. For any given $t \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}(\exp(t\varepsilon|\theta)) &= \int c(x) \exp\left(\frac{\theta x - b(\theta)}{\phi}\right) \exp\left(t(x - b'(\theta))\right) dx \\
&= \int c(x) \exp\left(\frac{(\theta + \phi t)x - b(\theta + \phi t) + b(\theta + \phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right) dx \\
&= \exp\left(\frac{b(\theta + \phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right) \\
&\leq \exp\left(\frac{\phi U t^2}{2}\right),
\end{aligned}$$

where $c(\cdot)$ and $b(\cdot)$ are known functions in the exponential family corresponding to y, and the last line uses the fact that $\sup_{\theta \in \mathbb{R}} b''(\theta) \leq U$. Therefore, ε is sub-Gaussian-(ϕU). □

Definition A.1 (α-convexity)**.** A real-valued function $f \colon \mathcal{S} \to \mathbb{R}$ is called α-convex, if

$$f(x_1) \geq f(x_2) + \langle \nabla_x f(x_2),\ x_1 - x_2 \rangle + \alpha \|x_1 - x_2\|_F^2, \text{ for all } x_1, x_2 \in \mathcal{S}.$$

Lemma A.1 (Convexity under linear transformation)**.** Suppose $f \colon \mathbb{R}^{d \times \cdots \times d} \to \mathbb{R}$ is a α-convex function. Define a function $g \colon \mathbb{R}^{p \times \cdots \times p} \to \mathbb{R}$ by $g(\mathcal{B}) = f(\mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\})$ for all $\mathcal{B} \in \mathbb{R}^{p \times \cdots \times p}$. Then, g is a $(\underline{\gamma}^2 \alpha)$-convex function.

Proof of Lemma A.1. By the definition of α-convexity, we have

$$f(\Theta_1) \geq f(\Theta_2) + \langle \nabla_\Theta f(\Theta_2), \Theta_1 - \Theta_2 \rangle + \alpha \|\Theta_1 - \Theta_2\|_F^2, \text{ for all } \Theta_1, \Theta_2 \in \mathbb{R}^{d \times \cdots \times d}, \quad (17)$$

where $\nabla_\Theta f(\cdot)$ denotes the derivative of f with respect to $\Theta \in \mathbb{R}^{d \times \cdots \times d}$. For any $\mathcal{B}_1, \mathcal{B}_2 \in \mathbb{R}^{p \times \cdots \times p}$, we notice that $\mathcal{B}_i \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \in \mathbb{R}^{d \times \cdots \times d}$ for $i = 1, 2$. Applying (17) to this setting gives

$$f(\mathcal{B}_1 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\})$$

$$
\begin{aligned}
&\geq f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) + \langle \nabla_\Theta f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}), (\mathcal{B}_1 - \mathcal{B}_2) \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \rangle \\
&\quad + \alpha \|(\mathcal{B}_1 - \mathcal{B}_2) \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}\|_F^2 \\
&\geq f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) + \langle \nabla_\Theta f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) \times \{\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T\}, (\mathcal{B}_1 - \mathcal{B}_2) \rangle \\
&\quad + \alpha \underline{\gamma}^2 \|\mathcal{B}_1 - \mathcal{B}_2\|_F^2. \qquad (18)
\end{aligned}
$$

By the definition of g and the linearity from $\mathcal{B}$ to Θ, we have

$$
\nabla g_{\mathcal{B}}(\mathcal{B}_2) = \nabla f_\Theta(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) \times \{\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T\}. \qquad (19)
$$

The convexity of g directly follows by plugging (19) into (18),

$$
g(\mathcal{B}_1) \geq g(\mathcal{B}_2) + \langle \nabla g_{\mathcal{B}}(\mathcal{B}_2), \mathcal{B}_1 - \mathcal{B}_2 \rangle + \alpha \underline{\gamma}^2 \|\mathcal{B}_1 - \mathcal{B}_2\|_F^2.
$$

□

Proof of Proposition A.2. We first prove the strong concavity by viewing the log-likelihood as a function of the linear predictor Θ. Write

$$
\bar{\mathcal{L}}(\Theta) = \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1, \ldots, i_K} b(\theta_{i_1, \ldots, i_K}).
$$

Direct calculation shows that the Hession of $\bar{\mathcal{L}}(\Theta)$ can be expressed as

$$
\frac{\partial \bar{\mathcal{L}}(\Theta)}{\partial \theta_{i_1, \ldots, i_K} \partial \theta_{j_1, \ldots, j_K}} = \begin{cases} -b''(\theta_{i_1, \ldots, i_K}) < -L < 0, & \text{if } (i_1, \ldots, i_K) = (j_1, \ldots, j_K), \\ 0, & \text{otherwise,} \end{cases}
$$

Therefore, the Hession matrix of $\bar{\mathcal{L}}(\Theta)$ is strictly negative definite with eigenvalues upper bounded by $-L < 0$. By Taylor expansion, $-\bar{\mathcal{L}}(\Theta)$ is $L/2$-convex with respect to Θ. Note that $\bar{\mathcal{L}}(\Theta) = \mathcal{L}(\mathcal{B})$ via the linear mapping $\Theta = \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$. Therefore, by Lemma A.1, $\mathcal{L}(\mathcal{B})$ is $(\gamma^2 L/2)$-convex with respect to $\mathcal{B}$.

To prove the second part of Proposition A.2, we note

$$\langle \nabla \mathcal{L}(\mathcal{B}_{\text{true}}), \mathcal{B} \rangle = \langle \nabla \bar{\mathcal{L}}(\Theta_{\text{true}}) \times \{\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T\},\ \mathcal{B} \rangle = \langle \mathcal{Y} - b'(\Theta_{\text{true}}),\ \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \rangle.$$

By Proposition A.1, $\mathcal{Y} - b'(\Theta_{\text{true}})$ is a random tensor consisting of i.i.d. sub-Gaussian-$(U\phi)$ entries under Assumption 2. We write $\mathcal{E} = \mathcal{Y} - b'(\Theta_{\text{true}})$ and consider the sub-Gaussian maxima

$$\text{Err}_{\text{ideal}}(\boldsymbol{r}) = \sup_{\|\mathcal{B}\|_F = 1, \mathcal{B} \in \mathcal{P}(r)} \langle \mathcal{E}, \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \rangle.$$

The quantity $\text{Err}_{\text{ideal}}(\boldsymbol{r})$ is closely related to the localized Gaussian width (Chen et al., 2019; Han et al., 2020) that measures the model complexity of $\mathcal{P}(\boldsymbol{r})$. By adapting Han et al. (2020, Lemma E.5) in our context, we have

$$\text{Err}_{\text{ideal}}(\boldsymbol{r}) \lesssim \sqrt{\phi U (r^K + Kpr)} \prod_{k \in [K]} \sigma_{\max}(\boldsymbol{X}_k) \leq \bar{\gamma} \sqrt{\phi U (r^K + Kpr)},$$

with probability at least $1 - \exp(-p)$. □

The following Lemma is adopted from Wang and Song (2017, Theorem 6.1) in our contexts.

Lemma A.2 (Wedin's $\sin\Theta$ Theorem)**.** Let $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$ be two $m \times n$ real matrix SVDs $\boldsymbol{B} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ and $\hat{\boldsymbol{B}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^T$. If $\sigma_{\min}(\boldsymbol{B}) > 0$ and $\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_F \ll \sigma_{\min}(\boldsymbol{B})$, then

$$\sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}) \leq \frac{\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B})}{\sigma_{\min}(\boldsymbol{B})} \leq \frac{\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_F}{\sigma_{\min}(\boldsymbol{B})}.$$

The following theorem Zhang and Xia (2018) provides the statistical guarantees for unsupervised tensor decomposition based on alternating least square algorithm. For simplicity, we consider the balanced dimension $p_1 = \cdots = p_K = p$ and $r_1 = \cdots = r_K = r$.

Lemma A.3 (Theorem 1 in Zhang and Xia (2018))**.** Consider the Gaussian tensor model

$$\mathcal{Y} = \mathcal{S}_{\text{true}} + \mathcal{E},$$

where $\mathcal{S}_{\text{true}} = \mathcal{C}_{\text{true}} \times \{\boldsymbol{U}_{1,\text{true}}, \ldots, \boldsymbol{U}_{K,\text{true}}\}$ is an unknown signal tensor, $\mathcal{C}_{\text{true}} \in \mathbb{R}^{r \times \cdots \times r}$ is a full rank core tensor, $\boldsymbol{U}_{k,\text{true}} \in \mathbb{O}(p, r)$ are orthornomal matrices, and $\mathcal{E} \in \mathbb{R}^{p \times \cdots \times p}$ is a Gaussian noise tensor consisting of i.i.d entries from $N(0, \sigma)$. Let λ denote the smallest singular value of matrices $\text{Unfold}_k(\mathcal{S}_{\text{true}})$ over all possible k,

$$\lambda' = \min_{k \in [K]} \sigma_{\min}(\text{Unfold}_k(\mathcal{S}_{\text{true}})).$$

Then, the following two properties hold whenever $\lambda'/\sigma \geq C_{\text{gap}} p^{K/4}$ for some universal constant $C_{\text{gap}} > 0$.

(a) With probability at least $1 - \exp(-p)$, the spectral initialization $\hat{\boldsymbol{U}}_k^{(0)}$ has

$$\max_{k \in [K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}) \leq c \frac{p^{K/2}}{\lambda'^2/\sigma^2},$$

for some constant $c > 0$.

(b) Let $t = 1, 2, \ldots,$ denote the iteration in HOOI algorithm. With probability at least $1 - \exp(-p)$, the alternating optimization $\hat{\boldsymbol{U}}_k^{(t)}$ satisfies

$$\max_{k \in [K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \lesssim \frac{\sqrt{p}}{\lambda'/\sigma} + \left(\frac{1}{2}\right)^t \max_{k \in [K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}),$$

(c) When $t \gtrsim \log \frac{p^{(K-1)/2}}{\lambda'}$, the tensor estimate $\hat{\mathcal{S}}^{(t)}$ from HOOI satisfies

$$\|\hat{\mathcal{S}}^{(t)} - \mathcal{S}_{\text{true}}\|_F^2 \lesssim \sigma^2(r^K + Kpr),$$

with probability at least $1 - \exp(-p)$.

Lemma A.4 (Angle distance under linear transformation)**.** Let $\boldsymbol{U}$ and $\hat{\boldsymbol{U}}$ be two $m \times n$ real matrices where $m > n$. Let $\boldsymbol{R}$ be an $m \times m$ invertible matrix. If $\sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}) \leq L$ for some constant $L \in [0, 1]$, then

$$\sin\Theta(\boldsymbol{R}\boldsymbol{U}, \boldsymbol{R}\hat{\boldsymbol{U}}) \leq \left(\frac{\sigma_{\max}(\boldsymbol{R})}{\sigma_{\min}(\boldsymbol{R})}\right)^2 L.$$

Proof. Suppose that orthonormal basis of $\mathrm{Span}(\boldsymbol{U})$ and $\mathrm{Span}(\hat{\boldsymbol{U}}^{\perp})$ are $\{\mu_1, \ldots, \mu_n\}$ and $\{\nu_{n+1}, \ldots, \nu_m\}$ respectively. By definition,

$$\sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}) = \max_{\sum_{i=1}^n a_i^2 = \sum_{j=n+1}^m b_j^2 = 1} \left\langle \sum_{i=1}^n a_i \mu_i, \sum_{j=n+1}^m b_j \nu_j \right\rangle \leq L.$$

We write $\boldsymbol{x} = \boldsymbol{R}\sum_{i=1}^n a_i\mu_i$ and $\boldsymbol{y} = \boldsymbol{R}\sum_{j=n+1}^m b_j\nu_j$ for any $\boldsymbol{x} \in \mathrm{Span}(\boldsymbol{RU})$ and $\boldsymbol{y} \in \mathrm{Span}((\boldsymbol{R}\hat{\boldsymbol{U}})^{\perp})$. Then,

$$\begin{aligned}
\frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2} &= \frac{\langle \boldsymbol{R}\sum_{i=1}^n a_i\mu_i, \boldsymbol{R}\sum_{j=n+1}^m b_j\nu_j\rangle}{\|\boldsymbol{R}\sum_{i=1}^n a_i\mu_i\|_2\|\boldsymbol{R}\sum_{j=n+1}^m b_j\nu_j\|_2} \\
&\leq \frac{\sigma_{\max}(\boldsymbol{R}^T\boldsymbol{R})\langle\sum_{i=1}^n a_i\mu_i, \sum_{j=n+1}^m b_j\nu_j\rangle}{\sigma_{\min}^2(\boldsymbol{R})\sqrt{\sum_{i=1}^n a_i^2}\sqrt{\sum_{j=n+1}^m b_j^2}} \\
&\leq \left(\frac{\sigma_{\max}(\boldsymbol{R})}{\sigma_{\min}(\boldsymbol{R})}\right)^2 \sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}).
\end{aligned}$$

□

B Additional simulation results

B.1 Detailed simulation setup for Figure 6a-b

We generate data from **Envelope** model (Li and Zhang, 2017) with slight modification. We simulate response tensor $\mathcal{Y} \in \mathbb{R}^{d\times d\times d}$ from the following model with envelope dimension (u_1, u_2),

$$\begin{aligned}
&\mathcal{Y}|\boldsymbol{X} = \mathcal{B}\times_3 \boldsymbol{X} + \mathcal{E} = \mathcal{C}\times\{\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{X}\} + \mathcal{E}, \\
&\text{with} \quad \mathcal{E} \sim \mathcal{TN}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{I}), \quad \boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Omega}_k\boldsymbol{\Gamma}_k^T + \boldsymbol{\Gamma}_{0k}\boldsymbol{\Omega}_{0k}\boldsymbol{\Gamma}_{0k}^T + \boldsymbol{I}, \quad k = 1, 2,
\end{aligned} \tag{20}$$

where $\boldsymbol{X} \in \mathbb{R}^{d\times p}$ is the feature matrix, $\mathcal{B} = \mathcal{C}\times\{\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{I}\} \in \mathbb{R}^{d\times d\times p}$ is the coefficient tensor, $\mathcal{C} \in \mathbb{R}^{\mu_1\times\mu_2\times p}$ is a full-rank core tensor, $\mathcal{TN}(\cdot,\ \cdot,\ \cdot)$ represents zero-mean tensor normal distribution with Kronecker structured covariance, $\boldsymbol{\Gamma}_k \in \mathbb{O}(d, u_k)$ consists of orthogonal columns, $\boldsymbol{\Gamma}_{0k} \in \mathbb{O}(d, d-u_k)$ is the orthogonal complement of $\boldsymbol{\Gamma}_k$, and $\boldsymbol{\Omega}_k = \boldsymbol{A}_k\boldsymbol{A}_k^T$, $\boldsymbol{\Omega}_{0k} = \boldsymbol{A}_{k0}\boldsymbol{A}_{k0}^T$ with $\boldsymbol{A}_k \in \mathbb{R}^{u_k\times u_k}$, $\boldsymbol{A}_{k0} \in \mathbb{R}^{(d-u_k)\times(d-u_k)}$.

The entries of $\boldsymbol{X}$ are i.i.d. drawn from $\mathcal{N}(0,1)$, the entries of $\boldsymbol{A}_k$, $\boldsymbol{A}_{k0}$ are i.i.d. drawn from Uniform$[-\gamma,\gamma]$, and the entries of core tensor $\mathcal{C}$ are i.i.d. drawn from Uniform$[-3,3]$. We call γ the *correlation level*. Note that the only distinction between model (20) and standard **Envelope** model is the additional identity matrix $\boldsymbol{I}$ in the expression of $\boldsymbol{\Sigma}_k$. When $\gamma = 0$, the model (20) reduces to our **STD** model with rank $\boldsymbol{r} = (u_1, u_2, p)$. We set $d = 20, p = 5$ in our simulation.

B.2 Detailed simulation setup for Figure 6c-d

We generate the data from **GLSNet** model (Zhang et al., 2018) with slight modification. We simulate the binary response tensor $\mathcal{Y} \in \{0,1\}^{d\times d\times d}$ from the following model

$$\mathbb{E}[\mathcal{Y}|\boldsymbol{X}] = f(\mathbf{1}\otimes\boldsymbol{\Theta} + \mathcal{B}\times_3 \boldsymbol{X}),$$

where $f(\cdot)$ is the logistic link, $\boldsymbol{X} \in \mathbb{O}(d,p)$ is the feature matrix with orthonormal columns, $\boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{A}^T \in \mathbb{R}^{d\times d}$ is a rank-R intercept matrix, where the entries of $\boldsymbol{A} \in \mathbb{R}^{d\times R}$ are simulated from i.i.d. standard normal. Unlike original **GLSNet** model, we generate joint sparse and low-rank structure to the coefficient tensor $\mathcal{B}$ as follows.

To generate $\mathcal{B}$, we firstly generate a low-rank tensor $\mathcal{B}_0$ as

$$\mathcal{B}_0 = \mathcal{C}\times\boldsymbol{M}_1\times\boldsymbol{M}_2\times\boldsymbol{M}_3,$$

where $\mathcal{C} \in \mathbb{R}^{R\times R\times R}$ is a full-rank core tensor, $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{d\times R}$ and $\boldsymbol{M}_3 \in \mathbb{R}^{p\times R}$ are the factor matrices with orthonormal columns. We simulate i.i.d. uniform entries in $\mathcal{C}$ and rescale the tensor $\mathcal{B}_0$ such that $\|\mathcal{B}_0\|_{\max} = 2$. Last, we obtain a sparse $\mathcal{B}$ by randomly setting sd^2p entries in $\mathcal{B}_0$ to zero. We call s the *sparsity level* which quantifies the proportion of zero's in $\mathcal{B}$. Hence, the generated tensor $\mathcal{B}$ is of sparsity level s and of low-rank (R,R,R). We set $d = 20$, $p = 5$ and consider the combination of rank $R = 2$ (low), 4 (high) and sparsity level $s = \{0, 0.3, 0.5\}$ in the simulation.

B.3 Comparison with GLMs under stochastic block models

We investigate the performance of our model under correlated feature effects. We mimic the scenario of brain imaging analysis. A sample of $d_3 = 50$ networks are simulated, one for each individual. Each network measures the connections between $d_1 = d_2 = 20$ brain nodes. We simulate $p = 5$ features for the each of the 50 individuals. These features may represent, for example, age, gender, cognitive score, etc. Recent study has suggested that brain connectivity networks often exhibit community structure represented as a collection of subnetworks, and each subnetwork is comprised of a set of spatially distributed brain nodes. To accommodate this structure, we utilize the stochastic block model (Abbe, 2017) to generate the effect size. Specifically, we partition the nodes into r blocks by assigning each node to a block with uniform probability. Edges within a same block are assumed to share the same feature effects, where the effects are i.i.d. drawn from $N(0, 1)$. We then apply our tensor regression model to the network data using the BIC-selected rank. Note that in this case, the true model rank is unknown; the rank of a r-block network is not necessarily equal to matrix rank r (Wang and Zeng, 2019).

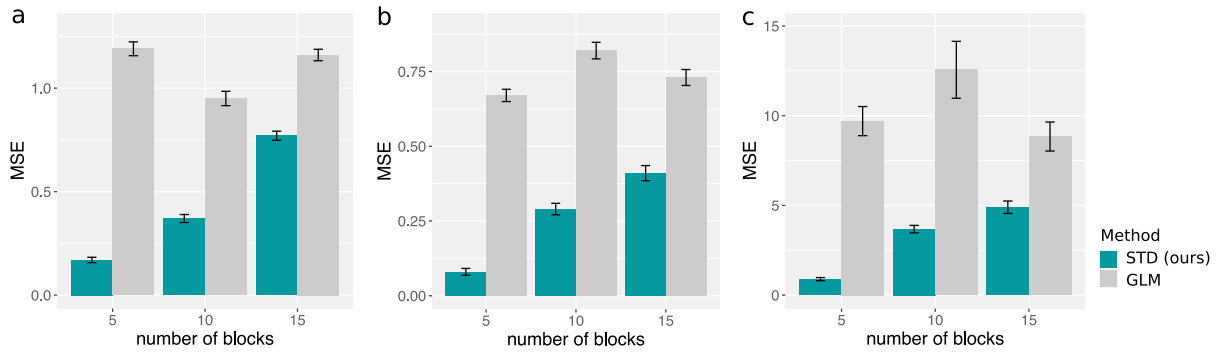

Figure S1: Performance comparison under stochastic block models. The three panels plot the MSE when the response tensors are generated from (a) Gaussian (b) Poisson and (c) Bernoulli models. The x-axis represents the number of blocks in the networks.

Figure S1 compares the MSE of our method with a multiple-response GLM approach. The multiple-response GLM is to regress the dyadic edges, one at a time, on the features, and this model is repeatedly fitted for each edge. As we find in Figure S1, our tensor regression method achieves significant error reduction in all three data types considered.

The outperformance is substantial in the presence of large communities; even in the less structured case ($\sim 20/15 = 1.33$ nodes per block), our method still outer-performs GLM. The possible reason is that the multiple-response GLM approach does not account for the correlation among the edges, and suffers from overfitting. In contrast, the low-rankness in our modeling incorporates the shared information across entries. By selecting the rank in a data-driven way, our method achieves accurate estimation in a wide range of settings.

C Additional results on data application

C.1 Rank selection for Nations data

Table S1 summarizes the BIC results in the grid search $\boldsymbol{r} \in \{3, 4, 5\}^3$. We set $r_1 = r_2$ due to the symmetry in the dataset. Table S1 shows that $(r_1, r_2) = (4, 4)$ consistently provides the minimal BIC under a range of r_3. Because multiple values of r_3 give similar BIC, we choose r_3 based on the interpretability of the results. Tables S2-S4 compare the clustering results for $r_3 = 3, 4, 5$. For ease of visualisation, we list only the subset of relations for which the three configurations yield incoherent clustering. We find that the clustering with $r_3 = 4$ (Table S3) provides the cleanest results. Table S2 with $r_3 = 3$ mixes the categories Economics with Organization and Military. Table S4 with $r_3 = 5$ mixes Economics with Organization, while splitting Military and Territory into different clusters. Therefore, we choose the rank $\boldsymbol{r} = (4, 4, 4)$ in the main paper. The running time for the rank selection via grid search is 95 secs in total, on an iMac macOS High Sierra 10.13.6 with Intel Core i5 3.8 GHz CPU and 8 GB RAM. This indicates the BIC is feasible in the considered setting.

r_3	$r_3 = 3$			$r_3 = 4$			$r_3 = 5$		
(r_1, r_2)	$(3,3)$	$(4,4)$	$(5,5)$	$(3,3)$	$(4,4)$	$(5,5)$	$(3,3)$	$(4,4)$	$(5,5)$
BIC	11364	**11194**	11701	12275	**11897**	12365	17652	**12666**	18146

Table S1: BIC results for *Nations* data under different tensor rank. Bold number indicates the minimal BIC with a certain r_3.

Cluster	Relations
I	exportbooks, relexportbooks, protests, tourism, reltourism, relintergovorgs relngo, intergovorgs3, ngoorgs3, militaryalliance,commonbloc1
II	militaryactions, severdiplomatic,expeldiplomats, commonbloc0, aidenemy attackembassy, lostterritory, blockpositionindex
III	tourism3, exports, relexports, exports3, intergovorgs, ngo ,embassy, reldiplomacy, commonbloc2

Economics Military Organization Territory

Table S2: K-mean relations clustering with $r_3 = 3$. For visualization purpose, only a subset of relations are presented. See texts for details.

Cluster	Relations
I	aidenemy, attackembassy, lostterritory
II	militaryactions, severdiplomatic, expeldiplomats, protests, commonbloc0, blockpositionindex, commonbloc1
III	relintergovorgs, relngo, intergovorgs3, ngoorgs3, militaryalliance, commonbloc2
IV	exportbooks, relexportbooks, tourism, reltourism, tourism3 exports, relexports, exports3, intergovorgs, ngo, embassy, reldiplomacy

Economics Military Organization Territory

Table S3: K-mean relations clustering with $r_3 = 4$. For visualization purpose, only a subset of relations are presented. See texts for details.

Cluster	Relations
I	exportbooks, relexportbooks, tourism, reltourism, tourism3, exports, relexports, exports3 intergovorgs, relintergovorgs, ngo, relngo, intergovorgs3, ngoorgs3, embassy, reldiplomacy
II	attackembassy
III	commonbloc0, blockpositionindex
IV	militaryalliance, commonbloc2
V	militaryactions, severdiplomatic, expeldiplomats, aidenemy, lostterritory, protests, commonbloc1

Economics Military Organization Territory

Table S4: K-mean relations clustering with $r_3 = 5$. For visualization purpose, only a subset of relations are presented. See texts for details.

C.2 Comparison with unsupervised decomposition

We compare the supervised vs. unsupervised decomposition in the *Nations* data analysis. Table S5 shows the clustering results based on classical unsupervised Tucker decomposition without the feature matrices. Table S6 shows the clustering results based on supervised tensor decomposition (**STD**). Compared with supervised decomposition, the unsupervised

clustering loses some interpretation. Similar relations *exports* and *relexports*, *ngo* and *relngo* are separated into different clusters.

Cluster	Relations
I	economicaid, releconomicaid, exportbooks, relexportbooks, weightedunvote, unweightedunvote, tourism, reltourism, tourism3, exports, intergovorgs, ngo, militaryalliance
II	warning, violentactions, militaryactions, duration, severdiplomatic, expeldiplomats, boycottembargo, aidenemy, negativecomm, accusation, protests, unoffialacts, attackembassy, relemigrants, timesincewar, lostterritory, dependent
III	timesinceally, independence, commonbloc0, blockpositionindex
IV	treaties, reltreaties, officialvisits, conferences, booktranslations, relbooktranslations negativebehavior, nonviolentbehavior, emigrants, emigrants3, students, relstudents, relexports, exports3 relintergovorgs, relngo, intergovorgs3, ngoorgs3, embassy, reldiplomacy, commonbloc1, commonbloc2

Economics Military Organization Territory

Table S5: Clustering of relations based on unsupervised tensor decomposition.

Category	Relations
I	warning, violentactions, militaryactions, duration, negativebehavior, protests, severdiplomatic timesincewar, commonbloc0, commonbloc1, blockpositionindex, expeldiplomats
II	emigrants, emigrants3, relemigrants, accusation, nonviolentbehavior, ngoorgs3, commonbloc2, intergovorgs3 releconomicaid, relintergovorgs, relngo, students, relstudents, economicaid, negativecomm, militaryalliance
III	treaties, reltreaties, officialvisits, exportbooks, relexportbooks, booktranslations, relbooktranslations boycottembargo, weightedunvote, unweightedunvote, reltourism, tourism, tourism3, exports, exports3 relexports, intergovorgs, ngo, embassy, reldiplomacy, timesinceally, independence, conferences, dependent
IV	aidenemy, lostterritory, unoffialacts, attackembassy

Economics Military Organization Territory

Table S6: Clustering of relations based on supervised tensor decomposition.

C.3 How different are supervised vs. unsupervised factors in general?

It is helpful to realize that the unsupervised and methods address different aspects of the problem. The unsupervised decomposition identifies factors that explain most variation in the tensor, whereas the supervised decomposition identifies factors that are most attributable to side features.

We provide a simple example here for illustration.

Example C.1. Consider the following data tensor $\mathcal{Y}$ and one-sided feature matrix $\boldsymbol{X}$,

$$\mathcal{Y} = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + 10\boldsymbol{e}_2 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_2, \quad \boldsymbol{X} = \boldsymbol{e}_1,$$

where $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is the ith canonical basis vector in $\mathbb{R}^d$ for $i = 1, 2$. Now, consider the unsupervised vs. supervised decomposition of $\mathcal{Y}$ with rank $\boldsymbol{r} = (1, 1, 1)$. Then,

the top supervised and unsupervised factors are perpendicular to each other,

$$\boldsymbol{M}_{\text{sup},k} \perp \boldsymbol{M}_{\text{unsup},k}, \quad \text{for all } k = 1, 2, 3,$$

where $\boldsymbol{M}_{\text{sup},k}$, $\boldsymbol{M}_{\text{unsup},k}$ denote the mode-k factors from supervised and unsupervised decompositions, respectively.

Remark C.1. This example shows complementary information between factors from supervised vs. unsupervised decompositions. In general, one could construct examples such that these two methods return **arbitrarily different** factors.

References

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

Chen, H., Raskutti, G., and Yuan, M. (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208.

Han, R., Willett, R., and Zhang, A. (2020). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics, In press. arXiv preprint arXiv:2002.11255*.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146.

Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific.

Wang, M. and Song, Y. (2017). Tensor decompositions via two-mode higher-order SVD (HOSVD). In *Artificial Intelligence and Statistics*, pages 614–622.

Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019). arXiv:1906.03807*.

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

Zhang, J., Sun, W. W., and Li, L. (2018). Network response regression for modeling population of networks with covariates. *arXiv preprint arXiv:1810.03192*.