Supervised Tensor Decomposition with Interactive Side Information

September 7, 2020

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 side information as interactive features. Unlike unsupervised tensor decomposition, our supervised decomposition captures the effective dimension reduction of the data tensor confined to feature space on each mode. An efficient alternating optimization algorithm 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 social network data. We identify the key global connectivity pattern and pinpoints the local regions that are associated with available features. Our method will help the practitioners efficiently analyze tensor datasets in various areas. Toward this end, the package and data used are available at https://CRAN.R-project.org/package=tensorregress.

Keywords: Applications and case studies, Tensor data analysis, Supervised dimension reduction, Exponential family distribution, Generalized multilinear model

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 (Sun and Li, 2017; Zhou et al., 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 1a). Another example is in the field of network analysis (Baldin and Berthet, 2018; Hoff, 2005). A typical social network consists of nodes that represent people and edges that represent the friendships. Side information such as people's demographic information and friendship types are often available. In both examples, it is of keen scientific interest to identify the variation in the tensor data (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 instances of tensor datasets consist of non-Gaussian measurements. Examples include the political interaction dataset (Hu et al., 2015) which measures action counts between countries under various relationships, and the brain connectivity network dataset (Wang et al., 2019) which is a collection of binary adjacency matrices. Classical tensor decomposition methods are based on minimizing the Frobenious norm of the reconstruction error, leading to suboptimal predictions for binary- or count-valued response variables. A number of supervised tensor methods have been proposed (Narita et al., 2012; Zhao et al., 2012; Yu and Liu, 2016) to address the tensor regression problem in various forms (e.g. scalar-to-tensor regression, 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 appli-

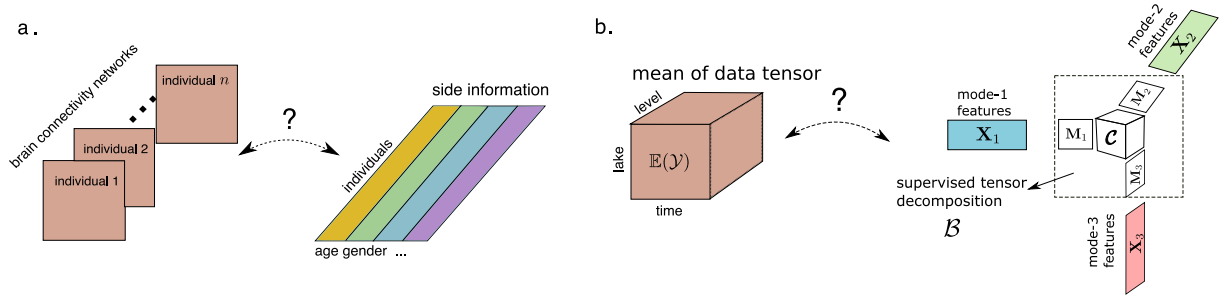

Figure 1: Examples of supervised tensor decomposition with interactive side information. (a) Network population model. (b) Spatial-temporal growth model.

cations of our interest. The gap in theory and practice suggests a great opportunity for improved modeling paradigms that better capture the complexity in tensor data with side information.

We present a general model and associated method for decomposing a data tensor with exponential family entries and interactive 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 interactive features. Figure 1b illustrates our model in the special case of order-3 tensors. A low-rank structure is imposed to the conditional mean tensor, 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.

In addition, we use tools from generalized linear model (GLM) to allow heteroscedacity due to the mean-variance relationship in the non-Gaussian data. This flexibility is important in practice. In classical GLM, the sample size and feature dimension are well defined;

however, in the tensor data analysis, we observe only one single realization of tensor response but up to K interactive feature matrices, where K is the number of tensor modes. Both the sample size and feature dimension grow exponentially in K. We use a low-rank constraint in the conditional mean tensor (with a suitable transformation) to mitigate the curse of high dimensionality. The statistical convergence of our estimator is established, and we quantify the gain in prediction through several 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 Tucker and CP decomposition (De Lathauwer et al., 2000; Kolda and Bader, 2009; Zhang and Xia, 2018; Hong et al., 2020) that aims to find the best low-rank representation of a data tensor. 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.

The second line of work studies the scalar-to-tensor regression, in which the response is a scalar and the predictor is a tensor (Zhou et al., 2013; Chen et al., 2019). Our proposal is orthogonal to this line because we treat the data tensor as random variables. The difference in model settings yields different focuses and interpretations. Consider the neuroimaging analysis as an example. The scalar-to-tensor regression aims to predict the individual response based on brain connectivity network, while tensor-to-scalar model focuses on understanding the variation in the brain connectivity attributable to the individual features such as age, gender, and disease status. The scalar-to-tensor model is insufficient for tensor dimension reduction because it lacks the modeling of stochastic noise in the data tensor.

The third line of work studies the tensor-on-tensor regression, in which both the response and predictor are tensors (Raskutti et al., 2015; Lock, 2018; Gahrooei et al., 2020).

Our model shares a common ground but provides a more efficient context-specific solutions than earlier approaches. As we show in Section 4.4, the supervised tensor decomposition has an interesting interpretation as a special tensor-on-tensor regression. Nevertheless, our work improves from previous work in scope and applicability. Previous methods (Gahrooei et al., 2020; Lock, 2018) mainly focus on Gaussian tensors. The Frobenius norm used in the objective function is statistical suboptimal for general exponential family tensors. Maximum likelihood estimator (MLE) is studied in Raskutti et al. (2015) and a convex relaxation algorithm is proposed to solve for low-rank tensor coefficients. However, in the tensor case, convex MLE suffers from both computational intractability and statistical sub-optimality. We advocate a non-convex approach and provide strong evidence for its success in our setting. Most previous tensor regression focuses on prediction (Lock, 2018; Raskutti et al., 2015; Gahrooei et al., 2020), and we go step further by finding the sufficient dimension reduction (Adragni and Cook, 2009), Span($\boldsymbol{M}_k$), that facilitates the identification of effective features in prediction (see Figure 1b). The latter approach greatly improves the *interpretability* in prediction. In this regards, our method opens up new opportunities for tensor data analysis in a wider range of possible applications.

[FIXME (Miaoyan): One more paragraph. Current literature on tensor decomposition with auxiliary side information...]

The remainder of this paper is organized as follows. Section 2 introduces tensor notation and preliminaries. Section 3 presents the main model and three motivating examples for supervised tensor decomposition. Section 4 describes a rank-constrained likelihood-based estimation and associated alternating optimization algorithm. In Section 5, we present numerical experiments and assess the performance in comparison to alternative methods. In Section 6, we apply the method to diffusion tensor imaging data from human connectome

project and multi-relational social network data. We conclude in Section 7 with discussions about our findings and avenues of future work.

2 Preliminaries

We introduce the basic tensor properties used in the paper; more details on tensor notation can be found in Kolda and Bader (2009). We use $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ **[FIXME (Miaoyan): perhaps use** y_ω**]** to denote an order-K $(d_1, \ldots, d_K)$-dimensional tensor, where K corresponds to the number of modes of $\mathcal{Y}$ and is 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

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

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 Frobenius norm and maximum norm of $\mathcal{Y}$ are defined as

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

A higher-order tensor can be reshaped into a lower-order object. We use $\text{vec}(\cdot)$ to denote the operation that reshapes the tensor into a vector, and $\text{Unfold}_k(\cdot)$ the operation that

reshapes the tensor along mode-k into a matrix of size d_k-by-$\prod_{i\neq k} d_i$. The multilinear rank of an order-K tensor $\mathcal{Y}$ is defined as a length-K vector $\boldsymbol{r} = (r_1, \ldots, r_K)$, where r_k is the rank of matrix $\text{Unfold}_k(\mathcal{Y})$, $k = 1, \ldots, K$. Given a matrix $\boldsymbol{M} \in \mathbb{R}^{d\times r}$, we use $\text{Span}(\boldsymbol{M})$ to denote the space spanned by columns of $\boldsymbol{M}$. We use lower-case letters (e.g., a, b, c) for scalars and vectors, upper-case boldface letters (e.g., $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$) for matrices, and calligraphy letters (e.g., $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$) for tensors of order three or greater. We let $\boldsymbol{I}_d$ denote the $d \times d$ identity matrix and $[d]$ denote the d-set $\{1, \ldots, d\}$. 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.

3 Motivation and model

3.1 General framework for supervised 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 $(i, j) \in [d_k] \times [p_k]$, $p_k \leq d_k$. 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, and the conditional mean tensor admits the form

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) = f(\Theta), \quad \text{with} \quad \Theta = \mathcal{B} \times \{\boldsymbol{X}_1, \ \ldots, \ \boldsymbol{X}_K\}, \tag{1}$$

where $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the multilinear predictor, $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the unknown parameter tensor, $f(\cdot)$ 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 $\exp(\cdot)$ 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 omit the dispersion parameter in this section for ease of presentation.

In classical tensor decomposition, tensor factorization is performed on either data tensor $\mathcal{Y}$ or mean tensor $\mathbb{E}(\mathcal{Y})$. In the context of supervised tensor decomposition, we propose to factorize the latent parameter tensor $\mathcal{B}$,

$$\mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \ \ldots, \ \boldsymbol{M}_K\}, \tag{2}$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is a full-rank core tensor, and $\boldsymbol{M}_k \in \mathbb{R}^{p_k \times r_k}$ are factor matrices consisting of orthonormal columns, where $r_k \leq p_k$ for all $k \in [K]$. By the definition of multilinear rank, model equations (1) and (2) imply the low-rankness $\boldsymbol{r} = (r_1, \ldots, r_K)$ of the conditional mean tensor under the link function. We now reach our final model for supervised tensor decomposition,

$$\begin{aligned} \mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) &= f\left(\mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1, \ \ldots, \ \boldsymbol{X}_K\boldsymbol{M}_K\}\right), \\ &\text{where} \ \ \boldsymbol{M}_k\boldsymbol{M}_k^T = \boldsymbol{I}_{p_k}, \ \boldsymbol{M}_k \in \mathbb{R}^{p_k \times r_k} \text{ for all } k = 1, \ldots, K, \end{aligned} \tag{3}$$

where the parameters of interest are $\boldsymbol{M}_k$ and $\mathcal{C}$. Note that model (3) assumes a fixed, known rank $\boldsymbol{r} = (r_1, \ldots, r_K)$; the adaptation to unknown rank will be described in Sec-

tion 4.3. Figure 1b provides a schematic illustration of our model. The features $\boldsymbol{X}_k$ affect the distribution of tensor entries in $\mathcal{Y}$ through the form $\boldsymbol{X}_k\boldsymbol{M}_k$, which are r_k linear combinations of features on mode k. We call $\boldsymbol{X}_k\boldsymbol{M}_k$ the "supervised tensor factors" or "sufficient features" (Adragni and Cook, 2009), and call $\boldsymbol{M}_k$ the "dimension reduction matrix." The core tensor $\mathcal{C}$ collects the interaction effects between sufficient features across K modes. Our goal is to find $\boldsymbol{M}_k$ and the corresponding $\mathcal{C}$. Note that $\boldsymbol{M}_k$ and $\mathcal{C}$ are identifiable only up to orthonormal transformations.

3.2 Three examples

We give three concrete examples of supervised tensor decomposition model (3) that arises in practice.

Example 1 (Spatio-temporal growth model)**.** The growth curve model (Gabriel, 1998; Potthoff and Roy, 1964; Srivastava et al., 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 model can be represented as

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) = \mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1,\ \boldsymbol{X}_2\boldsymbol{M}_2,\ \boldsymbol{X}_3\boldsymbol{M}_3\}, \tag{4}$$

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 of coherent sizes. The factors $\boldsymbol{X}_k\boldsymbol{M}_k$ are sufficient features for the mean model (4). The spatial-temporal model is a special case of our supervised tensor decomposition model, with features available on each of the three modes.

Example 2 (Network population model)**.** Network response model (Rabusseau and Kadri, 2016; Zhang et al., 2018) is recently developed for neuroimanig 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 symmetric 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

$$\text{logit}(\mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{x}_i)) = \mathcal{B} \times_3 \boldsymbol{x}_i, \quad \text{for } i = 1, \ldots, n, \tag{5}$$

where $\mathcal{B} \in \mathbb{R}^{d\times d\times p}$ is a coefficient tensor of rank $r(\mathcal{B}) = (r, r, r')$.

The model (5) 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 (5) can be written as

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

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

Example 3 (Dyadic data with node attributes)**.** Dyadic dataset consists of measurements on pairs of objects or under a pair of conditions. 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 $i \in V$. 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

$$\text{logit}\left(\mathbb{P}\left((i,j) \in E\right)\right) = \boldsymbol{x}_i^T \boldsymbol{B} \boldsymbol{x}_j = \langle \boldsymbol{B},\ \boldsymbol{x}_i^T \boldsymbol{x}_j \rangle, \tag{6}$$

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 the networks (Hoff, 2005). We show that our supervised tensor decompostiion (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 (6) can be expressed as

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

where $\boldsymbol{C} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{M} \in \mathbb{R}^{p\times r}$ are based on 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 available features. The regression coefficient $\mathcal{B}$ in our model model (1) serves this goal by collecting the feature effects and the interaction thereof. To encourage the sharing among effects, decomposition (2) assumes that the coefficient tensor $\mathcal{B}$ lies in a low-dimensional parameter space. The low-rank assumption is plausible in many scientific applications. In the Example 2 of brain imaging analysis, for instance, it is often believed that the brain nodes can be grouped into fewer communities, and the numbers of communities are much smaller than the number of nodes. The low-rank structure encourages the shared information across tensor entries, thereby greatly improving the estimation stability.

Our supervised tensor decomposition (3) is able to incorporate arbitrary numbers of feature matrices. We set $\boldsymbol{X}_k = \boldsymbol{I}_{d_k}$ if certain mode k has no available side information. In particular, our model (3) reduces to classical unsupervised tensor decomposition (De Lathauwer et al., 2000; Hong et al., 2019) if no side information is available; i.e., $\boldsymbol{X}_k = \boldsymbol{I}_{d_k}$ for all $k \in [K]$.

3.3 Connection to sufficient dimension reduction and tensor-on-tensor regression

An important feature of our method is that we allow high-dimensionality in both tensor dimension d_k and feature dimension p_k. In most applications, the rank is much smaller than tensor dimension d_k and the feature dimension p_k. In such a case, the matrices $\boldsymbol{M}_k$

serve the role of simultaneous dimension reduction of the data tensor and features. The implications of the sufficient features $\boldsymbol{X}_k\boldsymbol{M}_k$ can be seen in the following two conditional independence assumptions in model (3),

$$\mathcal{Y} \perp\!\!\!\perp \{\boldsymbol{X}_k\} \,\big|\, \{\boldsymbol{X}_k\boldsymbol{M}_k\} \quad \text{(independence between the tensor and multiple features)},$$
$$y_\omega \perp\!\!\!\perp y_{\omega'} \,\big|\, \{\boldsymbol{X}_k\boldsymbol{M}_k\} \quad \text{(independence within the tensor)},$$

where the second line holds for all $\omega \neq \omega' \in [d_1] \times \cdots \times [d_K]$, and $\perp\!\!\!\perp$ denotes the independence. The first property highlights the "decorrelation" role of $\boldsymbol{M}_k$ akin to the sufficient dimension reduction (Adragni and Cook, 2009) in high-dimensional supervised learning, whereas the second property highlights the tensor dimension reduction in the usual unsupervised sense (consider, for example, $\boldsymbol{X}_k = \mathbf{I}_{d_k}$ for all $k \in [K]$).

Our model (3) also has an interesting connection with tensor-on-tensor regression. We give an example on order-3 tensors for illustration. Let $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}$ denote feature matrices and consider the sufficient features $\boldsymbol{M}_1\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$, $\boldsymbol{M}_2\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2]$, $\boldsymbol{M}_3\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2]$. Denote $d_{\text{total}} = d_1d_2d_3$ and $p_{\text{total}} = p_1p_2p_3$. Let $\boldsymbol{y} = \text{vec}(\mathcal{Y}) \in \mathbb{R}^d$

Denote $d_{\text{total}} = \prod_k d_k$ and $p_{\text{total}} = \prod_k p_k$. Let $\mathcal{X} = \boldsymbol{X}_1 \otimes \cdots \otimes \boldsymbol{X}_K \in \mathbb{R}^{d_{\text{total}} \times p_{\text{total}}}$ be the tensor product of feature spaces.

Unfold$(\mathcal{X}) \in \mathbb{R}^{d_{\text{total}} \times p_{\text{total}}}$.

$$\mathbb{E}(\text{vec}(\mathcal{Y})|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = \beta_{111}\boldsymbol{x}^1\boldsymbol{z}^1\boldsymbol{w}^1 + \beta_{121}\boldsymbol{x}^1\boldsymbol{z}^2\boldsymbol{w}^1 + \beta_{112}\boldsymbol{x}^1\boldsymbol{z}^1\boldsymbol{w}^2 + \cdots + \beta_{222}\boldsymbol{x}^2\boldsymbol{z}^2\boldsymbol{w}^2.$$

where each $\boldsymbol{M}_1\boldsymbol{X} = [\boldsymbol{x}^1, \boldsymbol{x}^2]$, $\boldsymbol{M}_2\boldsymbol{Z} = [\boldsymbol{z}^1, \boldsymbol{z}^2]$, $\boldsymbol{M}_3\boldsymbol{W} = [\boldsymbol{w}^1, \boldsymbol{w}^2]$ are sufficient features. ....Full interaction model with rank-1 features. bridge these two ideas and achieve interpretable prediction.

4 Estimation

4.1 Rank-constrained M-estimator

We develop a likelihood-based procedure to estimate $\mathcal{C}$ and $\boldsymbol{M}_k$ in (3). We adopt the exponential family as a flexible framework for different data types. In a classical generalized linear model (GLM) 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. Note that the canonical link function f is chosen to be $f(\cdot) = b'(\cdot)$. Table 1 summarizes the canonical link functions for common types of 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	$\exp(\theta)$	$(1 + \exp(-\theta))^{-1}$

Table 1: Canonical links for common distributions.

In our context, we model the the entries in the response tensor y_{ijk} conditional on θ_{ijk} as independent draws from an exponential family. The quasi log-likelihood of (3) is equal

(ignoring constant) to Bregman distance between $\mathcal{Y}$ and $b'(\Theta)$:

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

We assume that we have an additional information on an upper bound $\alpha > 0$ such that $\|\Theta\|_\infty \leq \alpha$. We propose a constrained maximum likelihood estimator (M-estimator)

$$(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_1, \ldots, \hat{\boldsymbol{M}}_k) = \underset{(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K) \in \mathcal{P}}{\arg\max} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K), \tag{7}$$

where the parameter space $\mathcal{P}$ is described as

$$\mathcal{P} = \left\{ \mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K},\ \boldsymbol{M}_k \in \mathbb{R}^{p_k \times r_k} \text{ for all } k \in [K] \;\middle|\; \boldsymbol{M}_k\boldsymbol{M}_k^T = \mathbf{I}_{p_k}, \|\Theta(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K)\|_\infty \leq \alpha \right\}.$$

In the following theoretical analysis, we assume the rank $\boldsymbol{r} = (r_1, \ldots, r_K)$ is known and fixed. The adaptation of unknown $\boldsymbol{r}$ will be addressed in Section 4.3.

4.2 Alternating optimization

We propose an alternating optimization algorithm to solve (7). The decision variables in the objective function (7) consist of $K + 1$ blocks of variables, one for the core tensor $\mathcal{C}$ and K for the factor matrices $\boldsymbol{M}_k$'s. We notice that, if any K out of the $K + 1$ blocks of variables are known, then the optimization with respect to the last block of variables reduced to a simple GLM. This observation suggests that we can iteratively update one

Algorithm 1 Supervised Tensor Decomposition with Interactive 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 Tucker rank $\boldsymbol{r} = (r_1, \ldots, r_K)$, link function f, maximum norm bound α

Output: Low-rank estimation for the coefficient tensor $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$.

1: Calculate $\check{\mathcal{B}} = \mathcal{Y} \times_1 \left[(\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T\right] \times_2 \cdots \times_K \left[(\boldsymbol{X}_K^T \boldsymbol{X}_K)^{-1} \boldsymbol{X}_K^T\right]$.

2: Initialize the iteration index $t = 0$. Initialize the core tensor $\mathcal{C}^{(0)}$ and factor matrices $\boldsymbol{M}_k^{(0)} \in \mathbb{R}^{p_k \times r_k}$ via rank-$\boldsymbol{r}$ Tucker approximation of $\check{\mathcal{B}}$, in the least-square sense.

3: **while** the relative increase in objective function $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})$ is less than the tolerance **do**

4: Update iteration index $t \leftarrow t + 1$.

5: **for** $k = 1$ to K **do**

6: Obtain the factor matrix $\tilde{\boldsymbol{M}}_k^{(t+1)} \in \mathbb{R}^{p_k \times r_k}$ by a GLM with link function f.

7: Perform QR factorization on $\tilde{\boldsymbol{M}}_k^{(t+1)} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q} \in \mathbb{R}^{p_k \times r_k}$ has orthogonal columns.

8: Update $\boldsymbol{M}_k^{(t+1)} \leftarrow \boldsymbol{Q}$ and core tensor $\mathcal{C}^{(t+1)} \leftarrow \mathcal{C}^{(t+1)} \times_k \boldsymbol{R}$.

9: **end for**

10: Obtain the core tensor $\mathcal{C}^{(t+1)} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ by solving a GLM with $\mathrm{vec}(\mathcal{Y})$ as response, $\otimes_{k=1}^{K} [\boldsymbol{X}_k \boldsymbol{M}_k^{(t)}]$ as features, and f as link function. Here $\otimes$ denotes the kronecker product of matrices.

11: Rescale the core tensor subject to the maximum norm constraint.

12: Update $\mathcal{B}^{(t+1)} \leftarrow \mathcal{C}^{(t+1)} \times_1 \boldsymbol{M}_1^{(t+1)} \times_2 \cdots \times_K \boldsymbol{M}_K^{(t+1)}$.

13: **end while**

block at a time while keeping others fixed. After each iteration, we rescale the core tensor $\mathcal{C}^{(t+1)}$ subject to the maximum norm constraint. This post-processing in principle may not guarantee the monotonic increase of the objective, but we found that in our experiment this simple step appears robust for obtaining a desirable solution. The full algorithm is described in Algorithm 1.

Note that the feasible set $\mathcal{P}$ is a non-convex set. Therefore, the optimization (7) is a non-convex problem, and the Algorithm 1 usually not theoretically possesses the global optimality. **[FIXME (Miaoyan): This is wrong. Proof of Theorem 4.1 is based on global optimality condition. Please relax the proof.]** However, as mentioned

in Section 4.4, the desired convergence rate (9) holds for the valid estimators satisfying $\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) \geq \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{true})$, which indicates the global optimality is not necessarily a serious concern in our context, as long as the convergent objective of local optimums are large enough. Fortunately, we find the Algorithm 1 often gives a satisfactory convergence point $\hat{\mathcal{B}}$, when we initialize the parameters via continuous-valued Tucker decomposition. Figure 2 shows the trajectory of the objective function for order-3 tensors under the balanced setting, where $\alpha = 10, d_k = d, p_k = 0.4d, r_k = r$ for $k = 1, 2, 3, p \in \{25, 30\}$ and $r \in \{3, 6\}$. We consider the inputs 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, and the objective value at convergent points are close to or larger than the value at true parameters.

4.3 Rank selection and time complexity

Algorithm 1 takes the rank $\boldsymbol{r}$ as an input. Estimating an appropriate rank given the data is of practical importance. We propose to use Bayesian information criterion (BIC) and choose the rank that minimizes BIC; i.e.

$$\begin{aligned} \hat{\boldsymbol{r}} &= \underset{\boldsymbol{r}=(r_1,\ldots,r_K)}{\arg\min}\ \mathrm{BIC}(\boldsymbol{r}) \\ &= \underset{\boldsymbol{r}=(r_1,\ldots,r_K)}{\arg\min} \left[-2\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) + p_e(\boldsymbol{r}) \log\left(\prod_k d_k\right) \right], \end{aligned} \tag{8}$$

where $p_e(\boldsymbol{r}) \stackrel{\text{def}}{=} \sum_k (p_k - r_k - 1) r_k + \prod_k r_k$ is the effective number of parameters in the model. We choose $\hat{\boldsymbol{r}}$ that minimizes $\mathrm{BIC}(\boldsymbol{r})$ 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.

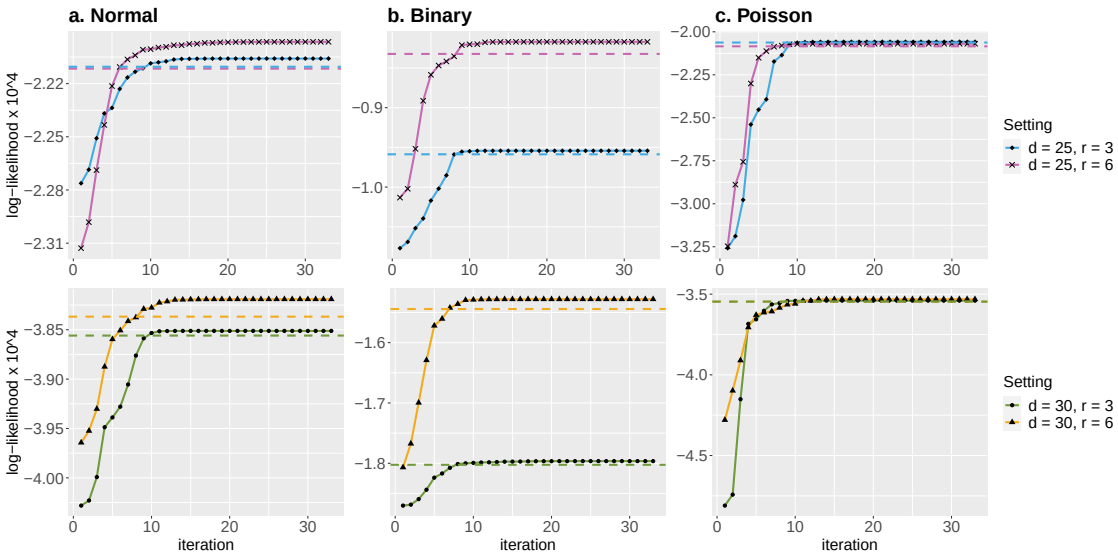

Figure 2: Trajectory of the objective function with various dimension d and rank r under (a) Gaussian (b) Bernoulli (c) Poisson models. The dashed line represents the objective value at true parameter $\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{true})$. A micro-iteration refers to the update for one of the K blocks. Four micro-iterations consist of a complete "while" iteration in Algorithm 1.

We test its empirical performance in Section 5.

The computational complexity of our alternating optimization algorithm is $O\left(d\sum_k p_k^3\right)$ for each loop of iterations, where $d = \prod_k d_k$ is the total size of the response tensor. More precisely, the update of core tensor costs $O(r^3 d)$, where $r = \prod_k r_k$ is the total size of the core tensor. The update of each factor matrix $\boldsymbol{M}_k$ involves a GLM with a d-length response, and d-by-$(r_k p_k)$ feature matrix. Solving such a GLM requires $O(d r_k^3 p_k^3)$, and therefore the cost for updating K factors in total is $O(d\sum_k r_k^3 p_k^3)$.

4.4 Statistical properties

We assess the estimation accuracy using the deviation in the Frobenius norm. Add angle deviation in dimension reduction matrices. Add identifiability condition. For the true coefficient tensor $\mathcal{B}_{\text{true}}$ and its estimator $\hat{\mathcal{B}}$, define

$$\text{Loss}(\mathcal{B}_{\text{true}},\ \hat{\mathcal{B}}) = \|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}\|_F^2.$$

In modern applications, the response tensor and features are often large-scale. We are particularly interested in the high-dimensional region in which both d_k and p_k diverge; i.e. $d_k \to \infty$ and $p_k \to \infty$, while $\frac{p_k}{d_k} \to \gamma_k \in [0, 1)$. As the size of problem grows, and so does the number of unknown parameters. As such, the classical MLE theory does not directly apply. 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 $\sigma_{max}(\cdot)$ denotes the smallest and largest singular values, respectively.

A2. There exist two positive constants $L,\ U > 0$ such that $L\phi \leq Var(y_{i_1,\ldots,i_K}|\theta_{i_1,\ldots,i_K}) \leq U\phi$ for all $|\theta_{i_1,\ldots,i_K}| \leq \alpha$.

A2'. Equivalently, there exists two positive constants $L,\ U > 0$ such that $L \leq b''(\theta) \leq U$ for all $|\theta| \leq \alpha$, where α is the upper bound of the linear predictor.

The assumptions are fairly mild. Assumption A1 guarantees the non-singularity of the feature matrices, and Assumption A2 ensures the log-likelihood $\mathcal{Y}(\Theta)$ is strictly concave in

the linear predictor Θ. Assumption A2 and A2' are equivalent, because $\mathrm{Var}(y_{i_1,\ldots,i_K}|\mathcal{X},\mathcal{B}) = \phi b''(\theta_{i_1,\ldots,i_K})$ when $y_{i_1,\ldots,i_K}$ belongs to an exponential family (McCullagh and Nelder, 1989).

Theorem 4.1 (Statistical convergence)**.** *Consider an order-K dimensional-$(d_1,\ldots,d_K)$ tensor $\mathcal{Y}$ and multiple feature matrices $\boldsymbol{X}_k \in \mathbb{R}^{d_k\times p_k}$, for $k\in[K]$. Suppose that, conditional on the features, the entries in $\mathcal{Y}$ are independent realizations from an exponential family distribution with conditional mean model* (3)*. Let $r_{total} = \prod_k r_k$, $r_{\max} = \max_k r_k$, and $\lambda_{\min}(\boldsymbol{M}_k)$ be the minimal non-zero singular value of $\boldsymbol{M}_k$. Assume that $\lambda_{\min}(\boldsymbol{M}_k) > 0$ and $\|\boldsymbol{X}_k\|_F \asymp \Omega(\sqrt{d_k})$, for all $k\in[K]$. Under Assumption 1, there exist two constants $C_1, C_2 > 0$, such that, with probability at least $1-\exp(-C_1\sum_k p_k)$,*

$$\sin^2\Theta(\boldsymbol{M}_k,\ \hat{\boldsymbol{M}}_k) \le \frac{C_2 r_{\text{total}}}{r_{\max}\lambda_{\min}(\boldsymbol{M}_k)}\frac{\sum_k p_k}{\prod_k d_k},\quad \textit{for all } k\in[K], \tag{9}$$

and

$$\|\mathcal{B}_{true} - \hat{\mathcal{B}}\|_F^2 \le \frac{C_2 r_{\text{total}}}{r_{\max}}\frac{\sum_k p_k}{\prod_k d_k}.$$

Here, $C_2 = C_2(\boldsymbol{r},\alpha,K) > 0$ is a constant independent of the dimensions $\{d_k\}$ and $\{p_k\}$.

Theorem 4.1 establishes the statistical convergence for the estimator (7). **[FIXME (Miaoyan): Wrong statement based on current proof]**Actually, the proof in section A shows that the statistically optimal rate holds, not only for the MLE (7), but also any local estimators $\hat{\mathcal{B}}$ in the level set $\{\hat{\mathcal{B}}\in\mathcal{P}\colon \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) \ge \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{true})\}$ satisfies the upper bound 4.1. In Section 4.2, we implement empirical studies to verify the algorithmic stability to find a local optimum.

To gain further insight on the bound (9), we consider a special case when tensor dimensions are equal at each of the modes, i.e., $d_k = d$, $p_k = \gamma d$, $\gamma\in[0,1)$ for all $k\in[K]$, and the feature matrices $\boldsymbol{X}_k$ are Gaussian design with i.i.d. $N(0,1)$ entries. To put the

context in the framework of Theorem 4.1, we rescale the feature matrices into $\check{\boldsymbol{X}}_k = \frac{1}{\sqrt{d}}\boldsymbol{X}_k$ so that the singular values of $\check{\boldsymbol{X}}_k$ are bounded by $1 \pm \sqrt{\gamma}$. The result in (9) implies that the estimated coefficient has a convergence rate $\mathcal{O}(\frac{p}{d^K})$ in the scale of the original feature $\{\boldsymbol{X}_k\}$. Therefore, our estimation is consistent as the dimension grows, and the convergence becomes especially favorably as the order of tensor data increases.

As immediate applications, we obtain the convergence rate for the three examples mentioned in Section 3. Without loss of generality, we assume that the singular values of the d_k-by-p_k feature matrix $\boldsymbol{X}_k$ are bounded by $\sqrt{d_k}$.

Example 4 (Spatio-temporal growth model)**.** The estimated type-by-time-by-space coefficient tensor converges at the rate $\mathcal{O}\left(\frac{p+r+s}{dmn}\right)$ where $p \leq d$, $r \leq m$ and $s \leq n$. The estimation achieves consistency as long as the dimension grows in either of the three modes.

Example 5 (Network population model)**.** The estimated node-by-node-by-feature tensor converges at the rate $\mathcal{O}\left(\frac{2d+p}{d^2n}\right)$ where $p \leq n$. The estimation achieves consistency as the number of individuals or the number of nodes grows.

Example 6 (Dyadic data with node attributes)**.** The estimated feature-by-feature matrix converges at the rate $\mathcal{O}\left(\frac{p}{d^2}\right)$ where $p \leq d$. Again, our estimation achieves consistency as the number of nodes grows.

We conclude this section by providing the prediction accuracy, measured in KL divergence, for the response distribution.

Theorem 4.2 (Prediction error)**.** *Assume the same set-up as in Theorem 4.1. Let* $\mathbb{P}_{\mathcal{Y}_{true}}$ *and* $\mathbb{P}_{\hat{\mathcal{Y}}}$ *denote the distributions of* $\mathcal{Y}$ *given the true parameter* $\mathcal{B}_{true}$ *and estimated parameter*

$\hat{\mathcal{B}}$, *respectively. Then, we have, with probability at least* $1 - \exp(C_1 \sum_k p_k)$,

$$KL(\mathbb{P}_{\mathcal{Y}_{true}},\ \mathbb{P}_{\hat{\mathcal{Y}}}) \leq \frac{C_4 r_{\text{total}}}{r_{\max}} \frac{\sum_k p_k}{\prod_k d_k},$$

where $C_4 = C_4(\boldsymbol{r}, \alpha, K) > 0$ *is a constant that independent of dimensions* $\{d_k\}$ *and* $\{p_k\}$.

5 Numerical experiments

Simulation needs to be re-done, since code has been changed.

We evaluate the empirical performance of our generalized tensor regression through simulations. We consider order-3 tensors with a range of distribution types. The coefficient tensor $\mathcal{B}$ is generated using the factorization form (**??**) where both the core and factor matrices are drawn i.i.d. from Uniform[-1,1]. The linear predictor is then simulated from $\mathcal{U} = \mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$, where $\boldsymbol{X}_k$ is either an identity matrix (i.e. no feature available) or Gaussian random matrix with i.i.d. entries from $N(0, \sigma_k^2)$. We set $\sigma_k = d_k^{-1/2}$ to ensure the singular values of $\boldsymbol{X}_k$ are bounded as d_k increases. The $\mathcal{U}$ is scaled such that $\|\mathcal{U}\|_\infty = 1$. Conditional on the linear predictor $\mathcal{U} = [\![u_{ijk}]\!]$, the entries in the tensor $\mathcal{Y} = [\![y_{ijk}]\!]$ are drawn independently according to one of the following three probabilistic models:

(a) (Gaussian). Continuous data $y_{ijk} \sim N(\alpha u_{ijk}, 1)$.

(b) (Poisson). Count data $y_{ijk} \sim \text{Poi}(e^{\alpha u_{ijk}})$.

(c) (Bernoulli). Binary data $y_{ijk} \sim \text{Ber}\left(\frac{e^{\alpha u_{ijk}}}{1+e^{\alpha u_{ijk}}}\right)$.

Here $\alpha > 0$ is a scalar controlling the magnitude of the effect size. In each simulation study, we report the mean squared error (MSE) for the coefficient tensor averaged across

True Rank $\boldsymbol{r}$	Dimension (Gaussian tensors) $d = 20$	$d = 40$	Dimension (Poisson tensors) $d = 20$	$d = 40$
(3, 3, 3)	(2.1, 2.0, 2.0)	(**3**, **3**, **3**)	(2.0, 2.2, 2.1)	(**3**, **3**, **3**)
(4, 4, 6)	(3.2, 3.1, 5.0)	(**4**, **4**, **6**)	(**4.0**, **4.0**, 5.2)	(**4**, **4**, **6**)
(6, 8, 8)	(5.1, 7.0, 6.9)	(**6**, **8**, **8**)	(5.0, 6.1, 7.1)	(**6**, **8**, **8**)

Table 2: Rank selection via BIC. Bold number indicates no significant difference between the estimate and the ground truth, based on a z-test with a level 0.05.

$n_{\text{sim}} = 30$ replications.

5.1 Finite-sample Performance

The first experiment assesses the selection accuracy of our BIC criterion (8). We consider the balanced situation where $d_k = d$, $p_k = 0.4d_k$ for $k = 1, 2, 3$. We set $\alpha = 10$ and consider various combinations of dimension d and rank $\boldsymbol{r} = (r_1, r_2, r_3)$. For each combination, we simulate tensor data following Gaussian, Bernoulli, and Poisson models. We then minimize BIC using a grid search over three dimensions. The hyper-parameter α is set to infinity in the fitting, which essentially imposes no prior on the coefficient magnitude. Table 2 reports the selected rank averaged over $n_{\text{sim}} = 30$ replicates for Gaussian and Poisson models. We found that when $d = 20$, the selected rank is slightly smaller than the true rank, and the accuracy improves immediately when the dimension increases to $d = 40$. This agrees with our expectation, as in tensor regression, the sample size is related to the number of entries. A larger d implies a larger sample size, so the BIC selection becomes more accurate.

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 25 to 50. Our theoretical analysis suggests that $\hat{\mathcal{B}}$ has a convergence rate $\mathcal{O}(d^{-2})$ in this setting. Figure 1 plots the estimation error versus the "effective sample size", d^2, under three

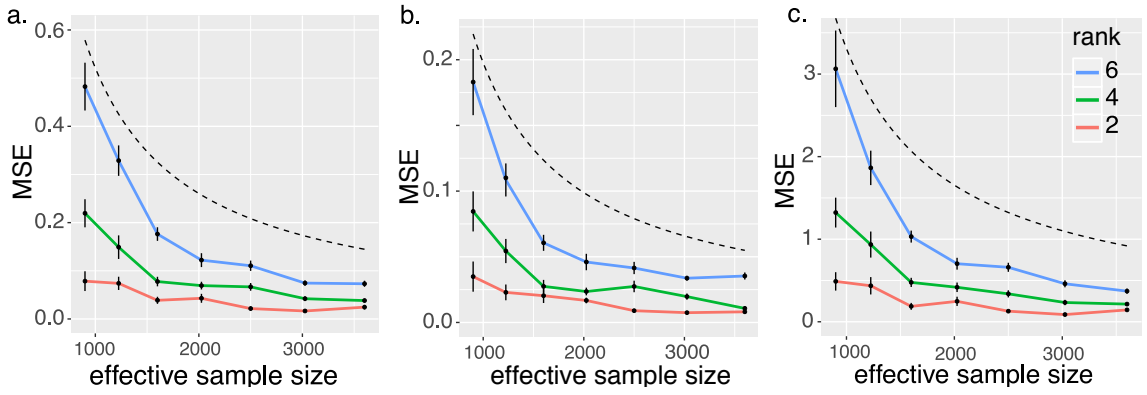

Figure 3: Estimation error against effective sample size. The three panels depict the MSE when the response tensors are generated form (a) Gaussian (b) Poisson and (c) Bernoulli models. The dashed curves correspond to $\mathcal{O}(1/d^2)$.

different distribution models. We found that the empirical MSE decreases roughly at the rate of $1/d^2$, which is consistent with our theoretical ascertainment. We also observed that, tensors with higher ranks 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. Similar behaviors can be observed in the non-Gaussian data in Figure 3b-c.

The third experiment investigates our model's ability in handling correlation among coefficients. 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 (Robinson et al., 2015) 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

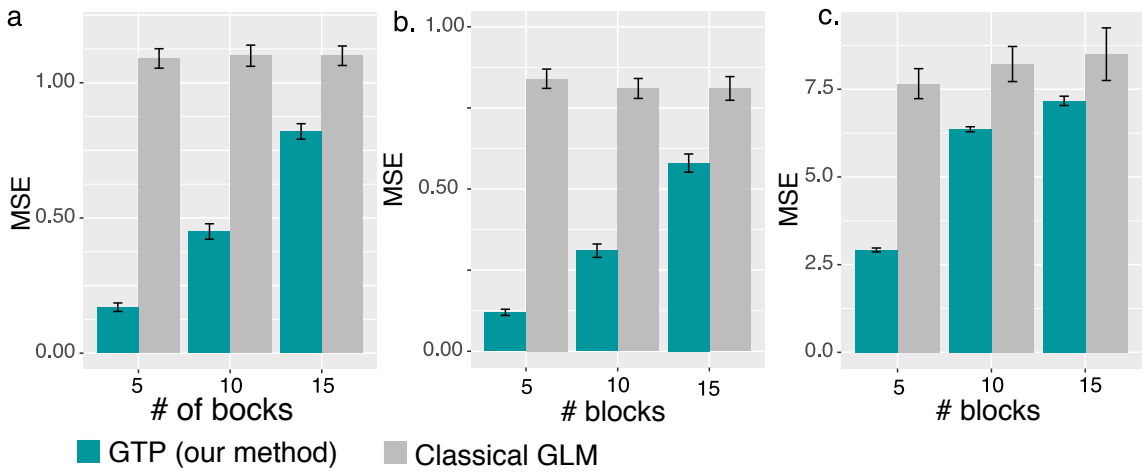

Figure 4: Performance comparison when the networks have block structure. The three panels depict the MSE when the response tensors are generated form (a) Gaussian (b) Poisson and (c) Bernoulli models. The x-axis represents the number of blocks in the networks.

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 drawn i.i.d. 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 matrix is not necessarily equal to r (Wang and Zeng, 2019).

Figure 4 compares the MSE of our method with a classical GLM approach. A classical GLM is to regress the dyadic edges, one at a time, on the features, and this model is repeatedly fitted for each edge. This repeated approach, however, does not account for the correlation among the edges, and may suffer from overfitting. As we can see in Figure 4, out tensor regression method achieves significant error reduction in all three models considered.

The outer-performance is significant in the presence of large communities, and even in the less structured case ($\sim 20/15 = 1.33$ nodes per block), our method still outer-performs GLM. This is because the low-rankness in our modeling automatically identifies the shared information across entries. By selecting the rank in a data-driven way, our method is able to achieve accurate estimation with improved interpretability.

5.2 Comparison with alternative methods

We compare our generalized tensor regression (**GTR**) with three other supervised tensor methods:

- Higher-order low-rank regression (**HOLRR**, (Rabusseau and Kadri, 2016)) is a least-square based tensor regression that allows features on a single mode.
- Higher-order partial least square (**HOPLS**, (Zhao et al., 2012)) is a dimension-reduction method that jointly models a tensor response and a tensor feature.
- Subsampled tensor projected gradient (**TPG**, (Yu and Liu, 2016)) tackles the same question as **HOLRR** but instead uses a different algorithm to solve the problem.

These three methods are the closest algorithms to ours, in that they relate a tensor response to features using a low-rank structure. All the three methods allow only Gaussian data, whereas ours is applicable to any exponential family distribution including Gaussian, Bernoulli, Multinomial, etc. For fair comparison, we consider only Gaussian response in the simulation. We measure the accuracy using mean squared prediction error, $\text{MSPE} = \sqrt{\sum_k d_k}\|\hat{\mathcal{Y}} - \mathbb{E}(\mathcal{Y}|\mathcal{X})\|_F$, where $\hat{\mathcal{Y}}$ is the fitted value from each of the methods.

The comparison was assessed from three aspects: (a) benefit of incorporating features from multiple modes; (b) prediction error with respect to sample size; (c) sensitivity of

accuracy with respect to model complexity. We use similar simulation setups as in our experiment II, but consider combinations of rank ($\boldsymbol{r} = (3,3,3)$ vs. $(4,5,6)$), noise ($\sigma = 1/2$ vs. $1/4$), and dimension (d ranging from 20 to 100 for modes with features, $d = 20$ for modes without features).

Figure 5 shows the averaged prediction error across 30 replicates. We see that our **GTR** outperforms others, especially in the high-rank high-noise setting. As the number of informative modes (i.e. modes with available features) increases, the **GTR** exhibits a reduction in error whereas others have increased errors. This showcases the benefit toward prediction via incorporation of interactive features. Note that our method **GTR** is most comparable to **HOLRR** when there is only a single informative mode. In such a case, both methods share a same cost function but have different algorithms. **GTR** alternates between informative and non-informative modes, whereas **HOLRR** approximates the non-informative modes via unfolded response alone. The accuracy gain in Figure 5 demonstrates the benefit of alternating algorithm – having informative modes also improves the estimation along non-informative modes.

Figure 6 compares the prediction error with respect to sample size. The sample size is the total number of entries in the tensor. In the low-rank setting, our method has similar performance as **HOLRR**, and the improvement becomes more pronounced when the rank increases. Neither **HOPLS** nor **TPG** has satisfactory performance in high-rank or high-noise settings. One possible reason is that a higher rank implies a higher inter-mode complexity, and our **GTR** method lends itself well to this context.

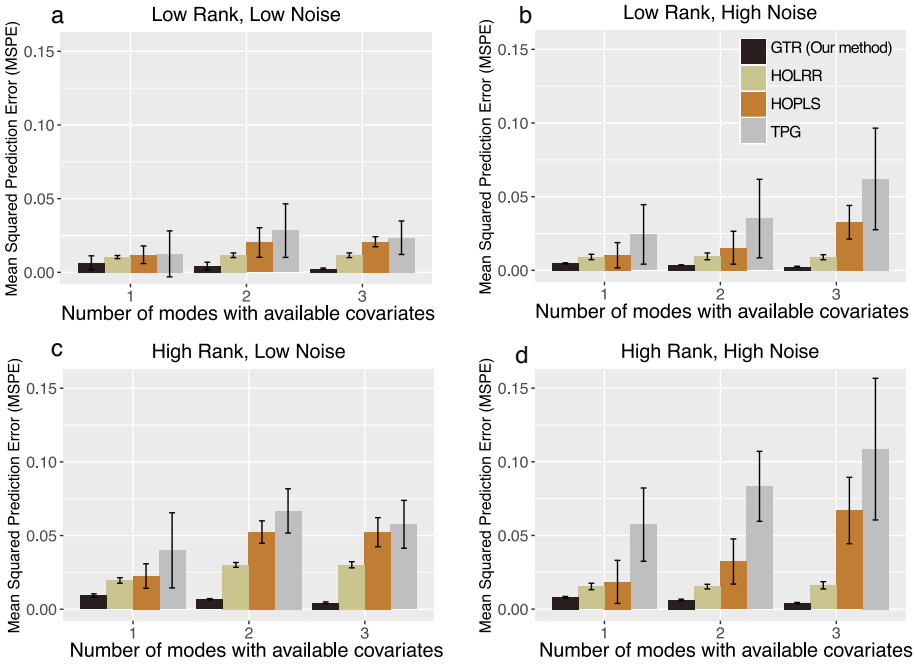

Figure 5: Comparison of MSPE versus the number of modes with features. We consider rank $\boldsymbol{r} = (3, 3, 3)$ (low), $\boldsymbol{r} = (4, 5, 6)$ (high), and noise $\sigma = 1/2$ (high), $\sigma = 1/4$ (low).

6 Data analysis

We apply our tensor regression model to two datasets. The first application concerns the brain network modeling 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).

6.1 Analysis of 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 (Geddes, 2016).

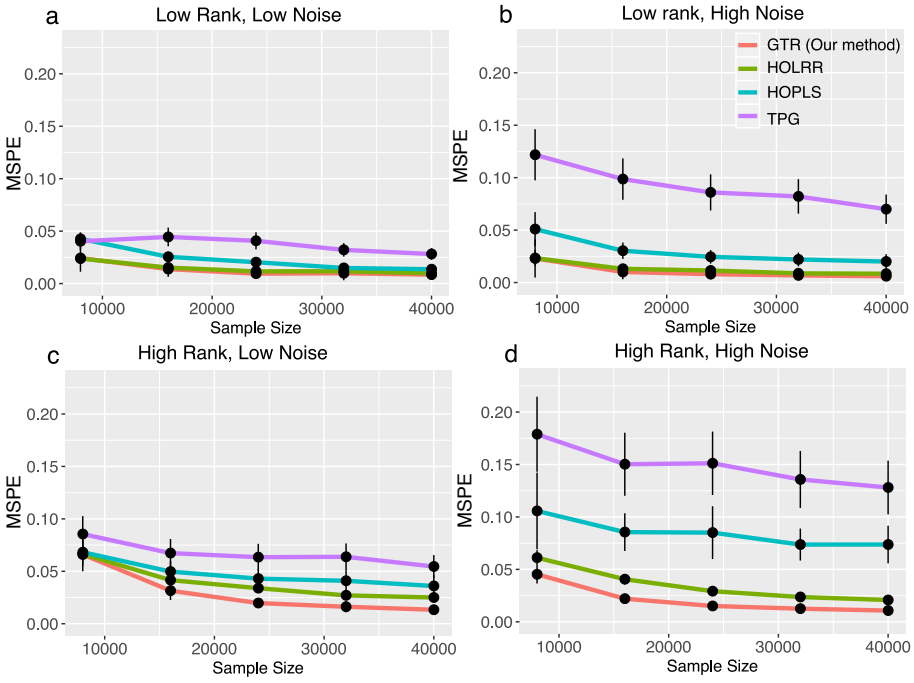

Figure 6: Comparison of MSPE versus sample size. We consider rank $\boldsymbol{r} = (3, 3, 3)$ (low), $\boldsymbol{r} = (4, 5, 6)$ (high), and noise $\sigma = 1/2$ (high), $\sigma = 1/4$ (low).

We take a subset of HCP data that 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 goal is to identify the connection edges that are affected by the individual feature. A key challenge in brain network is that the edges are correlated; for example, two edges may stem out from a same brain region, and it is of importance to take into account the within-dyad dependence.

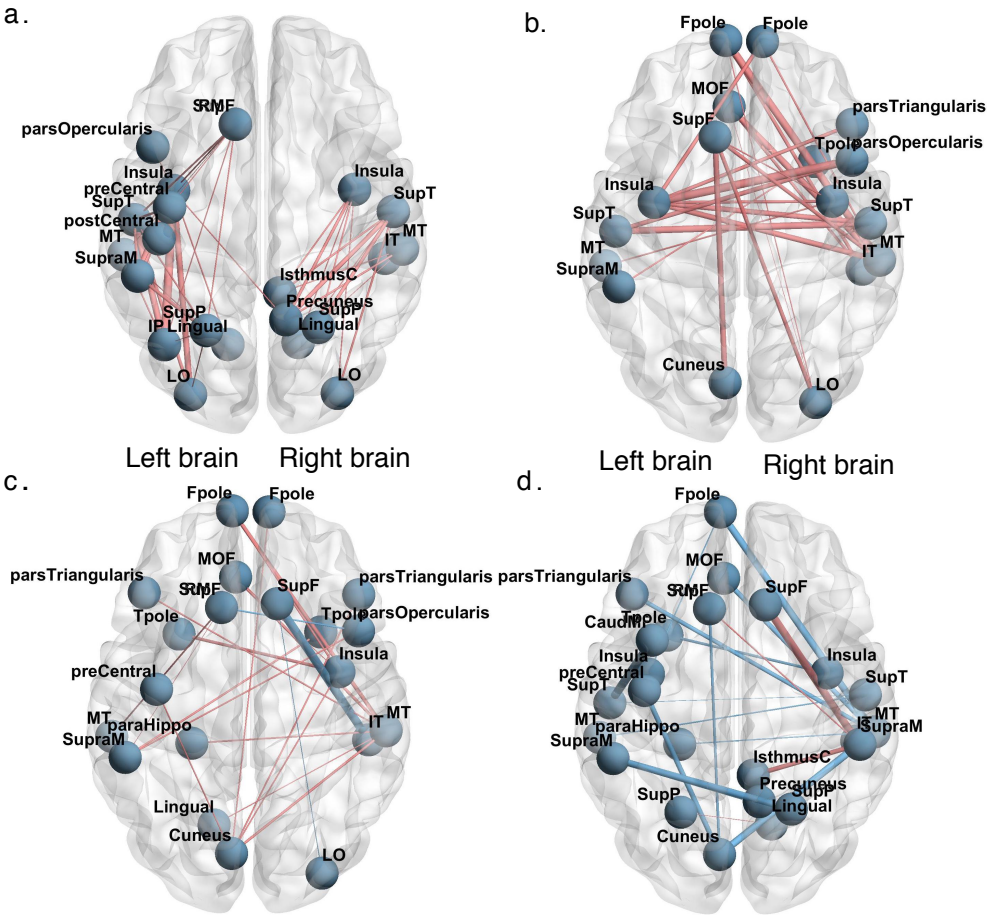

Figure 7: Top edges with large effects. Red edges represent relatively strong connections and blue edges represent relatively weak connections. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+.

We fit the tensor regression model to the HCP data. The response is a binary tensor $\mathcal{Y} \in \{0,1\}^{68\times68\times136}$ and the features are of dimension 4 along the 3$^{\text{rd}}$ mode. The BIC selection suggests a rank $\boldsymbol{r} = (10, 10, 4)$ with log-likelihood $\mathcal{L}_{\mathcal{Y}} = -174654.7$. Figure 7 shows the top edges with high effect size, overlaid on the Desikan atlas brain template (Desikan et al., 2006; Xia et al., 2013). We utilize the sum-to-zero contrasts in the effects coding and

depicted only the top 3% edges whose connections are non-constant across the sample. It is observed that the global connection exhibits clear spatial separation, and that the nodes within each hemisphere are more densely connected with each other (Figure 7a). 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 7b). This is in agreement with a recent study showing that female brains are optimized for inter-hemispheric communication (Ingalhalikar et al., 2014). We also found several edges with declined connection in the group Age 31+. Notably, those edges involve Frontal-pole ($Fploe$), superior-frontal ($SupF$) and Cuneus nodes. The Frontal-pole region has long been known for its importance in memory and cognition, and the detected decline with age further highlights its biological importance.

Figure 8 compares the estimated coefficients from our method (tensor regression) with those from classical GLM approach. A classical GLM is to regress the brain edges, one at a time, on the individual-level features, and this logistic model is repeatedly fitted for every edge $\in [68] \times [68]$. As we can see in the figure, our tensor regression shrinkages the coefficients towards center, thereby enforcing the sharing between coefficient entries.

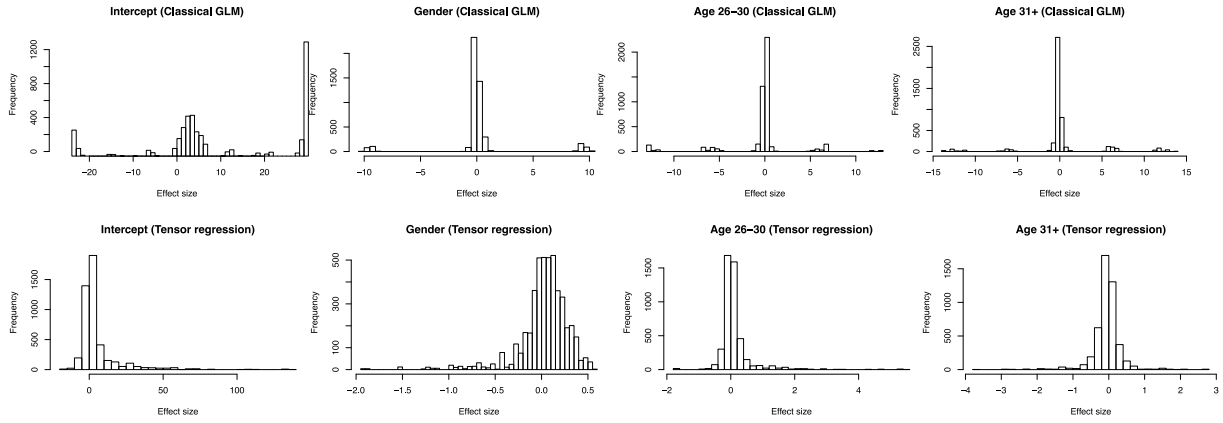

Figure 8: Comparison of coefficient estimation in the HCP data.

6.2 Analysis of political relational data

The second application concerns the multi-relational network analysis with node-level attributes. We consider *Nations* dataset (Nickel et al., 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 a connection, such as "sending tourist to", "export", "import", between countries. The 56 relations span the fields of politics, economics, military, religion, and so on. In addition, country-level attributes are also available, and we focus on the following six features: *constitutional, catholics, lawngos, politicalleadership, geographyx*, and *medicinengo*. The goal is to identify the variation in connections due to country-level attributes and interactions thereof. One of the key features is that the 56 relations are correlated, and we would like to take that into account in assessing the feature effects.

We applied our tensor regression model to the *Nations* data. The multi-relational network $\mathcal{Y} \in \{0,1\}^{14\times14\times56}$ was treated as the response tensor, and the country attributes

$\boldsymbol{M} \in \mathbb{R}^{14\times 6}$ were treated as features on both the 1st and 2nd modes. The BIC criterion suggests a rank $\boldsymbol{r} = (4, 4, 4)$ for the coefficient tensor $\mathcal{B} \in \mathbb{R}^{6\times 6\times 56}$. Table Section 8 shows the K-mean clustering of the 56 relations based on the 3^{rd} mode factor $\boldsymbol{M}_3 \in \mathbb{R}^{56\times 4}$. We found that the relations reflecting the similar aspects of international affairs are grouped together. In particular, Cluster I consists of political relations such as *officialvisits, intergovorgs*, and *militaryactions*; Clusters II and III capture the economical relations such as *economicaid, booktranslations, tourism*; and Cluster IV represents the Cold War alliance blocs. The similarity among entities in each cluster suggests the plausibility of our dimension reduction.

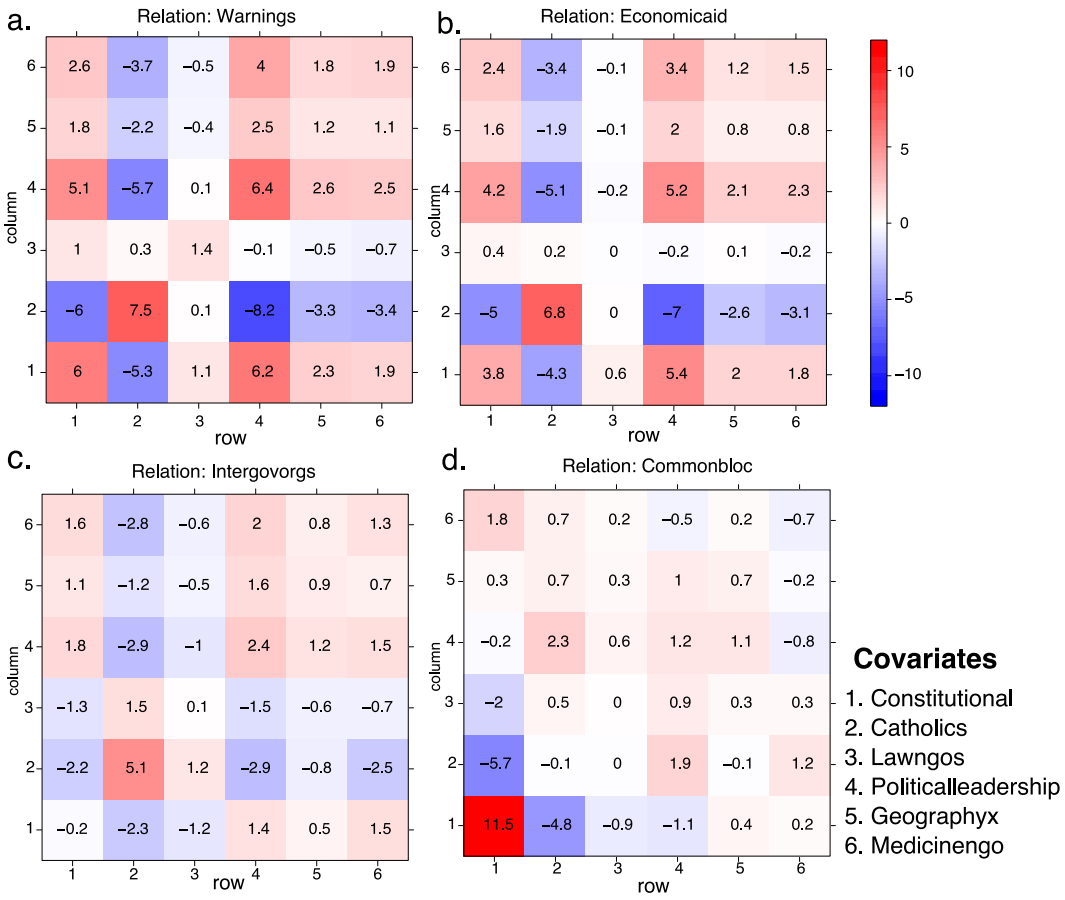

Figure 9: Effect estimation in the *Nations* data. Panels (a)-(d) represent the estimated effects of country-level attributes towards the connection probability, for relations *warnning*, *economicaid*,*intergovorg*, and *commonblock*, respectively.

Cluster I	Government	officialvisits, intergovorgs, militaryactions, negativebehavior, boycottembargo, aidenemy, negativecomm, protestsunoffialacts, nonviolentbehavior, emigrants, timesincewar, commonbloc2, rintergovorgs3, relintergovorgs violentactions, duration,accusation,relexports
Cluster II	Economics	economicaid, booktranslations, tourism, conferences, severdiplomatic, expeldiplomats, attackembassy, reltourism, tourism3, relemigrants, emigrants3, students, exports, exports3, lostterritory, dependent, militaryalliance relbooktranslations, releconomicaid, unweightedunvote, warning, relstudents,
Cluster III	treaties	treaties, reltreaties, exportbooks, relexportbooks, relngo, ngoorgs3, embassy, reldiplomacy, timesinceally, independence, commonbloc1,weightedunvote, ngo,
Cluster IV	Politics	commonbloc0, blockpositionindex

Table 3: K-means clustering of relations based on factor matrix in the coefficient tensor.

To investigate the effects of dyadic attributes towards connections, we depicted the estimated coefficients $\hat{\mathcal{B}} = [\![\hat{b}_{ijk}]\!]$ for several relation types (Figure 9). Note that entries $\hat{b}_{ijk}$ can be interpreted as the contribution, at the logit scale, of feature pair (i, j) (ith feature for the "sender" country and jth feature for the "receiver" country) towards the connection of relation k. Several interesting findings emerge from the observation. We found that relations belonging to a same cluster tend to have similar feature effects. For example, the relations *warnings* and *ecnomicaid* are classified into Cluster II, and both exhibit similar feature pattern (Figure 9a-b). Moreover, the majority of the diagonal entries $\hat{\mathcal{B}}(i, i, k)$ positively contribute to the connection. This suggests that countries with coherent attributes tend to interact more often than others. We also found that the *constitutional* attribute is an important predictor for the *commonbloc* relation, whereas the effect is weaker for other relations (Figure 9d). This is not surprising, as the block partition during Cold

War is associated with the *constitutional* attribute.

7 Conclusion

We have developed a generalized tensor regression with features on multiple modes. A fundamental feature of tensor-valued data is the statistical interdependence among entries. Our proposed rank-constrained estimation achieves high accuracy with sound theoretical guarantees. The estimation accuracy is quantified via deviation in the Frobenius norm and K-L divergence. Other measures of accuracy may also be desirable, such as the spectral norm or the maximum norm of the deviation. Exploiting the properties and benefits of different error quantification warrants future research.

Add three more paragraphs:

1. Optimization

2. Assumption on (conditional) independence assumption on the tensor entries.

3. Inference. ... uncertainly quantification... bootstrap .... Testing for Interaction + main effects.

Acknowledgements

This research was supported by NSF grant DMS-1915978 and the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

References

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

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(1906):4385–4405.

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

Baldin, N. and Berthet, Q. (2018). Optimal link prediction with matrix logistic regression. *arXiv preprint arXiv:1803.07054*.

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.

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

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

Fan, J., Gong, W., and Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*.

Gabriel, K. R. (1998). Generalised bilinear regression. *Biometrika*, 85(3):689–700.

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*, pages 1–23.

Geddes, L. (2016). Human brain mapped in unprecedented detail. *Nature*.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295.

Hong, D., Kolda, T. G., and Duersch, J. A. (2019). Generalized canonical polyadic tensor decomposition. *SIAM Review, in press. arXiv:1808.07452*.

Hong, D., Kolda, T. G., and Duersch, J. A. (2020). Generalized canonical polyadic tensor decomposition. *SIAM Review*, 62(1):133–163.

Hu, C., Rai, P., Chen, C., Harding, M., and Carin, L. (2015). Scalable bayesian non-negative tensor factorization for massive count data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 53–70. Springer.

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(2):823–828.

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

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics*, 27(3):638–647.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Narita, A., Hayashi, K., Tomioka, R., and Kashima, H. (2012). Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery*, 25(2):298–324.

Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*, volume 11, pages 809–816.

Potthoff, R. F. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3-4):313–326.

Rabusseau, G. and Kadri, H. (2016). Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems*, pages 1867–1875.

Raskutti, G., Yuan, M., and Chen, H. (2015). Convex regularization for high-dimensional multi-response tensor regression. *arXiv preprint arXiv:1512.01215*.

Robinson, L. F., Atlas, L. Y., and Wager, T. D. (2015). Dynamic functional connectivity using state-based dynamic community structure: Method and application to opioid analgesia. *NeuroImage*, 108:274–291.

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(4):357–370.

Sun, W. W. and Li, L. (2017). STORE: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944.

Tomioka, R. and Suzuki, T. (2014). Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*.

Wang, L., Zhang, Z., Dunson, D., et al. (2019). Common and individual structure of brain networks. *The Annals of Applied Statistics*, 13(1):85–112.

Wang, M., Duc, K. D., Fischer, J., and Song, Y. S. (2017). Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520:44–66.

Wang, M. and Li, L. (2018). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *arXiv preprint arXiv:1811.05076*.

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

Xia, M., Wang, J., and He, Y. (2013). Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910.

Yu, R. and Liu, Y. (2016). Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381.

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

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

Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Nagasaka, Y., Fujii, N., Zhang, L., and Cichocki, A. (2012). Higher order partial least squares (HOPLS): a generalized multilinear regression method. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1660–1673.

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

SUPPLEMENTARY MATERIAL

A Proofs

Proof of Theorem 4.1. Define $\ell(\mathcal{B}) = \mathbb{E}(\mathcal{L}_{\mathcal{Y}}(\mathcal{B}))$, where the expectation is taken with respect to $\mathcal{Y} \sim \mathcal{B}_{\text{true}}$ under the model with true parameter $\mathcal{B}_{\text{true}}$. We first prove the following two conclusions:

C1. There exists two positive constants C_1, $C_2 > 0$, such that, with probability at least $1 - \exp(-C_1 \log K \sum_k p_k)$, the stochastic deviation, $\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) - \ell(\mathcal{B})$, satisfies

$$|\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) - \ell(\mathcal{B})| = |\langle \mathcal{E},\ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle| \leq C_2 \|\mathcal{B}\|_F \sqrt{\frac{\prod_k r_k}{\max_k r_k} \sum_k p_k}.$$

C2. The inequality $\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}}) \leq -\frac{L}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2$ holds, where $L > 0$ is the lower bound for $\min_{|\theta| \leq \alpha} |b''(\theta)|$.

To prove C1, we note that the stochastic deviation can be written as:

$$\begin{aligned} \mathcal{L}_{\mathcal{Y}}(\mathcal{B}) - \ell(\mathcal{B}) &= \langle \mathcal{Y} - \mathbb{E}(\mathcal{Y}|\mathcal{X}),\ \Theta(\mathcal{B}) \rangle \\ &= \langle \mathcal{Y} - b'(\Theta^{\text{true}}),\ \Theta \rangle \\ &= \langle \mathcal{E} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T,\ \mathcal{B} \rangle, \end{aligned} \tag{10}$$

where $\mathcal{E} \stackrel{\text{def}}{=} \mathcal{Y} - b'(\Theta^{\text{true}})$, and the second line uses the property of exponential family that $\mathbb{E}(\mathcal{Y}|\mathcal{X}) = b'(\Theta^{\text{true}})$. Based on Proposition 2, the boundedness of $b''(\cdot)$ implies that $\mathcal{E}$ is a sub-Gaussian-(ϕU) tensor. Let $\check{\mathcal{E}} \stackrel{\text{def}}{=} \mathcal{E} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T$. By Proposition 1, $\check{\mathcal{E}}$ is

a $(p_1, \ldots, p_K)$-dimensional sub-Gaussian tensor with parameter bounded by $C = \phi U c_2^K$. Here $c_2 > 0$ is the upper bound of $\sigma_{\max}(\boldsymbol{X}_k)$. Applying Cauchy-Schwarz inequality to (10) yields

$$|\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) - \ell(\mathcal{B})| \leq \left\|\check{\mathcal{E}}\right\|_2 \|\mathcal{B}\|_*, \tag{11}$$

where $\|\cdot\|_2$ denotes the tensor spectral norm and $\|\cdot\|_*$ denotes the tensor nuclear norm. The nuclear norm $\|\mathcal{B}\|_*$ is bounded by $\|\mathcal{B}\|_* \leq \sqrt{\frac{\prod_k r_k}{\max_k r_k}} \|\mathcal{B}\|_F$ (c.f. (Wang and Li, 2018; Wang et al., 2017)). The spectral norm $\left\|\check{\mathcal{E}}\right\|_2$ is bounded by $\left\|\check{\mathcal{E}}\right\|_2 \leq C_2 \sqrt{\sum_k p_k}$ with probability at least $1 - \exp(-C_1 \log K \sum_k p_k)$ (c.f. (Wang and Li, 2018; Tomioka and Suzuki, 2014)). Combining these two bounds with (11), we have, with probability at least $1 - \exp(-C_1 \log K \sum_k p_k)$,

$$|\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) - \ell(\mathcal{B})| \leq C_2 \|\mathcal{B}\|_F \sqrt{\frac{\prod_k r_k}{\max_k r_k} \sum_k p_k},$$

where $C_2 > 0$ is a constant absorbing all factors that do not depend on $\{p_k\}$ and $\{r_k\}$.

Next we prove C2. Applying Taylor expansion to $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})$ around $\mathcal{B}_{\text{true}}$ yields

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) = \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) + \left\langle \frac{\partial \mathcal{L}_{\mathcal{Y}}(\mathcal{B})}{\mathcal{B}}\Big|_{\mathcal{B}=\mathcal{B}_{\text{true}}}, \mathcal{B} - \mathcal{B}_{\text{true}} \right\rangle + \frac{1}{2} \text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}})^T \mathcal{H}(\check{\mathcal{B}}) \text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}}),$$

where $\mathcal{H}_{\mathcal{Y}}(\check{\mathcal{B}})$ is the (non-random) Hession of $\frac{\partial \mathcal{L}^2_{\mathcal{Y}}(\mathcal{B})}{\partial^2 \mathcal{B}}$ evaluated at $\check{\mathcal{B}} = \alpha \text{vec}(\alpha\mathcal{B} + (1-\alpha)\mathcal{B}_{\text{true}})$ for some $\alpha \in [0, 1]$. Note that we have $\mathbb{E}\left(\frac{\partial \mathcal{L}_{\mathcal{Y}}(\mathcal{B})}{\partial \mathcal{B}}\Big|_{\mathcal{B}=\mathcal{B}_{\text{true}}}\right) = 0$. We take expectation with respect to $\mathcal{Y} \sim \mathcal{B}_{\text{true}}$ on both sides of (12) and obtain

$$\ell(\mathcal{B}) = \ell(\mathcal{B}_{\text{true}}) + \frac{1}{2} \text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}})^T \mathcal{H}(\check{\mathcal{B}}) \text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}}). \tag{12}$$

By the fact $\frac{\partial \mathcal{L}_{\mathcal{Y}}^2(\Theta)}{\partial^2 \Theta} = -b''(\Theta)$ and chain rule over $\Theta = \Theta(\mathcal{B}) = \mathcal{B} \times_1 \boldsymbol{X}_1 \cdots \times_K \boldsymbol{X}_K$, the equation (12) implies that

$$\ell(\mathcal{B}) - \ell(\mathcal{B}_{\text{true}}) = -\frac{1}{2} \sum_{i_1,\ldots,i_K} b''(\check{\theta}_{i_1,\ldots,i_K})(\theta_{i_1,\ldots,i_K} - \theta_{\text{true},i_1,\ldots,i_K})^2 \leq -\frac{L}{2}\|\Theta - \Theta^{\text{true}}\|_F^2,$$

holds for all $\mathcal{B} \in \mathcal{P}$, provided that $\min_{|\theta| \leq \alpha} |b''(\theta)| \geq L > 0$. In particular, the inequality (12) also applies to the constrained MLE $\hat{\mathcal{B}}$. So we have

$$\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}}) \leq -\frac{L}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2. \tag{13}$$

Now we have proved both C1 and C2. Note that $\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) \geq 0$ by the definition of $\hat{\mathcal{B}}$, This implies that

$$\begin{aligned}
0 &\leq \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) \\
&\leq \Big(\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \ell(\hat{\mathcal{B}})\Big) - (\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) - \ell(\mathcal{B}_{\text{true}})) + \Big(\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}})\Big) \\
&\leq \langle \mathcal{E},\ \Theta - \Theta^{\text{true}} \rangle - \frac{L}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2,
\end{aligned}$$

where the second line follows from (13). Therefore,

$$\begin{aligned}
\|\hat{\Theta} - \Theta^{\text{true}}\|_F &\leq \frac{2}{L}\Big\langle \mathcal{E},\ \frac{\hat{\Theta} - \Theta^{\text{true}}}{\|\hat{\Theta} - \Theta^{\text{true}}\|_F} \Big\rangle \\
&\leq \frac{2}{L} \sup_{\Theta:\|\Theta\|_F=1, \Theta=\mathcal{B}\times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K} \langle \mathcal{E},\ \Theta \rangle \\
&\leq \frac{2}{L} \sup_{\mathcal{B}\in\mathcal{P}:\|\mathcal{B}\|_F \leq \prod_k \sigma_{\min}^{-1}(\boldsymbol{X}_k)} \langle \mathcal{E},\ \ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle.
\end{aligned} \tag{14}$$

Combining (14) with C1 yields

$$\|\hat{\Theta} - \Theta^{\text{true}}\|_F \leq \frac{2C_2}{L} \prod_k \sigma_{\min}^{-1}(\boldsymbol{X}_k) \sqrt{\frac{\prod_k r_k}{\max r_k} \sum_k p_k}.$$

Therefore, the final conclusion follows by noting that

$$\|\hat{\mathcal{B}} - \mathcal{B}_{\text{true}}\|_F \leq \|\hat{\Theta} - \Theta^{\text{true}}\|_F \prod_k \sigma_{\min}^{-1}(\boldsymbol{X}_k) \leq C \sqrt{\sum_k p_k},$$

where $C = C(\boldsymbol{r}, \alpha, K, c_1, c_2) > 0$ is a constant that does not depend on the dimensions $\{d_k\}$ and $\{p_k\}$. □

Proposition 1 (sub-Gaussian tensors)**.** *Let $\mathcal{S}$ be a sub-Gaussian-(σ) tensor of dimension $(d_1, \ldots, d_K)$, and $\boldsymbol{X}_k \in \mathbb{R}^{p_k \times d_k}$ be non-random matrices for all $k \in [K]$. Then $\mathcal{E} = \mathcal{S} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K$ is a sub-Gaussian-(σ') tensor of dimension $(p_1, \ldots, p_K)$, where $\sigma' \leq \sigma \prod_k \sigma_{\max}(\boldsymbol{X}_k)$. Here $\sigma_{\max}(\cdot)$ denotes the largest singular value of the matrix.*

Proof. To show $\mathcal{E}$ is a sub-Guassian tensor, it suffices to show that the $\mathcal{E} \times_1 \boldsymbol{u}_1^T \times_2 \cdots \times_K \boldsymbol{u}_K^T$ is a sub-Gaussian scalar with parameter σ', for any unit-1 vector $\boldsymbol{u}_k \in \mathbb{R}^{p_k}$, $k \in [K]$.

Note that,

$$\begin{aligned}
\mathcal{E} \times_1 \boldsymbol{u}_1^T \times_2 \cdots \times_K \boldsymbol{u}_K^T &= \mathcal{S} \times_1 (\boldsymbol{u}_1^T \boldsymbol{X}_1) \times_2 \cdots \times_K (\boldsymbol{u}_K^T \boldsymbol{X}_K) \\
&= \left(\prod_k \|\boldsymbol{u}_k^T \boldsymbol{X}_k\|_2 \right) \underbrace{\left[\mathcal{S} \times_1 \frac{(\boldsymbol{u}_1^T \boldsymbol{X}_1)}{\|(\boldsymbol{u}_1^T \boldsymbol{X}_1)\|_2} \times_2 \cdots \times_K \frac{(\boldsymbol{u}_K^T \boldsymbol{X}_K)}{\|(\boldsymbol{u}_K^T \boldsymbol{X}_K)\|_2} \right]}_{\text{sub-Gaussian-}\sigma \text{ scalar}}.
\end{aligned}$$

Because $\|(\boldsymbol{u}_k^T \boldsymbol{X}_k)\|_2 \leq \sigma_{\max}(\boldsymbol{X}_k^T) \|\boldsymbol{u}_k\|_2 = \sigma_{\max}(\boldsymbol{X}_k)$, we conclude that $\mathcal{E} \times_1 \boldsymbol{u}_1^T \times_2 \cdots \times_K \boldsymbol{u}_K^T$ is a sub-Gaussian tensor with parameter $\sigma \prod_k \sigma_{\max}(\boldsymbol{X}_k)$. □

Proposition 2 (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* $(i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]$.

Proof. The proof is similar to Lemma 3 in (Fan et al., 2019). 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. Therefore, ε is sub-Gaussian-(ϕU). □

Proof of Theorem 4.2. The proof is similar to (Baldin and Berthet, 2018). We sketch the main steps here for completeness. Recall that $\ell(\mathcal{B}) = \mathbb{E}(\mathcal{L}_{\mathcal{Y}}(\mathcal{B}))$. By the definition of KL divergence, we have that,

$$\begin{aligned}
\ell(\hat{\mathcal{B}}) &= \ell(\mathcal{B}_{\text{true}}) - \sum_{(i_1,\ldots,i_K)} KL(\theta_{\text{true},i_1,\ldots,i_K}, \hat{\theta}_{i_1,\ldots,i_K}) \\
&= \ell(\mathcal{B}_{\text{true}}) - \text{KL}(\mathbb{P}_{\mathcal{Y}_{\text{true}}}, \mathbb{P}_{\hat{\mathcal{Y}}}),
\end{aligned}$$

where $\mathbb{P}_{\mathcal{Y}_{\text{true}}}$ denotes the distribution of $\mathcal{Y}|\mathcal{X}$ with true parameter $\mathcal{B}_{\text{true}}$, and $\mathbb{P}_{\hat{\mathcal{Y}}}$ denotes the

distribution with estimated parameter $\hat{\mathcal{B}}$. Therefore

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\mathcal{Y}_{\text{true}}},\ \mathbb{P}_{\hat{\mathcal{Y}}}) &= \ell(\mathcal{B}_{\text{true}}) - \ell(\hat{\mathcal{B}}) \\
&= \frac{1}{2} \sum_{i_1 \ldots, i_K} b''(\check{\theta}_{i_1,\ldots,i_K})(\theta_{i_1,\ldots,i_K} - \theta_{\text{true},i_1,\ldots,i_K})^2 \\
&\leq \frac{U}{2} \|\Theta - \Theta^{\text{true}}\|_F^2 \\
&\leq \frac{U}{2} c_2^{2K} \|\mathcal{B} - \mathcal{B}_{\text{true}}\|_F^2,
\end{aligned}$$

where the second line comes from (12), and $c_2 > 0$ is the upper bound for the $\sigma_{\max}(\boldsymbol{X}_k)$. The result then follows from Theorem 4.1. □

B Code

The source code and data used in the paper are available at https://CRAN.R-project.org/package=tensorregress.