Exponential Family Tensor Regression with Covariates on Multiple Modes

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Higher-order tensors have recently received increasing attention in many fields across science and engineering. Here, we present an exponential family of tensor-response regression models that incorporate covariates on multiple modes. Such problems are common in neuroimaging, network modeling, and spatial-temporal analysis. We propose a rank-constrained estimator and establish the theoretical accuracy guarantees. Unlike earlier methods, our approach allows covariates from multiple tensor modes whenever available. An efficient alternating updating 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. Our approach identifies the key global connectivity pattern and pinpoints the local regions that are associated with covariates.

1 Introduction

Many contemporary scientific and engineering studies collect multi-way array data, a.k.a. tensors, accompanied by additional covariates. One example is neuroimaging analysis [**?** **?**], in which the brain connectivity networks are collected from a sample of individuals. Researchers are often interested in identifying connection edges that are affected by individual characteristics such as age, gender, and disease status (see Figure 1a). Another example is in the field of network analysis [**?** **?**]. A typical social network consists of nodes that represent people and edges that represent friendships. In addition, features on nodes and edges are often available, such as people's personality and demographic location. It is of keen scientific interest to identify the variation in the connection patterns (e.g., transitivity, community) that can be attributable to the node features.

Our contributions. This paper presents a general treatment to these seemingly different problems. We formulate the learning task as a regression problem, with tensor observation serving as a response, and the node features and/or their interactions forming the predictor. Figure 1b illustrates the general set-up we consider. The regression approach allows the identification of variation in the data tensor that is explained by the covariates. Our model greatly improves the classical tensor regression [**?** **?**] by incorporating covariates from multiple modes and the interactions thereof. The statistical convergence of our estimator is established, and we quantify the gain in predictive power.

A related contribution is that our method allows a broad range of tensor types, including continuous, count, and binary observations. While previous tensor regression methods [**?** **?**] are able to analyze Gaussian responses, none of them is suitable for exponential distribution family of tensors. We develop a generalized tensor regression framework, and as a by product, our models allows heteroscedasticity by relating the variance of tensor entry to its mean. This flexibility is particularly important in practice, because social network, brain imaging, or gene expression datasets are often non-Gaussian.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

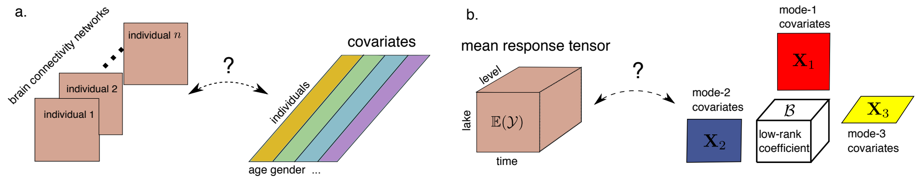

Figure 1: Examples of tensor response regression model with covariates on multiple modes. (a) Network population model. (b) Spatial-temporal growth model.

Related work. Our work is closely related to but also clearly distinctive from several lines of previous work. The first is a class of *unsupervised* tensor decomposition such as Tucker and CP decomposition [**? ? ?**]. Tucker decomposition is an unsupervised method that finds a low-rank representation of a data tensor. In contrast, our model is a *supervised* tensor model that identifies the association between a data tensor and multiple covariates. The low-rank structure is determined jointly by the tensor response and matrix covariates. The second line of work studies the network-response model [**? ?**]. Earlier development of this model focuses mostly on binary data in the presence of dyadic covariates [**?**]. We will demonstrate the enhanced accuracy as the order of data grows, and establish the general theory for exponential family which is arguably better suited to various data types.

2 Preliminaries

We begin by reviewing the basic properties about tensors [**?**]. We use $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_K}$ to denote an order-K $(d_1,\ldots,d_K)$-dimensional tensor. 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 shorthand notion $\mathcal{Y} \times \{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K\}$ 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 of tensor $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_F = \langle \mathcal{Y},\ \mathcal{Y} \rangle^{1/2}$. 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 Tucker 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 $\text{Unfold}_k(\mathcal{Y})$, $k = 1,\ldots,K$. We use lower-case letters (e.g., a, b, c) for scalars/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, $[d]$ denote the d-set $\{1,\ldots,d\}$, and allow an $\mathbb{R} \to \mathbb{R}$ function to be applied to tensors in an element-wise manner.

3 Motivation and model

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 we observe covariates on some of the K modes. Let $\boldsymbol{X}_k \in \mathbb{R}^{d_k\times p_k}$ denote the available covariates on the mode k, where $p_k \leq d_k$. We propose a multilinear structure on the conditional expectation of the tensor. Specifically,

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

where $f(\cdot)$ is a known link function, $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is the linear predictor, $\mathcal{B} \in \mathbb{R}^{p_1\times\cdots\times p_K}$ is the parameter tensor of interest, and $\times$ denotes the tensor Tucker product. The choice of link function depends on the distribution of the response data. Some common choices are identity link for Gaussian tensor, logistic link for binary tensor, and $\exp(\cdot)$ link for Poisson tensor (see Table 1).

We give three examples of tensor regression that arise in practice.

Example 1 (Spatio-temporal growth model)**.** 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 p

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.

types, with q 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 r and that the expected trend in time is a polynomial of order s. Then, the spatio-temporal growth model can be represented as

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

where $\mathcal{B}\in\mathbb{R}^{p\times(r+1)\times(s+1)}$ is the coefficient tensor of interest, $\boldsymbol{X}_1=\text{blockdiag}\{\mathbf{1}_q,\ldots,\mathbf{1}_q\}\in\{0,1\}^{d\times p}$ is the design matrix for lake types,

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

are the design matrices for spatial and temporal effects, respectively. The model (2) is a higher-order extension of the "growth curve" model originally proposed for matrix data [**? ? ?**]. Clearly, the spatial-temporal model is a special case of our tensor regression model, with covariates available on each of the three modes.

Example 2 (Network population model)**.** Network response model is recently developed in the context of neuroimanig analysis. The goal is to study the relationship between network-valued response and the individual covariates. Suppose we observe n i.i.d. observations $\{(\boldsymbol{Y}_i,\boldsymbol{x}_i):i=1,\ldots,n\}$, where $\boldsymbol{Y}_i\in\{0,1\}^{d\times d}$ is the brain connectivity network on the i-th individual, and $\boldsymbol{x}_i\in\mathbb{R}^p$ is the individual covariate such as age, gender, cognition, etc. The network-response model [**? ?**] has the form

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

where $\mathcal{B}\in\mathbb{R}^{d\times d\times p}$ is the coefficient tensor of interest. The model (3) is a special case of our tensor-response model, with covariates on the last mode of the tensor. Specifically, stacking $\{\boldsymbol{Y}_i\}$ together yields an order-3 response tensor $\mathcal{Y}\in\{0,1\}^{d\times d\times n}$, along with covariate matrix $\boldsymbol{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]^T\in\mathbb{R}^{n\times p}$. Then, the model (3) can be written as

$$\text{logit}(\mathbb{E}(\mathcal{Y}|\boldsymbol{X}))=\mathcal{B}\times_3\boldsymbol{X}=\mathcal{B}\times\{\boldsymbol{I}_d,\boldsymbol{I}_d,\boldsymbol{X}\}.$$

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 networks and graphs. Let $\mathcal{G}=(V,E)$ denote a network, 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 covariate $\boldsymbol{x}_i\in\mathbb{R}^p$ 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

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

The above model has demonstrated its success in modeling transitivity, balance, and communities in the networks [**?**]. We show that our tensor regression model (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

$$\text{logit}(\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}))=\mathcal{B}\times\{\boldsymbol{X},\boldsymbol{X}\}.$$

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 the covariates. The regression coefficient $\mathcal{B}$ in our model model (1) serves this goal by collecting the effects of covariates and the interaction thereof. To encourage the sharing among effects, we assume that the coefficient tensor $\mathcal{B}$ lies in a low-dimensional parameter space:

$$\mathcal{P}_{r_1,\ldots,r_K}=\{\mathcal{B}\in\mathbb{R}^{p_1\times\cdots\times p_K}:r_k(\mathcal{B})\le r_k\text{ for all }k\in[K]\},$$

where $r_k(\mathcal{B}) \leq p_k$ is the Tucker rank at mode k of the tensor. The low-rank assumption is plausible in many scientific applications. In 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. When no confusion arises, we drop the subscript $(r_1, \ldots, r_K)$ and write $\mathcal{P}$ for simplicity.

Our tensor regression model is able to incorporate covariates on any subset of modes, whenever available. Without loss of generality, we denote by $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$ the covariates in all modes and treat $\boldsymbol{X}_k = \boldsymbol{I}_{d_k}$ if the mode-k has no (informative) covariate. Then, the final form of our tensor regression model can be written as:

$$\mathbb{E}(\mathcal{Y}|\mathcal{X}) = f(\Theta), \quad \Theta = \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}, \quad \text{where rank}(\mathcal{B}) \leq (r_1, \ldots, r_K), \tag{5}$$

where the entries of $\mathcal{Y}$ are independent r.v.'s conditional on $\mathcal{X}$, and $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the low-rank coefficient tensor of interest. We comment that other forms of tensor low-rankness are also possible, and here we choose Tucker rank just for parsimony. Similar models can be derived using various notions of low-rankness based on CP decomposition [**?**] and train decomposition [**?**].

4 Rank-constrained likelihood-based estimation

We develop a likelihood-based procedure to estimate the coefficient tensor $\mathcal{B}$ in (5). 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 covariate $\boldsymbol{x}$, 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 distributions.

We model the entries in the response tensor y_{ijk} conditional on θ_{ijk} as independent draws from an exponential family. The quasi log-likelihood of (5) is equal (ignoring constant) to Bregman distance between $\mathcal{Y}$ and $b'(\Theta)$:

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

We assume that we have an additional information on an upper bound $\alpha > 0$ such that $\|\Theta\|_\infty \leq \alpha$. This is the case for many applications we have in mind such as brain network analysis where fiber connections are bounded. We propose a constrained maximum likelihood estimator (MLE) for the coefficient tensor:

$$\hat{\mathcal{B}} = \mathop{\arg\max}_{\text{rank}(\mathcal{B}) \leq \boldsymbol{r}, \|\Theta(\mathcal{B})\|_\infty \leq \alpha} \mathcal{L}_{\mathcal{Y}}(\mathcal{B}). \tag{6}$$

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 5.2.

4.1 Statistical properties

We assess the estimation accuracy using the deviation in the Frobenius norm. 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 covariates 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 $p_k/d_k \to \gamma_k \in [0, 1)$. 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 positive constants $c_1, c_2 > 0$ such that $c_1 \le \sigma_{\min}(\boldsymbol{X}_k) \le \sigma_{\max}(\boldsymbol{X}_k) \le c_2$ for all $k \in [K]$. Here $\sigma_{min}(\cdot)$ and $\sigma_{max}(\cdot)$ denotes the smallest and largest singular values, respectively.

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

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

The assumptions are fairly mild. Assumption A1 guarantees the non-singularity of the covariates, 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 $\text{Var}(y_{i_1,\ldots,i_K}|\mathcal{X}, \mathcal{B}) = \phi b''(\theta_{i_1,\ldots,i_K})$ when $y_{i_1,\ldots,i_K}$ belongs to an exponential family [**?**].

Theorem 4.1 (Statistical convergence)**.** *Consider a generalized tensor regression model with covariates on multiple modes $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$. Suppose the entries in $\mathcal{Y}$ are independent realizations of an exponential family distribution, and $\mathbb{E}(\mathcal{Y}|\mathcal{X})$ follows the low-rank tensor regression model* (5)*. Under Assumption 1, there exist constants $C_1, C_2 > 0$, such that, with probability at least $1 - \exp(-C_1 \sum_k p_k)$,*

$$\mathit{Loss}(\mathcal{B}_{\mathit{true}},\ \hat{\mathcal{B}}) \le \frac{C_2 \prod_k r_k}{\max_k r_k} \sum_k p_k, \tag{7}$$

where $C_2 = C_2(\alpha, K) > 0$ is a constant that does not depend on $\{d_k\}$, $\{r_k\}$, and $\{p_k\}$.

To gain insight on the bound (7), we consider a special case when tensor dimensions are equal at every mode, i.e., $d_k = d$, $p_k = \gamma d$, $\gamma \in [0, 1)$ for all $k \in [K]$, and the covariates $\boldsymbol{X}_k$ are Gaussian design matrices with i.i.d. $N(0, 1)$ entries. To put the context in the framework of Theorem 4.1, we rescale the covariates into $\check{\boldsymbol{X}}_k = d^{-1/2}\boldsymbol{X}_k$ so that the singular values of $\check{\boldsymbol{X}}_k$ are bounded by $1 \pm \sqrt{\gamma}$. The result (7) implies that the estimated coefficient has a convergence rate $\mathcal{O}(p/d^K)$ in the scale of the original covariates $\boldsymbol{X}_k$. 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, assume that the singular values of the d_k-by-p_k covariate matrix $\boldsymbol{X}_k$ are bounded by $\sqrt{d_k}$. In the network model, for example, the coefficient tensor estimate converges at the rate $\mathcal{O}((2d + p)/d^2 n)$. In the dyadic data model, the coefficient matrix estimate converges at the rate $\mathcal{O}\left(p/d^2\right)$. The estimates achieve consistency as the dimension 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 $\mathbb{P}_{\mathcal{Y}_{\mathit{true}}}$ and $\mathbb{P}_{\hat{\mathcal{Y}}}$ denote the distributions of $\mathcal{Y}$ given the true parameter $\mathcal{B}_{\mathit{true}}$ and estimated parameter $\hat{\mathcal{B}}$, respectively. Then, we have, with probability at least $1 - \exp(C_1 \sum_k p_k)$,*

$$\mathit{KL}(\mathbb{P}_{\mathcal{Y}_{\mathit{true}}},\ \mathbb{P}_{\hat{\mathcal{Y}}}) \le \frac{C_4 \prod_k r_k}{\max_k r_k} \sum_k p_k,$$

where $C_4 = C_4(\alpha, K) > 0$ is a constant that do not depend on $\{d_k\}$, $\{r_k\}$, and $\{p_k\}$.

5 Numerical implementation

5.1 Alternating optimization

We propose an alternating optimization to solve the non-convex problem (6). Briefly, we utilize the Tucker factor representation of the coefficient tensor $\mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K\}$, where $\mathcal{C}$ is a core tensor with a given rank (to be specified in the next subsection), and $\boldsymbol{M}_k$ are column-wise orthogonal matrices of coherent dimensions. The optimization (6) is equivalent to $(\hat{\mathcal{C}}, \{\hat{\boldsymbol{M}}_k\}) = \arg\max \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K)$ where

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

The alternating algorithm proceeds by literately updating one block at a time while keeping others fixed. We summarize the consistency property of the algorithm below.

Proposition 1 (Local convergence)**.** *Assume the solution to each block update in the alternating optimization exists and is unique. Let $\mathcal{B}^* = \mathcal{C}^* \times \{\boldsymbol{M}_1^*, \ldots, \boldsymbol{M}_K^*\}$ be a local maximizer of $\mathcal{L}_{\mathcal{Y}}$, and assume the Hessian is strictly negative definite with respect to the block variables module the orthogonal transformation of $\boldsymbol{M}_k^*$. Then, the sequence $\mathcal{B}^{(t)} = \mathcal{C}^{(t)} \times \{\boldsymbol{M}_1^{(t)}, \ldots, \boldsymbol{M}_K^{(t)}\}$ generated by the alternating algorithm linearly converges to $\mathcal{B}^*$; i.e.*

$$\|\mathcal{B}^{(t)} - \mathcal{B}^*\|_F^2 \leq \rho^t(\|\mathcal{C}^{(0)} - \mathcal{C}^*\|_F^2 + \textstyle\sum_k \|\boldsymbol{M}_k^{(0)} - \boldsymbol{M}_k^*\|_F^2),$$

for initialization $(\mathcal{C}^{(0)}, \{\boldsymbol{M}_k^{(0)}\})$ sufficiently close to $(\mathcal{C}^*, \{\boldsymbol{M}_k^*\})$. Here $t \in \mathbb{N}_+$ is the iteration number and $\rho \in (0,1)$ is a contraction parameter (specified in the supplement).

Furthermore, under mild conditions, our algorithm enjoys global convergence; i.e. any sequence of iterates generated by the alternating algorithm converges to a stationary point of $\mathcal{L}_{\mathcal{Y}}$. Although a stationary point is not guaranteed to be an optimum (it could be a saddle point), our numerical experiments have suggested high-quality solutions by multiple randomized initializations (see Section 6) .

5.2 Rank selection

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.

$$\hat{\boldsymbol{r}} = \mathop{\arg\min}_{\boldsymbol{r}=(r_1,\ldots,r_K)} \text{BIC}(\boldsymbol{r}) = \mathop{\arg\min}_{\boldsymbol{r}=(r_1,\ldots,r_K)} [-2\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) + p_e(\boldsymbol{r}) \log(\prod_k d_k)],$$

where $p_e(\boldsymbol{r}) \stackrel{\text{def}}{=} \sum_k (p_k - r_k) r_k + \prod_k r_k$ 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.

6 Simulation

We evaluate the empirical performance of our generalized tensor regression through simulations. We consider order-3 tensors from three probabilistic models: (a) continous entries $y_{ijk} \sim N(\alpha u_{ijk}, 1)$; (b) count entries $y_{ijk} \sim \text{Poisson}(e^{\alpha u_{ijk}})$; (c) binary entries $y_{ijk} \sim \text{Bernoulli}(e^{\alpha u_{ijk}}/(1 + e^{\alpha u_{ijk}}))$. Here $\alpha > 0$ is a scalar controlling the magnitude of the effect size, and $\mathcal{U} = [\![u_{ijk}]\!] = \mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$ is the linear predictor. In each simulation study, we report the mean squared error (MSE) for the coefficient tensor averaged across $n_{\text{sim}} = 30$ replications.

6.1 Finite-sample performance

The experiment I evaluates the accuracy when covariates are available on all modes. We set $\alpha = 10, d_k = d, p_k = 0.4 d_k, r_k = r \in \{2, 4, 6\}$ 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 2a plots the estimation error versus the "effective sample size", d^2, under three 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 Figures 2b-c.

The experiment II investigates the capability of our model 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 $p = 5$ covariates for the each of the 50 individuals. These covariates may represent, for example, age, 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 [**?**] 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 covariate 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 [**?**].

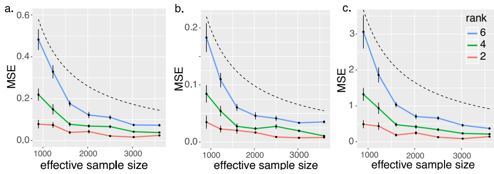

Figure 2: Mean squared error (MSE) against effective sample size. Responses are generated from Gaussian, Poisson and Bernoulli models. The dashed curves correspond to $\mathcal{O}(1/d^2)$.

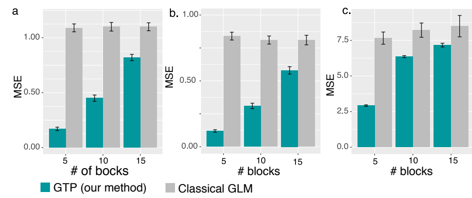

Figure 3: MSE when the networks have block structure. Responses are generated from Gaussian, Poisson and Bernoulli models. The x-axis represents the number of blocks in the networks.

Figure 3 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 covariates, 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 3, our 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.

6.2 Comparison with alternative methods

We compare our generalized tensor regression (**GTR**) with three other supervised tensor methods: Higher-order low-rank regression (**HOLRR** [**?**]), Higher-order partial least square (**HOPLS** [**?**]) and Subsampled tensor projected gradient (**TPG** [**?**]). All the three methods allow only Gaussian data, whereas ours is applicable to arbitrary exponential family distributions. For fair comparison, we consider only Gaussian response in the simulation. We measure the accuracy using mean squared prediction error, MSPE $= (\sum_k d_k)^{-1/2} \|\hat{\mathcal{Y}} - \mathbb{E}(\mathcal{Y}|\mathcal{X})\|_F$, where $\hat{\mathcal{Y}}$ is the fitted value from each method. We use similar simulation setups as in our experiment II, but consider combinations of rank $\boldsymbol{r} = (3,3,3)$ (low) vs. $(4,5,6)$ (high), noise $\sigma = 1/4$ (low) vs. $1/2$ (high), and dimension d ranging from 20 to 100 for modes with covariates, $d = 20$ for modes without covariates.

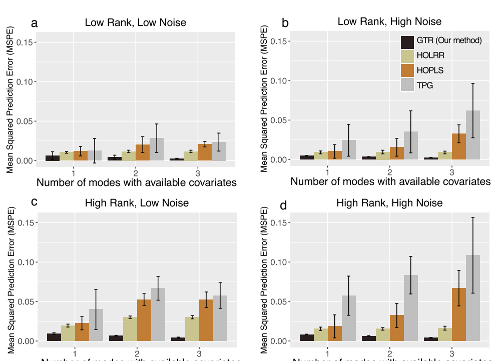

Figure 4: Comparison of MSPE versus the number of modes with covariates. Four combinations of rank/signal settings are considered.

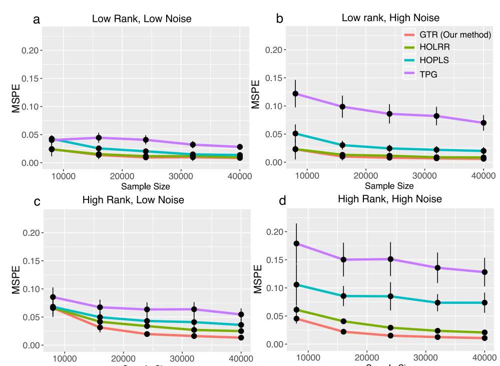

Figure 5: Comparison of MSPE versus sample size. Four combinations of rank/signal settings are considered.

Figure 4 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 covariates) increases, the **GTR** exhibits a reduction in error whereas others have increased errors. This showcases the benefit toward prediction via incorporation of multiple covariates. The accuracy gain in Figure 4 demonstrates the benefit of alternating algorithm – having informative modes also improves the estimation along non-informative modes.

Figure 5 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. Note that a higher rank implies a higher inter-mode complexity, and the results suggest the adaptation of our **GTR** method to richer models.

7 Data analysis

We apply our method to two real datasets. The first application concerns the brain network modeling in response to individual attributes (i.e. covariate on one mode), and the second application focuses on multi-relational network analysis with dyadic attributes (i.e. covariates on two modes).

We fit the tensor regression model to the Human connectome project (HCP, [**?**]) data. The HCP aims to build a network map that characterizes the anatomical and functional connectivity within healthy human brains. We take 136 individuals' brain structural networks. Each brain network is represented as a binary matrix, where the entries encode the presence or absence of fiber connections between 68 brain regions. We consider four individual-covariates: gender, age 22-25, age 26-30, and age 31+. The BIC suggests a rank $\boldsymbol{r} = (10, 10, 4)$. Figure 6 shows the top edges with high effect size, overlaid on the Desikan atlas brain template [**?**]. We depict only the top 3% edges whose connections are non-constant across samples. Figure 6a shows that the global connection exhibits clear spatial separation, and that the nodes within each hemisphere are more densely connected with each other. In particular, the superior-temporal (*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 6b). This is in agreement with a recent study showing that female brains are optimized for inter-hemispheric communication [**?**].

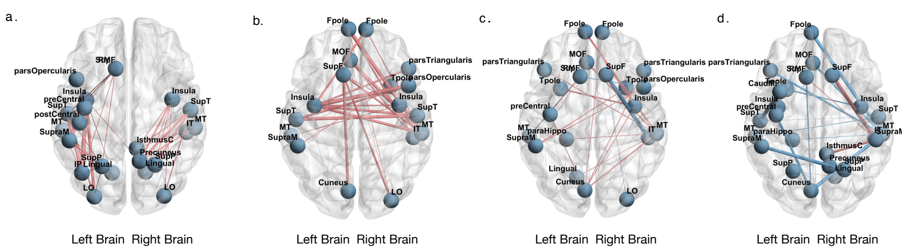

Figure 6: Top edges with large effects. Red edges refer relatively strong connections and blue edges refer relatively weak connections. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+.

The second application examines the multi-relational network analysis with node-level attributes. We consider *Nations* dataset [**?**] 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. Our tensor regression results show 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; and Cluster IV represents the Cold War alliance blocs. Detailed results and analyses are in supplements.

8 Conclusion

We have developed a generalized tensor regression with covariates 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.

Broader impact

Our exponential family tensor regression is widely applicable to spatial-temporal model, network analysis, dyadic data analysis, and recommendation systems. We have shown the improved predictive power and enhanced interpretability by incorporating interactions between multiple covariate matrices. We acknowledge the opportunities and challenges in the tensor regression modeling. We hope the method developed in this paper will open up new research directions towards richer tensor-based learning. All code and data are publically available, and we encourage members of the machine learning community to participate in these exciting problems.

References

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

[2] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

[3] Nicolai Baldin and Quentin Berthet. Optimal link prediction with matrix logistic regression. *arXiv preprint arXiv:1803.07054*, 2018.

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

[5] Guillaume Rabusseau and Hachem Kadri. Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems*, pages 1867–1875, 2016.

[6] Rose Yu and Yan Liu. Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381, 2016.

[7] Qibin Zhao, Cesar F Caiafa, Danilo P Mandic, Zenas C Chao, Yasuo Nagasaka, Naotaka Fujii, Liqing Zhang, and Andrzej Cichocki. Higher order partial least squares (HOPLS): a generalized multilinear regression method. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1660–1673, 2012.

[8] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[9] David Hong, Tamara G Kolda, and Jed A Duersch. Generalized canonical polyadic tensor decomposition. *SIAM Review, in press. arXiv:1808.07452*, 2019.

[10] Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 2018.

[11] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.

[12] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[13] Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520:44–66, 2017.

[14] K Ruben Gabriel. Generalised bilinear regression. *Biometrika*, 85(3):689–700, 1998.

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

[16] Muni S Srivastava, Tatjana von Rosen, and Dietrich Von Rosen. Models with a kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17(4):357–370, 2008.

[17] Jingfei Zhang, Will Wei Sun, and Lexin Li. Network response regression for modeling population of networks with covariates. *arXiv preprint arXiv:1810.03192*, 2018.

[18] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

[19] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

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

[21] Lucy F Robinson, Lauren Y Atlas, and Tor D Wager. Dynamic functional connectivity using state-based dynamic community structure: Method and application to opioid analgesia. *NeuroImage*, 108:274–291, 2015.

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

[23] Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019). arXiv:1906.03807*, 2019.

[24] Linda Geddes. Human brain mapped in unprecedented detail. *Nature*, 2016.

[25] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

[26] Madhura Ingalhalikar, Alex Smith, Drew Parker, Theodore D Satterthwaite, Mark A Elliott, Kosha Ruparel, Hakon Hakonarson, Raquel E Gur, Ruben C Gur, and Ragini Verma. Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828, 2014.

[27] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*, volume 11, pages 809–816, 2011.