

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
054

---

# Exponential Family Tensor Regression with Covariates on Multiple Modes

---

## 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 multi-relational social network data and diffusion tensor imaging data from human connection project. 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 (Sun & Li, 2017; Zhou et al., 2013), 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 (Baldin & Berthet, 2018; Hoff, 2005). 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.

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. In contrast to earlier work (Rabusseau & Kadri, 2016; Yu & Liu, 2016), our method allows the covariates from multiple modes, whenever available. We utilize a low-rank constraint in the regression coefficient to encourage the sharing among tensor entries. The statistical convergence of our estimator is established, and we quantify the gain in predictive power by taking multiple covariates into account.

A secondary contribution is that our method allows a broad range of tensor types, including continuous, count, and binary observations. While previous tensor regression methods (Zhao et al., 2012; Yu & Liu, 2016) 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.

**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 (De Lathauwer et al., 2000; Hong et al., 2019; Zhang & Xia, 2018) that aims to find a low-rank representation of a data tensor. In contrast, our model can be viewed a *supervised* tensor learning, which aims to identify the association between a data tensor and covariates. The second related line (Zhou et al., 2013; Chen et al., 2019) tackles tensor regression where the response is a scalar and the *predictor* is a tensor. Our proposal is orthogonal to theirs because we treat the tensor as a *response*. The tensor-response model is appealing for high-dimensional analysis when both the response and the covariate dimensions grow. The last line of work studies the network-response model (Rabusseau & Kadri, 2016; Li & Zhang, 2017). The earlier development of this model focuses mostly on binary data in the presence of dyadic covariates (Hoff, 2005). 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.

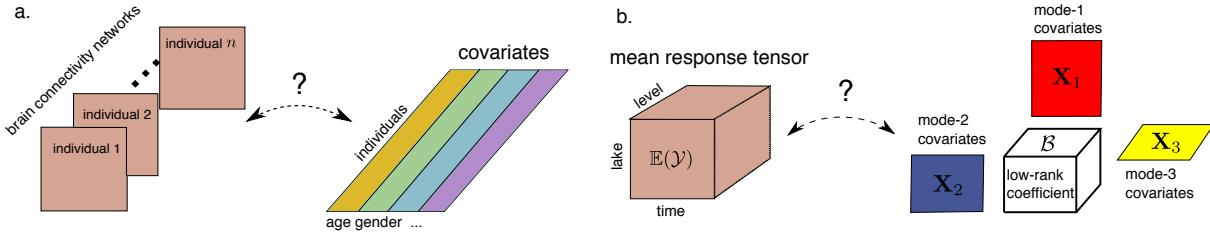


Figure 1. Examples of tensor response regression model with covariates on multiple modes. (a) Network population model. (b) Spatial-temporal growth model.

## 2. Preliminaries

We begin by reviewing the basic properties about tensors (Kolda & Bader, 2009). We use  $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!] \in \mathbb{R}^{d_1 \times \dots \times d_K}$  to denote an order- $K$  ( $d_1, \dots, d_K$ )-dimensional tensor. The multilinear multiplication of a tensor  $\mathcal{Y} \in \mathbb{R}^{d_1 \times \dots \times d_K}$  by matrices  $\mathbf{X}_k = [\![x_{i_k, j_k}^{(k)}]\!] \in \mathbb{R}^{p_k \times d_k}$  is defined as

$$\mathcal{Y} \times_1 \mathbf{X}_1 \dots \times_K \mathbf{X}_K = [\! [ \sum_{i_1, \dots, i_K} y_{i_1, \dots, i_K} x_{j_1, i_1}^{(1)} \dots x_{j_K, i_K}^{(K)} ]\!],$$

which results in an order- $K$  ( $p_1, \dots, p_K$ )-dimensional tensor. For ease of presentation, we use shorthand notion  $\mathcal{Y} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\}$  to denote the tensor-by-matrix product. For any two tensors  $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!]$ ,  $\mathcal{Y}' = [\![y'_{i_1, \dots, i_K}]\!]$  of identical order and dimensions, their inner product is defined as  $\langle \mathcal{Y}, \mathcal{Y}' \rangle = \sum_{i_1, \dots, i_K} y_{i_1, \dots, i_K} y'_{i_1, \dots, i_K}$ . 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 (Wang et al., 2017). 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  $\mathbf{r} = (r_1, \dots, r_K)$ , where  $r_k$  is the rank of matrix  $\text{Unfold}_k(\mathcal{Y})$ ,  $k = 1, \dots, K$ . We use lower-case letters (e.g.,  $a, b, c$ ) for scalars/vectors, upper-case boldface letters (e.g.,  $\mathbf{A}, \mathbf{B}, \mathbf{C}$ ) for matrices, and calligraphy letters (e.g.,  $\mathcal{A}, \mathcal{B}, \mathcal{C}$ ) for tensors of order three or greater. We let  $\mathbf{I}_d$  denote the  $d \times d$  identity matrix,  $[d]$  denote the  $d$ -set  $\{1, \dots, d\}$ , and allow an  $\mathbb{R} \rightarrow \mathbb{R}$  function to be applied to tensors in an element-wise manner.

## 3. Motivation and model

Let  $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!] \in \mathbb{R}^{d_1 \times \dots \times d_K}$  denote an order- $K$  data tensor. Suppose we observe covariates on some of the  $K$  modes. Let  $\mathbf{X}_k \in \mathbb{R}^{d_k \times p_k}$  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,

$$\mathbb{E}(\mathcal{Y} | \mathbf{X}_1, \dots, \mathbf{X}_K) = f(\Theta), \text{ with} \quad (1)$$

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 $f(\theta)$	$\theta$	$\exp(\theta)$	$(1 + \exp(-\theta))^{-1}$

Table 1. Canonical links for common distributions.

$$\Theta = \mathcal{B} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\},$$

where  $f(\cdot)$  is a known link function,  $\Theta \in \mathbb{R}^{d_1 \times \dots \times d_K}$  is the linear predictor,  $\mathcal{B} \in \mathbb{R}^{p_1 \times \dots \times p_K}$  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 concrete 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  $p$  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

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

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

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

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 (Gabriel, 1998; Potthoff & Roy, 1964; Srivastava et al.,

2008). 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 neuroimaging analysis. The goal is to study the relationship between network-valued response and the individual covariates. Suppose we observe  $n$  i.i.d. observations  $\{(\mathbf{Y}_i, \mathbf{x}_i) : i = 1, \dots, n\}$ , where  $\mathbf{Y}_i \in \{0, 1\}^{d \times d}$  is the brain connectivity network on the  $i$ -th individual, and  $\mathbf{x}_i \in \mathbb{R}^p$  is the individual covariate such as age, gender, cognition, etc. The network-response model (Rabusseau & Kadri, 2016; Zhang et al., 2018) has the form

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

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  $\{\mathbf{Y}_i\}$  together yields an order-3 response tensor  $\mathcal{Y} \in \{0, 1\}^{d \times d \times n}$ , along with covariate matrix  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ . Then, the model (3) can be written as

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

**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  $\mathbf{x}_i \in \mathbb{R}^p$  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

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

The above model has demonstrated its success in modeling transitivity, balance, and communities in the networks (Hoff, 2005). 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  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ . Then, the graph model (4) can be expressed as

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

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:

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

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  $(r_1, \dots, r_K)$  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  $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_K\}$  the covariates in all modes and treat  $\mathbf{X}_k = \mathbf{I}_{d_k}$  if the mode- $k$  has no (informative) covariate. Then, the final form of our tensor regression model can be written as:

$$\begin{aligned} \mathbb{E}(\mathcal{Y} | \mathcal{X}) &= f(\Theta), \quad \Theta = \mathcal{B} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\}, \\ \text{where } \text{rank}(\mathcal{B}) &\leq (r_1, \dots, r_K), \end{aligned} \quad (5)$$

where the entries of  $\mathcal{Y}$  are independent r.v.'s conditional on  $\mathcal{X}$ , and  $\mathcal{B} \in \mathbb{R}^{p_1 \times \dots \times p_K}$  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 (Hitchcock, 1927) and train decomposition (Oseledets, 2011).

## 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  $\mathbf{x}$ , the density is expressed as:

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

where  $b(\cdot)$  is a known function,  $\theta$  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.

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

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) = \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1, \dots, i_K} b(\theta_{i_1, \dots, i_K}),$$

where  $\Theta = \mathcal{B} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\}$ .

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:

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

In the following theoretical analysis, we assume the rank  $r = (r_1, \dots, r_K)$  is known and fixed. The adaptation of unknown  $r$  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.  $d_k \rightarrow \infty$  and  $p_k \rightarrow \infty$ , while  $\frac{p_k}{d_k} \rightarrow \gamma_k \in [0, 1]$ . 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  $c_1 \leq \sigma_{\min}(\mathbf{X}_k) \leq \sigma_{\max}(\mathbf{X}_k) \leq c_2$  for all  $k \in [K]$ . Here  $\sigma_{\min}(\cdot)$  and  $\sigma_{\max}(\cdot)$  denotes the smallest and largest singular values, respectively.*
- A2. *There exist positive constants  $L, U > 0$  such that  $L\phi \leq \text{Var}(y_{i_1, \dots, i_K} | \theta_{i_1, \dots, i_K}) \leq U\phi$  for all  $|\theta_{i_1, \dots, i_K}| \leq \alpha$ .*
- A2'. *Equivalently, there exists two positive constants  $L, U > 0$  such that  $L \leq b''(\theta) \leq U$  for all  $|\theta| \leq \alpha$ , where  $\alpha$  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  $\Theta$ . Assumption A2 and A2' are equivalent, because  $\text{Var}(y_{i_1, \dots, i_K} | \mathcal{X}, \mathcal{B}) = \phi b''(\theta_{i_1, \dots, i_K})$  when  $y_{i_1, \dots, i_K}$  belongs to an exponential family (McCullagh & Nelder, 1989).

**Theorem 4.1** (Statistical convergence). *Consider a generalized tensor regression model with covariates on multiple modes  $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_K\}$ . 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 two constants  $C_1, C_2 > 0$ , such that, with probability at least  $1 - \exp(-C_1 \sum_k p_k)$ ,*

$$\text{Loss}(\mathcal{B}_{\text{true}}, \hat{\mathcal{B}}) \leq C_2 \sum_k p_k. \quad (7)$$

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

To gain further insight on the bound (7), 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 covariates  $\mathbf{X}_k$  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  $\check{\mathbf{X}}_k = \frac{1}{\sqrt{d}} \mathbf{X}_k$  so that the singular values of  $\check{\mathbf{X}}_k$  are bounded by  $1 \pm \sqrt{\gamma}$ . The result in (7) implies that the estimated coefficient has a convergence rate  $\mathcal{O}(\frac{p}{d^K})$  in the scale of the original covariates  $\{\mathbf{X}_k\}$ . 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- $p_k$  covariate matrix  $\mathbf{X}_k$  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  $\mathcal{O}(\frac{p+r+s}{d^{mn}})$  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-covariate tensor converges at the rate  $\mathcal{O}(\frac{2d+p}{d^2n})$  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 covariate-by-covariate matrix converges at the rate  $\mathcal{O}(\frac{p}{d^2})$  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.

220 **Theorem 4.2** (Prediction error). Assume the same set-up as  
 221 in Theorem 4.1. Let  $\mathbb{P}_{\mathcal{Y}_{true}}$  and  $\mathbb{P}_{\hat{\mathcal{Y}}}$  denote the distributions of  
 222  $\mathcal{Y}$  given the true parameter  $\mathcal{B}_{true}$  and estimated parameter  
 223  $\hat{\mathcal{B}}$ , respectively. Then, we have, with probability at least  
 224  $1 - \exp(C_1 \sum_k p_k)$ ,

$$KL(\mathbb{P}_{\mathcal{Y}_{true}}, \mathbb{P}_{\hat{\mathcal{Y}}}) \leq C_4 \sum_k p_k,$$

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

## 5. Numerical implementation

### 5.1. Alternating optimization

235 In this section, we introduce an efficient algorithm to  
 236 solve (6). The objective function  $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})$  is concave in  $\mathcal{B}$   
 237 when the link  $f$  is the canonical link function. However, the  
 238 feasible set  $\mathcal{P}$  is non-convex, and thus the optimization (6) is  
 239 a non-convex problem. We utilize a Tucker factor represen-  
 240 tation of the coefficient tensor  $\mathcal{B}$  and turn the optimization  
 241 into a block-wise convex problem.

242 Specifically, write the rank- $r$  decomposition of coefficient  
 243 tensor  $\mathcal{B}$  as

$$\mathcal{B} = \mathcal{C} \times \{\mathbf{M}_1, \dots, \mathbf{M}_K\}, \quad (8)$$

247 where  $\mathcal{C} \in \mathbb{R}^{r_1 \times \dots \times r_K}$  is a full-rank core tensor,  $\mathbf{M}_k \in$   
 248  $\mathbb{R}^{p_k \times r_k}$  are factor matrices with orthogonal columns. Esti-  
 249 mating  $\mathcal{B}$  amounts to finding both the core tensor  $\mathcal{C}$  and the  
 250 factor matrices  $\mathbf{M}_k$ 's. The optimization (6) can be written  
 251 as  $(\hat{\mathcal{C}}, \{\hat{\mathbf{M}}_k\}) = \arg \max \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K)$ , where

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K) = \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1, \dots, i_K} b(\theta_{i_1, \dots, i_K}),$$

with  $\Theta = \mathcal{C} \times \{\mathbf{M}_1 \mathbf{X}_1, \dots, \mathbf{M}_K \mathbf{X}_K\}$ .

252 The decision variables in the above objective function con-  
 253 sist of  $K + 1$  blocks of variables, one for the core tensor  
 254  $\mathcal{C}$  and  $K$  for the factor matrices  $\mathbf{M}_k$ 's. We notice that, if  
 255 any  $K$  out of the  $K + 1$  blocks of variables are known, then  
 256 the optimization with respect to the last block of variables  
 257 reduced to a simple GLM. We therefore choose to iteratively  
 258 update one block at a time while keeping others fixed. We  
 259 leverage on a block relaxation algorithm for optimization,  
 260 and the classical (local) convergence for block algorithm  
 261 applies. Although a non-convex optimization of this type  
 262 usually has no guarantee on global optimality, our numeri-  
 263 cal experiments have suggested high-quality solutions (see  
 264 Section 6). The full algorithm is described in Algorithm 1.

### 5.2. Rank selection

272 Algorithm 1 takes the rank  $r$  as an input. Estimating an  
 273 appropriate rank given the data is of practical importance.

220 We propose to use Bayesian information criterion (BIC) and  
 221 choose the rank that minimizes BIC; i.e.

$$\begin{aligned} \hat{r} &= \arg \min_{\mathbf{r}=(r_1, \dots, r_K)} \text{BIC}(\mathbf{r}) \\ &= \arg \min_{\mathbf{r}=(r_1, \dots, r_K)} [-2\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) + p_e(\mathbf{r}) \log(\prod_k d_k)], \end{aligned} \quad (9)$$

228 where  $p_e(\mathbf{r}) \stackrel{\text{def}}{=} \sum_k (p_k - r_k)r_k + \prod_k r_k$  is the effective  
 229 number of parameters in the model. We choose  $\hat{r}$  that  
 230 minimizes  $\text{BIC}(\mathbf{r})$  via grid search. Our choice of BIC aims  
 231 to balance between the goodness-of-fit for the data and the  
 232 degree of freedom in the population model. We test its  
 233 empirical performance in Section 6.

## 6. Simulation

234 We evaluate the empirical performance of our generalized  
 235 tensor regression through simulations. We consider order-  
 236 3 tensors with a range of distribution types. The coeffi-  
 237 cient tensor  $\mathcal{B}$  is generated using the factorization form (8)  
 238 where both the core and factor matrices are drawn i.i.d. from  
 239 Uniform[-1,1]. The linear predictor is then simulated from  
 240  $\mathcal{U} = \mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ , where  $\mathbf{X}_k$  is either an identity  
 241 matrix (i.e. no covariate available) or Gaussian random  
 242 matrix with i.i.d. entries from  $N(0, \sigma_k^2)$ . We set  $\sigma_k = d_k^{-1/2}$   
 243 to ensure the singular values of  $\mathbf{X}_k$  are bounded as  $d_k$  in-  
 244 creases. The  $\mathcal{U}$  is scaled such that  $\|\mathcal{U}\|_\infty = 1$ . Conditional  
 245 on the linear predictor  $\mathcal{U} = [u_{ijk}]$ , the entries in the tensor  
 246  $\mathcal{Y} = [y_{ijk}]$  are drawn independently according to one of  
 247 the following three probabilistic models:

- (a) (Gaussian). Continuous entries  $y_{ijk} \sim N(\alpha u_{ijk}, 1)$ .
- (b) (Poisson). Count entries  $y_{ijk} \sim \text{Poi}(e^{\alpha u_{ijk}})$ .
- (c) (Bernoulli). Binary entries  $y_{ijk} \sim \text{Ber}\left(\frac{e^{\alpha u_{ijk}}}{1+e^{\alpha u_{ijk}}}\right)$ .

248 Here  $\alpha > 0$  is a scalar controlling the magnitude of the  
 249 effect size. In each simulation study, we report the mean  
 250 squared error (MSE) for the coefficient tensor averaged  
 251 across  $n_{\text{sim}} = 30$  replications.

### 6.1. Finite-sample performance

252 The experiment I assesses the selection accuracy of our  
 253 BIC criterion (9). We consider the balanced situation where  
 254  $d_k = d$ ,  $p_k = 0.4d_k$  for  $k = 1, 2, 3$ . We set  $\alpha = 10$   
 255 and consider various combinations of dimension  $d$  and rank  
 256  $\mathbf{r} = (r_1, r_2, r_3)$ . For each combination, we simulate tensor  
 257 data following Gaussian, Bernoulli, and Poisson models.  
 258 We then minimize BIC using a grid search over three di-  
 259 mensions. The hyper-parameter  $\alpha$  is set to infinity in the  
 260 fitting, which essentially imposes no prior on the coefficient  
 261 magnitude. Table 2 reports the selected rank averaged over  
 262  $n_{\text{sim}} = 30$  replicates for Gaussian and Poisson models. We  
 263 found that when  $d = 20$ , the selected rank is slightly smaller

---

275           **Algorithm 1** Generalized tensor response regression with covariates on multiple modes  
 276           **Input:** Response tensor  $\mathcal{Y} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ , covariate matrices  $\mathbf{X}_k \in \mathbb{R}^{d_k \times p_k}$  for  $k = 1, \dots, K$ , target Tucker  
 277           rank  $\mathbf{r} = (r_1, \dots, r_K)$ , link function  $f$ , infinity norm bound  $\alpha$   
 278           **Output:** Low-rank estimation for the coefficient tensor  $\mathcal{B} \in \mathbb{R}^{p_1 \times \dots \times p_K}$ .  
 279           1: Calculate  $\tilde{\mathcal{B}} = \mathcal{Y} \times_1 [(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T] \times_2 \dots \times_K [(\mathbf{X}_K^T \mathbf{X}_K)^{-1} \mathbf{X}_K^T]$ .  
 280           2: Initialize the iteration index  $t = 0$ . Initialize the core tensor  $\mathcal{C}^{(0)}$  and factor matrices  $\mathbf{M}_k^{(0)} \in \mathbb{R}^{p_k \times r_k}$  via  
 281           rank- $\mathbf{r}$  Tucker approximation of  $\tilde{\mathcal{B}}$ , in the least-square sense.  
 282           3: **while** the relative increase in objective function  $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})$  is less than the tolerance **do**  
 283           4:     Update iteration index  $t \leftarrow t + 1$ .  
 284           5:     **for**  $k = 1$  to  $K$  **do**  
 285           6:         Obtain the factor matrix  $\mathbf{M}_k^{(t+1)} \in \mathbb{R}^{p_k \times r_k}$  by solving  $p_k$  separate GLMs with link function  $f$ .  
 286           7:         Update the columns of  $\mathbf{M}_k^{(t+1)}$  by Gram-Schmidt orthogonalization.  
 287           8:     **end for**  
 288           9:     Obtain the core tensor  $\mathcal{C}^{(t+1)} \in \mathbb{R}^{r_1 \times \dots \times r_K}$  by solving a GLM with  $\text{vec}(\mathcal{Y})$  as response,  $\odot_{k=1}^K [\mathbf{X}_k \mathbf{M}_k^{(t)}]$   
 289           as covariates, and  $f$  as link function. Here  $\odot$  denotes the Khatri-Rao product of matrices.  
 290           10:    Rescale the core tensor subject to the infinity norm constraint.  
 291           11:    Update  $\mathcal{B}^{(t+1)} \leftarrow \mathcal{C}^{(t+1)} \times_1 \mathbf{M}_1^{(t+1)} \times_2 \dots \times_K \mathbf{M}_K^{(t+1)}$ .  
 292           12: **end while**

---

True Rank $r$	Dimension (Gaussian tensors)		Dimension (Poisson tensors)	
	$d = 20$	$d = 40$	$d = 20$	$d = 40$
(3, 3, 3)	(2.1, 2.0, 2.0)	<b>(3, 3, 3)</b>	(2.0, 2.2, 2.1)	<b>(3, 3, 3)</b>
(4, 4, 6)	(3.2, 3.1, 5.0)	<b>(4, 4, 6)</b>	(4.0, 4.0, 5.2)	<b>(4, 4, 6)</b>
(6, 8, 8)	(5.1, 7.0, 6.9)	<b>(6, 8, 8)</b>	(5.0, 6.1, 7.1)	<b>(6, 8, 8)</b>

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.

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 experiment II evaluates the accuracy when covariates 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 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 III 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

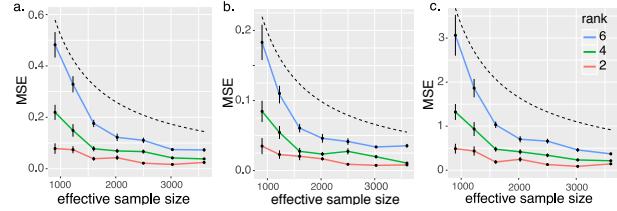


Figure 2. Mean squared error (MSE) against effective sample size. The three panels depict the MSE when the response tensors are generated from (a) Gaussian (b) Poisson and (c) Bernoulli models. The dashed curves correspond to  $\mathcal{O}(1/d^2)$ .

individuals. These covariates 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 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 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  $r$  (Wang & Zeng, 2019).

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

## Exponential family tensor regression

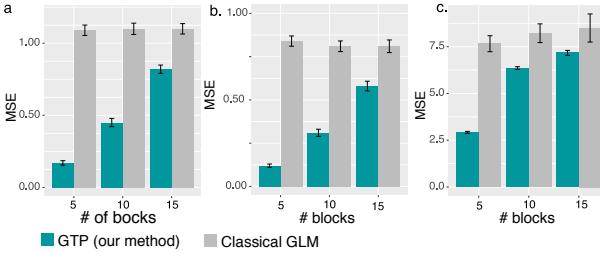


Figure 3. MSE when the networks have block structure. The three panels depict the MSE when the response tensors are generated from (a) Gaussian (b) Poisson and (c) Bernoulli models. The  $x$ -axis represents the number of blocks in the networks.

Figure 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**, Rabusseau & Kadri (2016)) is a least-square based tensor regression that allows covariates 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 covariate.
- Subsampled tensor projected gradient (**TPG**, Yu & 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 covariates 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,  $MSPE = \sqrt{\sum_k d_k} \|\hat{\mathcal{Y}} - \mathbb{E}(\mathcal{Y}|\mathcal{X})\|_F$ , where  $\hat{\mathcal{Y}}$  is the fitted value from each of the methods.

The comparison was assessed from three aspects: (a) benefit of incorporating covariates 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 ( $r = (3, 3, 3)$  vs.  $(4, 5, 6)$ ), noise ( $\sigma = 1/2$  vs.  $1/4$ ), and dimension ( $d$  ranging from 20 to

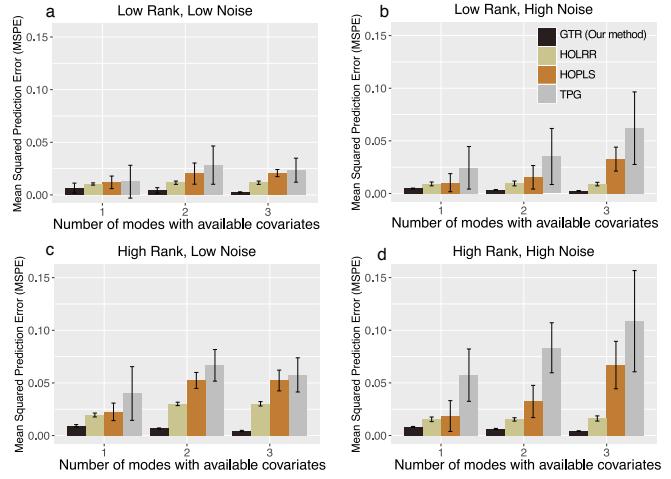


Figure 4. Comparison of MSPE versus the number of modes with covariates. We consider rank  $r = (3, 3, 3)$  (low),  $r = (4, 5, 6)$  (high), and noise  $\sigma = 1/2$  (high),  $\sigma = 1/4$  (low).

100 for modes with covariates,  $d = 20$  for modes without covariates).

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. 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 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. One possible reason is that a higher rank implies a higher inter-mode complexity, and our **GTR** method lends itself well to this context.

## 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).

## Exponential family tensor regression

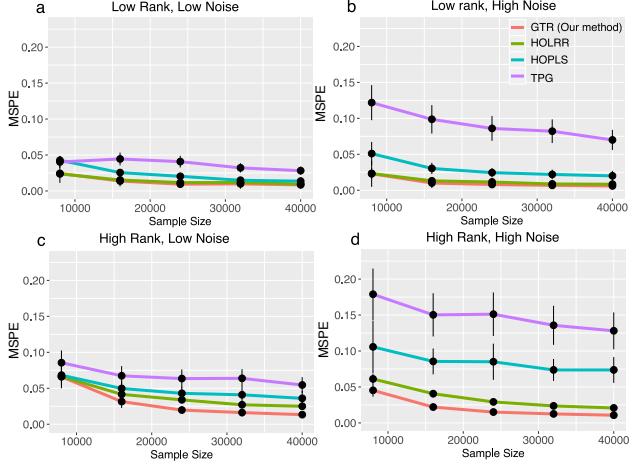


Figure 5. Comparison of MSPE versus sample size. We consider rank  $r = (3, 3, 3)$  (low),  $r = (4, 5, 6)$  (high), and noise  $\sigma = 1/2$  (high),  $\sigma = 1/4$  (low).

### 7.1. Human Connectome Project (HCP)

The Human connectome project (HCP, (Geddes, 2016)) aims to build a network map that characterizes the anatomical and functional connectivity within healthy human brains. 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 68 brain regions. We consider four individual-covariates: gender, age 22-25, age 26-30, and age 31+.

We fit the tensor regression model to the HCP data. The BIC suggests a rank  $r = (10, 10, 4)$  with log-likelihood  $\mathcal{L}_y = -174654.7$ . Figure 6 shows the top edges with high effect size, overlaid on the Desikan atlas brain template (Xia et al., 2013). We utilize the sum-to-zero contrasts in the effects coding and 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, parietal, and temporal lobes (Figure 6b). This is in agreement with a recent study showing that female brains are optimized for inter-hemispheric communication (Ingalhalikar et al., 2014). This result demonstrates the applicability of our method in detecting covariates signals.

### 7.2. Nations data

The second application examines the multi-relational network analysis with node-level attributes. We consider *Nations* dataset (Nickel et al., 2011) which records 56 relations

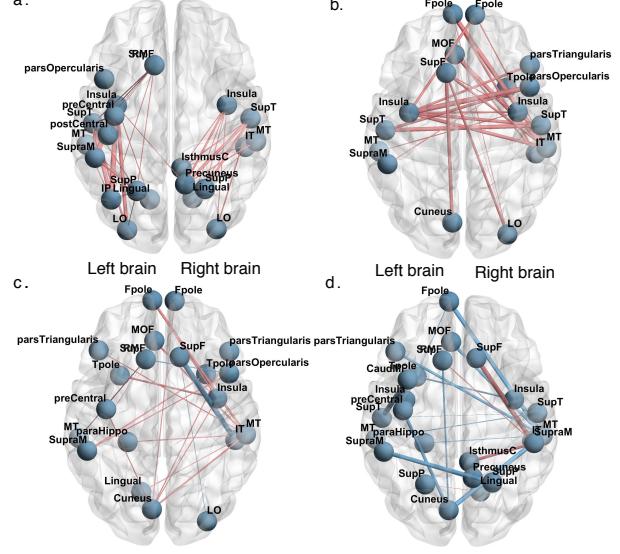


Figure 6. Top edges with large effects. Red edges represent relatively strong connections and blue edges represent relatively weak connections. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+.

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, etc.

We apply our tensor regression model to the *Nations* data. The BIC criterion suggests a rank  $r = (4, 4, 4)$  for the coefficient tensor  $\mathcal{B} \in \mathbb{R}^{6 \times 6 \times 56}$ . Table S1 shows the  $K$ -means clustering of the 56 relations based on the 3<sup>rd</sup> mode factor  $M_3 \in \mathbb{R}^{56 \times 4}$ . We find 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 annotation similarity among grouped entities indicates the clustering results.

## 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.

---

## References

- 440  
441 Abbe, E. Community detection and stochastic block models:  
442 recent developments. *The Journal of Machine Learning*  
443 *Research*, 18(1):6446–6531, 2017.  
444
- 445 Baldin, N. and Berthet, Q. Optimal link prediction with  
446 matrix logistic regression. *arXiv preprint arXiv:1803.07054*,  
447 2018.  
448
- 449 Chen, H., Raskutti, G., and Yuan, M. Non-convex projected  
450 gradient descent for generalized low-rank tensor regres-  
451 sion. *The Journal of Machine Learning Research*, 20(1):  
452 172–208, 2019.  
453
- 454 De Lathauwer, L., De Moor, B., and Vandewalle, J. A  
455 multilinear singular value decomposition. *SIAM Journal*  
456 *on Matrix Analysis and Applications*, 21(4):1253–1278,  
457 2000.  
458
- 459 Gabriel, K. R. Generalised bilinear regression. *Biometrika*,  
460 85(3):689–700, 1998.  
461
- 462 Geddes, L. Human brain mapped in unprecedented detail.  
*Nature*, 2016.  
463
- 464 Hitchcock, F. L. The expression of a tensor or a polyadic as  
465 a sum of products. *Journal of Mathematics and Physics*,  
466 6(1-4):164–189, 1927.  
467
- 468 Hoff, P. D. Bilinear mixed-effects models for dyadic data.  
*Journal of the American Statistical Association*, 100(469):  
469 286–295, 2005.  
470
- 471 Hong, D., Kolda, T. G., and Duersch, J. A. Generalized  
472 canonical polyadic tensor decomposition. *SIAM Review*,  
473 *in press*. *arXiv:1808.07452*, 2019.  
474
- 475 Ingallalikar, M., Smith, A., Parker, D., Satterthwaite, T. D.,  
476 Elliott, M. A., Ruparel, K., Hakonarson, H., Gur, R. E.,  
477 Gur, R. C., and Verma, R. Sex differences in the structural  
478 connectome of the human brain. *Proceedings of the*  
479 *National Academy of Sciences*, 111(2):823–828, 2014.  
480
- 481 Kolda, T. G. and Bader, B. W. Tensor decompositions and  
482 applications. *SIAM review*, 51(3):455–500, 2009.  
483
- 484 Li, L. and Zhang, X. Parsimonious tensor response regres-  
485 sion. *Journal of the American Statistical Association*, 112  
486 (519):1131–1146, 2017.  
487
- 488 McCullagh, P. and Nelder, J. *Generalized Linear Models*,  
489 *Second Edition*. Chapman and Hall/CRC Monographs  
490 on Statistics and Applied Probability Series. Chapman &  
491 Hall, 1989.  
492
- 493 Nickel, M., Tresp, V., and Kriegel, H.-P. A three-way model  
494 for collective learning on multi-relational data. In *Inter-  
495 national Conference on Machine Learning*, volume 11,  
496 pp. 809–816, 2011.  
497
- 498 Oseledets, I. V. Tensor-train decomposition. *SIAM Journal*  
499 *on Scientific Computing*, 33(5):2295–2317, 2011.  
500
- 501 Potthoff, R. F. and Roy, S. A generalized multivariate anal-  
502 ysis of variance model useful especially for growth curve  
503 problems. *Biometrika*, 51(3-4):313–326, 1964.  
504
- 505 Rabusseau, G. and Kadri, H. Low-rank regression with  
506 tensor responses. In *Advances in Neural Information  
507 Processing Systems*, pp. 1867–1875, 2016.  
508
- 509 Robinson, L. F., Atlas, L. Y., and Wager, T. D. Dynamic  
510 functional connectivity using state-based dynamic com-  
511 munity structure: Method and application to opioid anal-  
512 gesia. *NeuroImage*, 108:274–291, 2015.  
513
- 514 Srivastava, M. S., von Rosen, T., and Von Rosen, D. Models  
515 with a kronecker product covariance structure: estimation  
516 and testing. *Mathematical Methods of Statistics*, 17(4):  
517 357–370, 2008.  
518
- 519 Sun, W. W. and Li, L. STORE: sparse tensor response  
520 regression and neuroimaging analysis. *The Journal of  
521 Machine Learning Research*, 18(1):4908–4944, 2017.  
522
- 523 Wang, M. and Zeng, Y. Multiway clustering via tensor block  
524 models. *Advances in Neural Information Processing  
525 Systems 32 (NeurIPS 2019)*. *arXiv:1906.03807*, 2019.  
526
- 527 Wang, M., Duc, K. D., Fischer, J., and Song, Y. S. Operator  
528 norm inequalities between tensor unfoldings on the par-  
529 tition lattice. *Linear Algebra and Its Applications*, 520:  
530 44–66, 2017.  
531
- 532 Xia, M., Wang, J., and He, Y. Brainnet viewer: a network  
533 visualization tool for human brain connectomics. *PloS  
534 one*, 8(7):e68910, 2013.  
535
- 536 Yu, R. and Liu, Y. Learning from multiway data: Simple and  
537 efficient tensor regression. In *International Conference  
538 on Machine Learning*, pp. 373–381, 2016.  
539
- 540 Zhang, A. and Xia, D. Tensor SVD: Statistical and computa-  
541 tional limits. *IEEE Transactions on Information Theory*,  
542 2018.  
543
- 544 Zhang, J., Sun, W. W., and Li, L. Network response regres-  
545 sion for modeling population of networks with covariates.  
546 *arXiv preprint arXiv:1810.03192*, 2018.  
547
- 548 Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Na-  
549 gasaka, Y., Fujii, N., Zhang, L., and Cichocki, A. Higher  
550 order partial least squares (HOPLS): a generalized mul-  
551 tilinear regression method. *IEEE transactions on pat-  
552 tern analysis and machine intelligence*, 35(7):1660–1673,  
553 2012.  
554
- 555 Zhou, H., Li, L., and Zhu, H. Tensor regression with applica-  
556 tions in neuroimaging data analysis. *Journal of the Amer-  
557 ican Statistical Association*, 108(502):540–552, 2013.  
558