

---

# Exponential family tensor regression(V2)

---

**Anonymous Author(s)**

Affiliation  
Address  
email

## Abstract

1 Higher-order tensors have recently received increasing attention in many fields  
2 across science and engineering. Here, we present an exponential family of tensor-  
3 response regression models that incorporate covariates on multiple modes. Such  
4 problems are common in neuroimaging, network modeling, and spatial-temporal  
5 analysis. We propose a rank-constrained estimator and establish the theoretical  
6 accuracy guarantees. Unlike earlier methods, our approach allows covariates  
7 from multiple tensor modes whenever available. An efficient alternating updating  
8 algorithm is further developed. Our proposal handles a broad range of data types,  
9 including continuous, count, and binary observations. We apply the method to  
10 multi-relational social network data and diffusion tensor imaging data from human  
11 connection project. Our approach identifies the key global connectivity pattern and  
12 pinpoints the local regions that are associated with covariates.

13 **1 Introduction**

14 Many contemporary scientific and engineering studies collect multi-way array data, a.k.a. tensors,  
15 accompanied by additional covariates. One example is neuroimaging analysis [1, 2], in which  
16 the brain connectivity networks are collected from a sample of individuals. Researchers are often  
17 interested in identifying connection edges that are affected by individual characteristics such as age,  
18 gender, and disease status (see Figure 1a). Another example is in the field of network analysis [3, 4].  
19 A typical social network consists of nodes that represent people and edges that represent friendships.  
20 In addition, features on nodes and edges are often available, such as people’s personality and  
21 demographic location. It is of keen scientific interest to identify the variation in the connection  
22 patterns (e.g., transitivity, community) that can be attributable to the node features.

23 This paper presents a general treatment to these seemingly different problems. We formulate the  
24 learning task as a regression problem, with tensor observation serving as a response, and the node  
25 features and/or their interactions forming the predictor. Figure 1b illustrates the general set-up we  
26 consider. The regression approach allows the identification of variation in the data tensor that is  
27 explained by the covariates. In contrast to earlier work [5, 6], our method allows the covariates from  
28 multiple modes, whenever available. We utilize a low-rank constraint in the regression coefficient  
29 to encourage the sharing among tensor entries. The statistical convergence of our estimator is  
30 established, and we quantify the gain in predictive power by taking multiple covariates into account.  
31 A secondary contribution is that our method allows a broad range of tensor types, including continuous,  
32 count, and binary observations. While previous tensor regression methods [7, 6] are able to analyze  
33 Gaussian responses, none of them is suitable for exponential distribution family of tensors. We develop  
34 a generalized tensor regression framework, and as a by product, our models allows heteroscedasticity  
35 by relating the variance of tensor entry to its mean. This flexibility is particularly important in practice,  
36 because social network, brain imaging, or gene expression datasets are often non-Gaussian.

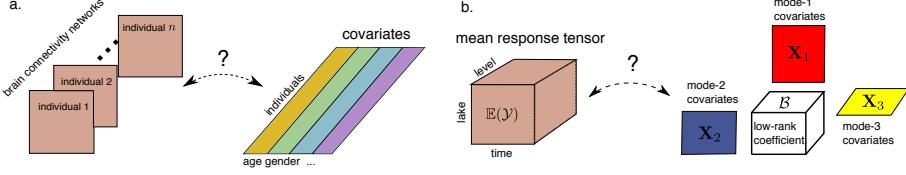


Figure 1: Examples of tensor response regression model with covariates on multiple modes. (a) Network population model. (b) Spatial-temporal growth model.

37 **Related work.** Our work is closely related to but also clearly distinctive from several lines of previous  
 38 work. The first is a class of *unsupervised* tensor decomposition [8, 9, 10] that aims to find a low-rank  
 39 representation of a data tensor. In contrast, our model can be viewed a *supervised* tensor learning,  
 40 which aims to identify the association between a data tensor and covariates. The second related  
 41 line [2, 11] tackles tensor regression where the response is a scalar and the *predictor* is a tensor. Our  
 42 proposal is orthogonal to theirs because we treat the tensor as a *response*. The tensor-response model  
 43 is appealing for high-dimensional analysis when both the response and the covariate dimensions grow.  
 44 The last line of work studies the network-response model [5, 12]. The earlier development of this  
 45 model focuses mostly on binary data in the presence of dyadic covariates [4]. We will demonstrate  
 46 the enhanced accuracy as the order of data grows, and establish the general theory for exponential  
 47 family which is arguably better suited to various data types.

## 48 2 Preliminaries

49 We begin by reviewing the basic properties about tensors [13]. We use  $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!] \in \mathbb{R}^{d_1 \times \dots \times d_K}$   
 50 to denote an order- $K$  ( $d_1, \dots, d_K$ )-dimensional tensor. The multilinear multiplication of a tensor  
 51  $\mathcal{Y} \in \mathbb{R}^{d_1 \times \dots \times d_K}$  by matrices  $\mathbf{X}_k = [\![x_{i_k, j_k}^{(k)}]\!] \in \mathbb{R}^{p_k \times d_k}$  is defined as

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

52 which results in an order- $K$  ( $p_1, \dots, p_K$ )-dimensional tensor. For ease of presentation, we use  
 53 shorthand notion  $\mathcal{Y} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\}$  to denote the tensor-by-matrix product. For any two tensors  
 54  $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!]$ ,  $\mathcal{Y}' = [\![y'_{i_1, \dots, i_K}]\!]$  of identical order and dimensions, their inner product is defined  
 55 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  $\|\mathcal{Y}\|_F =$   
 56  $\langle \mathcal{Y}, \mathcal{Y} \rangle^{1/2}$ . A higher-order tensor can be reshaped into a lower-order object [14]. We use  $\text{vec}(\cdot)$  to  
 57 denote the operation that reshapes the tensor into a vector, and  $\text{Unfold}_k(\cdot)$  the operation that reshapes  
 58 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  
 59  $\mathcal{Y}$  is defined as a length- $K$  vector  $\mathbf{r} = (r_1, \dots, r_K)$ , where  $r_k$  is the rank of matrix  $\text{Unfold}_k(\mathcal{Y})$ ,  
 60  $k = 1, \dots, K$ . We use lower-case letters (e.g.,  $a, b, c$ ) for scalars/vectors, upper-case boldface letters  
 61 (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.  
 62 We let  $I_d$  denote the  $d \times d$  identity matrix,  $[d]$  denote the  $d$ -set  $\{1, \dots, d\}$ , and allow an  $\mathbb{R} \rightarrow \mathbb{R}$   
 63 function to be applied to tensors in an element-wise manner.

## 64 3 Motivation and model

65 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  
 66 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  
 67  $p_k \leq d_k$ . We propose a multilinear structure on the conditional expectation of the tensor. Specifically,

$$68 \quad \mathbb{E}(\mathcal{Y} | \mathbf{X}_1, \dots, \mathbf{X}_K) = f(\Theta), \quad (1)$$

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

69 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  
 70 parameter tensor of interest, and  $\times$  denotes the tensor Tucker product. The choice of link function  
 71 depends on the distribution of the response data. Some common choices are identity link for Gaussian  
 72 tensor, logistic link for binary tensor, and  $\exp(\cdot)$  link for Poisson tensor (see Table 1).

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

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.

74 **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

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

80 where  $\mathcal{B} \in \mathbb{R}^{p \times (r+1) \times (s+1)}$  is the coefficient tensor of interest,  $\mathbf{X}_1 = \text{blockdiag}\{\mathbf{1}_q, \dots, \mathbf{1}_q\} \in$   
81  $\{0, 1\}^{d \times p}$  is the design matrix for lake types,

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

82 are the design matrices for spatial and temporal effects, respectively. The model (2) is a higher-order  
83 extension of the “growth curve” model originally proposed for matrix data [15, 16, 17]. Clearly, the  
84 spatial-temporal model is a special case of our tensor regression model, with covariates available on  
85 each of the three modes.

86 **Example 2** (Network population model). Network response model is recently developed in the  
87 context of neuroimaging analysis. The goal is to study the relationship between network-valued  
88 response and the individual covariates. Suppose we observe  $n$  i.i.d. observations  $\{(\mathbf{Y}_i, \mathbf{x}_i) : i =$   
89  $1, \dots, n\}$ , where  $\mathbf{Y}_i \in \{0, 1\}^{d \times d}$  is the brain connectivity network on the  $i$ -th individual, and  $\mathbf{x}_i \in \mathbb{R}^p$   
90 is the individual covariate such as age, gender, cognition, etc. The network-response model [5, 18]  
91 has the form

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

92 where  $\mathcal{B} \in \mathbb{R}^{d \times d \times p}$  is the coefficient tensor of interest.

93 The model (3) is a special case of our tensor-response model, with covariates on the last mode of  
94 the tensor. Specifically, stacking  $\{\mathbf{Y}_i\}$  together yields an order-3 response tensor  $\mathcal{Y} \in \{0, 1\}^{d \times d \times n}$ ,  
95 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

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

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

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

102 The above model has demonstrated its success in modeling transitivity, balance, and communities in  
103 the networks [4]. We show that our tensor regression model (1) also incorporates the graph model as a  
104 special case. Let  $\mathcal{Y} = [[y_{ij}]]$  be a binary matrix where  $y_{ij} = \mathbb{1}_{(i,j) \in E}$ . Define  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in$   
105  $\mathbb{R}^{n \times p}$ . Then, the graph model (4) can be expressed as

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

106 In the above three examples and many other studies, researchers are interested in uncovering the  
107 variation in the data tensor that can be explained by the covariates. The regression coefficient  $\mathcal{B}$   
108 in our model model (1) serves this goal by collecting the effects of covariates and the interaction

109 thereof. To encourage the sharing among effects, we assume that the coefficient tensor  $\mathcal{B}$  lies in a  
110 low-dimensional parameter space:

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

111 where  $r_k(\mathcal{B}) \leq p_k$  is the Tucker rank at mode  $k$  of the tensor. The low-rank assumption is plausible  
112 in many scientific applications. In brain imaging analysis, for instance, it is often believed that the  
113 brain nodes can be grouped into fewer communities, and the numbers of communities are much  
114 smaller than the number of nodes. The low-rank structure encourages the shared information across  
115 tensor entries, thereby greatly improving the estimation stability. When no confusion arises, we drop  
116 the subscript  $(r_1, \dots, r_K)$  and write  $\mathcal{P}$  for simplicity.

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

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

121 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  
122 coefficient tensor of interest. We comment that other forms of tensor low-rankness are also possible,  
123 and here we choose Tucker rank just for parsimony. Similar models can be derived using various  
124 notions of low-rankness based on CP decomposition [19] and train decomposition [20].

## 125 4 Rank-constrained likelihood-based estimation

126 We develop a likelihood-based procedure to estimate the coefficient tensor  $\mathcal{B}$  in (5). We adopt the  
127 exponential family as a flexible framework for different data types. In a classical generalized linear  
128 model (GLM) with a scalar response  $y$  and covariate  $\mathbf{x}$ , the density is expressed as:

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

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

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

$$\begin{aligned} \mathcal{L}_{\mathcal{Y}}(\mathcal{B}) &= \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1, \dots, i_K} b(\theta_{i_1, \dots, i_K}), \\ \text{where } \Theta &= \mathcal{B} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\}. \end{aligned}$$

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

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

142 In the following theoretical analysis, we assume the rank  $\mathbf{r} = (r_1, \dots, r_K)$  is known and fixed. The  
143 adaptation of unknown  $\mathbf{r}$  will be addressed in Section 5.3.

### 144 4.1 Statistical properties

145 We assess the estimation accuracy using the deviation in the Frobenius norm. For the true coefficient  
146 tensor  $\mathcal{B}_{\text{true}}$  and its estimator  $\hat{\mathcal{B}}$ , define

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

147 In modern applications, the response tensor and covariates are often large-scale. We are particularly  
 148 interested in the high-dimensional region in which both  $d_k$  and  $p_k$  diverge; i.e.  $d_k \rightarrow \infty$  and  
 149  $p_k \rightarrow \infty$ , while  $\frac{p_k}{d_k} \rightarrow \gamma_k \in [0, 1]$ . As the size of problem grows, and so does the number of  
 150 unknown parameters. As such, the classical MLE theory does not directly apply. We leverage the  
 151 recent development in random tensor theory and high-dimensional statistics to establish the error  
 152 bounds of the estimation.

153 **Assumption 1.** *We make the following assumptions:*

154 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  
 155  $k \in [K]$ . Here  $\sigma_{\min}(\cdot)$  and  $\sigma_{\max}(\cdot)$  denotes the smallest and largest singular values, respectively.*

156 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  
 157  $|\theta_{i_1, \dots, i_K}| \leq \alpha$ .*

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

160 The assumptions are fairly mild. Assumption A1 guarantees the non-singularity of the covariates,  
 161 and Assumption A2 ensures the log-likelihood  $\mathcal{Y}(\Theta)$  is strictly concave in the linear predictor  $\Theta$ .  
 162 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  $y_{i_1, \dots, i_K}$   
 163 belongs to an exponential family [21].

164 **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  $1 - \exp(-C_1 \sum_k p_k)$ ,*

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

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

170 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.

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

181 **Example 4** (Spatio-temporal growth model). The estimated type-by-time-by-space coefficient tensor  
 182 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  
 183 as long as the dimension grows in either of the three modes.

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

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

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

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

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

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

## 196 5 Numerical implementation

### 197 5.1 Alternating optimization

198 In this section, we introduce an efficient algorithm to solve (6). The objective function  $\mathcal{L}_Y(\mathcal{B})$  is  
 199 concave in  $\mathcal{B}$  when the link  $f$  is the canonical link function. However, the feasible set  $\mathcal{P}$  is non-convex,  
 200 and thus the optimization (6) is a non-convex problem. We utilize a Tucker factor representation of  
 201 the coefficient tensor  $\mathcal{B}$  and turn the optimization into a block-wise convex problem.

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

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

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

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

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

206 The decision variables in the above objective function consist of  $K+1$  blocks of variables, one for the  
 207 core tensor  $\mathcal{C}$  and  $K$  for the factor matrices  $\mathbf{M}_k$ 's. We notice that, if any  $K$  out of the  $K+1$  blocks  
 208 of variables are known, then the optimization with respect to the last block of variables reduced to a  
 209 simple GLM. We therefore choose to iteratively update one block at a time while keeping others fixed.  
 210 We leverage on a block relaxation algorithm for optimization, and the classical (local) convergence for  
 211 block algorithm applies. Although a non-convex optimization of this type usually has no guarantee on  
 212 global optimality, our numerical experiments have suggested high-quality solutions (see Section 6).  
 The full algorithm is described in Algorithm 1.

---

#### Algorithm 1 Generalized tensor response regression with covariates on multiple modes

**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  
 rank  $\mathbf{r} = (r_1, \dots, r_K)$ , link function  $f$ , infinity norm bound  $\alpha$

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

- 1: Calculate  $\tilde{\mathcal{B}} = \mathcal{Y} \times_1 [(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T] \times_2 \dots \times_K [(\mathbf{X}_K^T \mathbf{X}_K)^{-1} \mathbf{X}_K^T]$ .
  - 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  
 rank- $\mathbf{r}$  Tucker approximation of  $\tilde{\mathcal{B}}$ , in the least-square sense.
  - 3: **while** the relative increase in objective function  $\mathcal{L}_Y(\mathcal{B})$  is less than the tolerance **do**
  - 4:     Update iteration index  $t \leftarrow t + 1$ .
  - 5:     **for**  $k = 1$  to  $K$  **do**
  - 6:         Obtain the factor matrix  $\mathbf{M}_k^{(t+1)} \in \mathbb{R}^{p_k \times r_k}$  by solving  $p_k$  separate GLMs with link function  $f$ .
  - 7:         Update the columns of  $\mathbf{M}_k^{(t+1)}$  by Gram-Schmidt orthogonalization.
  - 8:     **end for**
  - 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,  $\odot_{k=1}^K [\mathbf{X}_k \mathbf{M}_k^{(t)}]$   
 as covariates, and  $f$  as link function. Here  $\odot$  denotes the Khatri-Rao product of matrices.
  - 10:    Rescale the core tensor subject to the infinity norm constraint.
  - 11:    Update  $\mathcal{B}^{(t+1)} \leftarrow \mathcal{C}^{(t+1)} \times_1 \mathbf{M}_1^{(t+1)} \times_2 \dots \times_K \mathbf{M}_K^{(t+1)}$ .
  - 12: **end while**
- 

213

### 214 5.2 Algorithmic Properties(Need Revision)

215 The convergence of objective function is guaranteed whenever  $\mathcal{L}_Y$  is upper bounded, because  
 216 Algorithm 1 monotonically increases the objective function. Now, we study the convergence  
 217 property of actual iterates  $\mathcal{A}^{(t)} = (\mathcal{C}^{(t)}, \{\mathbf{M}_k^{(t)}\})$  generated by Algorithm 1. To simplify the analysis,  
 218 we set the hyper-parameter  $\alpha$  to infinity.

219 Note that tucker decomposition for higher-order tensor is not unique. Consider the example.

$$\mathcal{B} = \mathcal{C} \times \{\mathbf{M}_1, \dots, \mathbf{M}_K\} = \mathcal{C} \times \{\mathbf{P}_1, \dots, \mathbf{P}_K\} \times \{\mathbf{M}_1 \mathbf{P}_1^{-1}, \dots, \mathbf{M}_K \mathbf{P}_K^{-1}\},$$

220 where  $\mathbf{P}_k \in \mathbb{R}^{r_k \times r_k}, \forall k \in [K]$ . Let  $\mathcal{C}' = (\mathcal{C} \times \{\mathbf{P}_1, \dots, \mathbf{P}_K\})$  and  $\{\mathbf{M}'_k\} = \{\mathbf{M}_k \mathbf{P}_k^T\}$ . Different  
221 decision variables  $(\mathcal{C}, \{\mathbf{M}_k\})$  and  $(\mathcal{C}', \{\mathbf{M}'_k\})$  lead to the same tensor representation  $\mathcal{B}$ . If  $\{\mathbf{P}_k\}$ s  
222 are orthogonal matrices, we describe that  $(\mathcal{C}, \{\mathbf{M}_k\})$  and  $(\mathcal{C}', \{\mathbf{M}'_k\})$  are equivalent.

223 **Proposition 1.** *Tucker decomposition for higher-order tensor is unique up to equivalence, if factor  
224 matrices are column-wise orthogonal.*

225 For column-wise orthogonal factor matrices, the tensor representation can no be identical if two sets  
226 of decision variables are not equivalent.

227 **Assumption 2.** We need following assumptions for algorithmic convergence results.

228 *B1. (Regularity Condition) The log-likelihood function  $\mathcal{L}(\mathcal{A})$  is continuous. The set  $\{\mathcal{A} : \mathcal{L}(\mathcal{A}) \geq$   
229  $\mathcal{L}(\mathcal{A}^{(0)})\}$  is compact.*

230 *B2. (Strictly local maximum condition) Each block update in Algorithm 1 is well-defined; i.e. the GLM  
231 solution for each update exists and is unique, the corresponding diagonal sub-block Hessian of  
232  $\mathcal{L}_{\mathcal{Y}}(\mathcal{A})$  is negative definite at the solution.*

233 *B3. (Local Uniqueness condition) The set of stationary points of  $\mathcal{L}_{\mathcal{Y}}(\mathcal{A})$  are isolated up to equivalence.*

234 Assumptions are mild for higher-order tensor. A1 ensures the maximum of likelihood function exists  
235 and log-likelihood is upper bounded due to compactness. Therefore, the stopping rule for Algorithm  
236 1 is well-defined. A2 ensures the negative-definiteness of sub-block Hessian only for  $\mathcal{C}$  or  $\mathbf{M}_k$ . Note  
237 that the full Hessian needs not to be negative definite for all decision variables simultaneously. We  
238 only require the solution for one variable( $\mathcal{C}$  or  $\mathbf{M}_k$ ) given by GLM is unique when other decision  
239 variables are fixed. A3 guarantees the uniqueness of stationary points up to equivalence.

240 **Proposition 2** (Algorithmic Convergence). Suppose assumption (B1)-(B3) hold.

241 *i. (Global Convergence) Any sequence  $\mathcal{A}^{(t)}$  generated by Algorithm 1 converges to equivalent  
242 stationary points of  $\mathcal{L}_{\mathcal{Y}}(\mathcal{A})$ .*

243 *ii. (Local Linear Convergence) Let  $\mathcal{A}^*$  be a local maximizer of  $\mathcal{L}(\mathcal{A})$ . There exists an  $\epsilon$ -neighborhood  
244 of  $\mathcal{A}^*$ , such that, for any  $\mathcal{A}^{(0)}$  in the neighborhood, the iterates  $\mathcal{A}^{(t)}$  generated by Algorithm 1  
245 linearly convergent to  $\mathcal{A}^*$ .*

$$\left\| \mathcal{A}^{(t)} - \mathcal{A}^* \right\|_F \leq \rho^t \left\| \mathcal{A}^{(0)} - \mathcal{A}^* \right\|_F$$

246 where  $\rho \in (0, 1)$  is a contraction parameter.

247 Though the stationary points is not necessarily a global maximum or even local maximum(may  
248 be saddle point), our alternating algorithm always converges to at least local maximum with  
249 good initialization in practice. The local linear convergences shows that iterates generated by  
250 algorithm are attracted by local maximums. The error of iterates exponentially decays near the  
251 local maximizers.

### 252 5.3 Rank selection

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

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

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

## 260 6 Simulation

261 We evaluate the empirical performance of our generalized tensor regression through simulations. We  
262 consider order-3 tensors with a range of distribution types. The coefficient tensor  $\mathcal{B}$  is generated using

263 the factorization form (8) where both the core and factor matrices are drawn i.i.d. from Uniform[-1,1].  
 264 The linear predictor is then simulated from  $\mathcal{U} = \mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ , where  $\mathbf{X}_k$  is either an identity  
 265 matrix (i.e. no covariate available) or Gaussian random matrix with i.i.d. entries from  $N(0, \sigma_k^2)$ . We  
 266 set  $\sigma_k = d_k^{-1/2}$  to ensure the singular values of  $\mathbf{X}_k$  are bounded as  $d_k$  increases. The  $\mathcal{U}$  is scaled such  
 267 that  $\|\mathcal{U}\|_\infty = 1$ . Conditional on the linear predictor  $\mathcal{U} = [u_{ijk}]$ , the entries in the tensor  $\mathcal{Y} = [y_{ijk}]$   
 268 are drawn independently according to one of the following three probabilistic models:

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

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

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

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

## 274 6.1 Finite-sample performance

275 The experiment I assesses the selection accuracy of our BIC criterion (9). We consider the balanced  
 276 situation where  $d_k = d$ ,  $p_k = 0.4d_k$  for  $k = 1, 2, 3$ . We set  $\alpha = 10$  and consider various  
 277 combinations of dimension  $d$  and rank  $\mathbf{r} = (r_1, r_2, r_3)$ . For each combination, we simulate tensor  
 278 data following Gaussian, Bernoulli, and Poisson models. We then minimize BIC using a grid search  
 279 over three dimensions. The hyper-parameter  $\alpha$  is set to infinity in the fitting, which essentially  
 280 imposes no prior on the coefficient magnitude. Table 2 reports the selected rank averaged over  
 281  $n_{\text{sim}} = 30$  replicates for Gaussian and Poisson models. We found that when  $d = 20$ , the selected rank  
 282 is slightly smaller than the true rank, and the accuracy improves immediately when the dimension  
 283 increases to  $d = 40$ . This agrees with our expectation, as in tensor regression, the sample size is  
 284 related to the number of entries. A larger  $d$  implies a larger sample size, so the BIC selection becomes  
 285 more accurate.

True Rank $\mathbf{r}$	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.

286 The experiment II evaluates the accuracy when covariates are available on all modes. We set  
 287  $\alpha = 10$ ,  $d_k = d$ ,  $p_k = 0.4d_k$ ,  $r_k = r \in \{2, 4, 6\}$  and increase  $d$  from 25 to 50. Our theoretical  
 288 analysis suggests that  $\hat{\mathcal{B}}$  has a convergence rate  $\mathcal{O}(d^{-2})$  in this setting. Figure 1 plots the estimation  
 289 error versus the “effective sample size”,  $d^2$ , under three different distribution models. We found that  
 290 the empirical MSE decreases roughly at the rate of  $1/d^2$ , which is consistent with our theoretical  
 291 ascertainment. We also observed that, tensors with higher ranks tend to yield higher estimation errors,  
 292 as reflected by the upward shift of the curves as  $r$  increases. Indeed, a larger  $r$  implies a higher model  
 293 complexity and thus greater difficulty in the estimation. Similar behaviors can be observed in the  
 294 non-Gaussian data in Figures 2b-c.

295 The experiment III investigates the capability of our model in handling correlation among coefficients.  
 296 We mimic the scenario of brain imaging analysis. A sample of  $d_3 = 50$  networks are simulated, one  
 297 for each individual. Each network measures the connections between  $d_1 = d_2 = 20$  brain nodes. We  
 298 simulate  $p = 5$  covariates for each of the 50 individuals. These covariates may represent, for  
 299 example, age, gender, cognitive score, etc. Recent study [22] has suggested that brain connectivity  
 300 networks often exhibit community structure represented as a collection of subnetworks, and each  
 301 subnetwork is comprised of a set of spatially distributed brain nodes. To accommodate this structure,  
 302 we utilize the stochastic block model [23] to generate the effect size. Specifically, we partition the  
 303 nodes into  $r$  blocks by assigning each node to a block with uniform probability. Edges within a same  
 304 block are assumed to share the same covariate effects, where the effects are drawn i.i.d. from  $N(0, 1)$ .  
 305 We then apply our tensor regression model to the network data using the BIC-selected rank. Note  
 306 that in this case, the true model rank is unknown; the rank of a  $r$ -block matrix is not necessarily equal  
 307 to  $r$  [24].

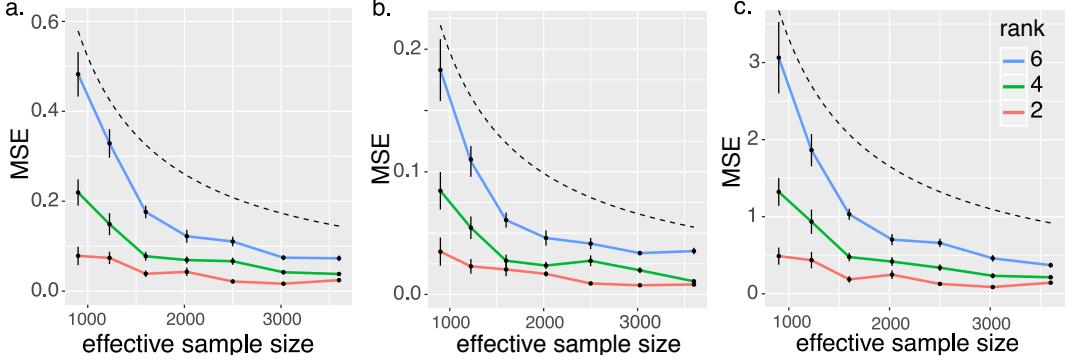


Figure 2: Mean squared error (MSE) against effective sample size. The three panels depict the MSE when the response tensors are generated from (a) Gaussian (b) Poisson and (c) Bernoulli models. The dashed curves correspond to  $\mathcal{O}(1/d^2)$ .

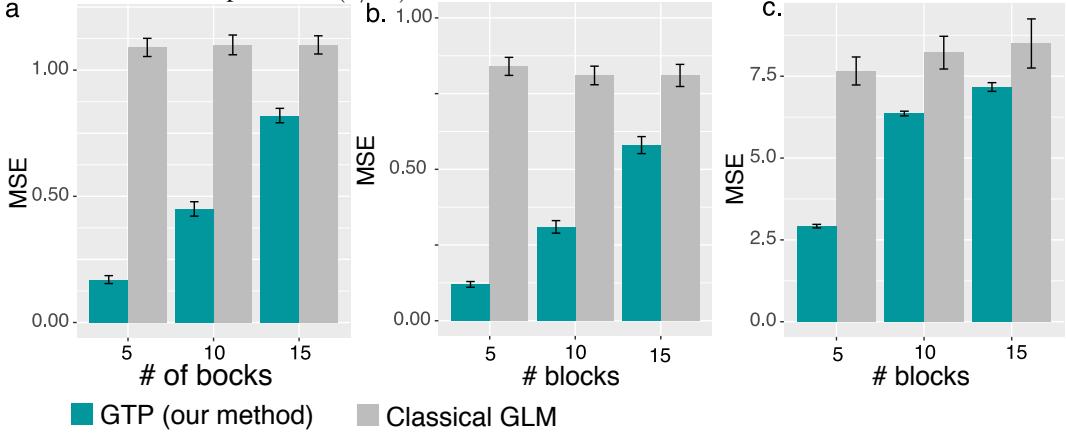


Figure 3: MSE when the networks have block structure. The three panels depict the MSE when the response tensors are generated from (a) Gaussian (b) Poisson and (c) Bernoulli models. The  $x$ -axis represents the number of blocks in the networks.

308 Figure 3 compares the MSE of our method with a classical GLM approach. A classical GLM is  
 309 to regress the dyadic edges, one at a time, on the covariates, and this model is repeatedly fitted for  
 310 each edge. This repeated approach, however, does not account for the correlation among the edges,  
 311 and may suffer from overfitting. As we can see in Figure 3, our tensor regression method achieves  
 312 significant error reduction in all three models considered. The outer-performance is significant in  
 313 the presence of large communities, and even in the less structured case ( $\sim 20/15 = 1.33$  nodes per  
 314 block), our method still outer-performs GLM. This is because the low-rankness in our modeling  
 315 automatically identifies the shared information across entries. By selecting the rank in a data-driven  
 316 way, our method is able to achieve accurate estimation with improved interpretability.

## 317 6.2 Comparison with alternative methods

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

- 319 • Higher-order low-rank regression (**HOLRR**, (author?) [5]) is a least-square based tensor regres-  
320 sion that allows covariates on a single mode.
- 321 • Higher-order partial least square (**HOPLS**, (author?) [7]) is a dimension-reduction method that  
322 jointly models a tensor response and a tensor covariate.
- 323 • Subsampled tensor projected gradient (**TPG**, (author?) [6]) tackles the same question as **HOLRR**  
324 but instead uses a different algorithm to solve the problem.

325 These three methods are the closest algorithms to ours, in that they relate a tensor response to  
 326 covariates using a low-rank structure. All the three methods allow only Gaussian data, whereas ours  
 327 is applicable to any exponential family distribution including Gaussian, Bernoulli, Multinomial, etc.  
 328 For fair comparison, we consider only Gaussian response in the simulation. We measure the accuracy

329 using mean squared prediction error,  $\text{MSPE} = \sqrt{\sum_k d_k} \|\hat{\mathcal{Y}} - \mathbb{E}(\mathcal{Y}|\mathcal{X})\|_F$ , where  $\hat{\mathcal{Y}}$  is the fitted value  
 330 from each of the methods.

331 The comparison was assessed from three aspects: (a) benefit of incorporating covariates from  
 332 multiple modes; (b) prediction error with respect to sample size; (c) sensitivity of accuracy with  
 333 respect to model complexity. We use similar simulation setups as in our experiment II, but consider  
 334 combinations of rank ( $r = (3, 3, 3)$  vs.  $(4, 5, 6)$ ), noise ( $\sigma = 1/2$  vs.  $1/4$ ), and dimension ( $d$  ranging  
 335 from 20 to 100 for modes with covariates,  $d = 20$  for modes without covariates).

336 Figure 4 shows the averaged prediction error across 30 replicates. We see that our **GTR** outperforms  
 337 others, especially in the high-rank high-noise setting. As the number of informative modes (i.e.  
 338 modes with available covariates) increases, the **GTR** exhibits a reduction in error whereas others  
 339 have increased errors. This showcases the benefit toward prediction via incorporation of multiple  
 340 covariates. Note that our method **GTR** is most comparable to **HOLRR** when there is only a single  
 341 informative mode. In such a case, both methods share a same cost function but have different  
 342 algorithms. **GTR** alternates between informative and non-informative modes, whereas **HOLRR**  
 343 approximates the non-informative modes via unfolded response alone. The accuracy gain in Figure 4  
 344 demonstrates the benefit of alternating algorithm – having informative modes also improves the  
 345 estimation along non-informative modes.

346 Figure 5 compares the prediction error with respect to sample size. The sample size is the total  
 347 number of entries in the tensor. In the low-rank setting, our method has similar performance as  
 348 **HOLRR**, and the improvement becomes more pronounced when the rank increases. Neither **HOPLS**  
 349 nor **TPG** has satisfactory performance in high-rank or high-noise settings. One possible reason is  
 350 that a higher rank implies a higher inter-mode complexity, and our **GTR** method lends itself well to  
 351 this context.

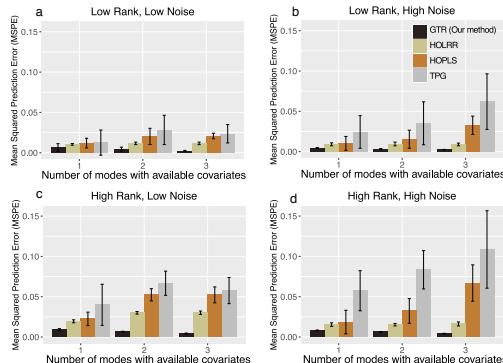


Figure 4: Comparison of MSPE versus the number of modes with covariates. We consider rank  $r = (3, 3, 3)$  (low),  $r = (4, 5, 6)$  (high), and noise  $\sigma = 1/2$  (high),  $\sigma = 1/4$  (low).

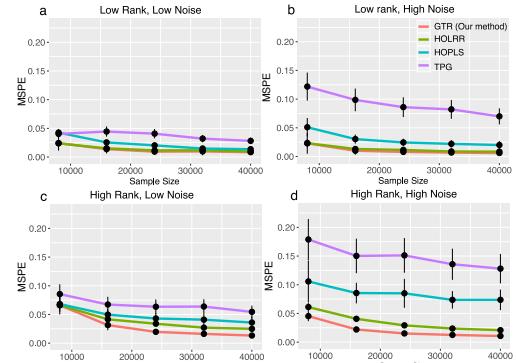


Figure 5: Comparison of MSPE versus sample size. We consider rank  $r = (3, 3, 3)$  (low),  
 $r = (4, 5, 6)$  (high), and noise  $\sigma = 1/2$  (high),  
 $\sigma = 1/4$  (low).

## 352 7 Data analysis

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

### 356 7.1 Human Connectome Project (HCP)

357 The Human connectome project (HCP, [25]) aims to build a network map that characterizes the  
 358 anatomical and functional connectivity within healthy human brains. We take a subset of HCP  
 359 data that consists of 136 brain structural networks, one for each individual. Each brain network is  
 360 represented as a 68-by-68 binary matrix, where the entries encode the presence or absence of fiber  
 361 connections between 68 brain regions. We consider four individual-covariates: gender, age 22-25,  
 362 age 26-30, and age 31+.

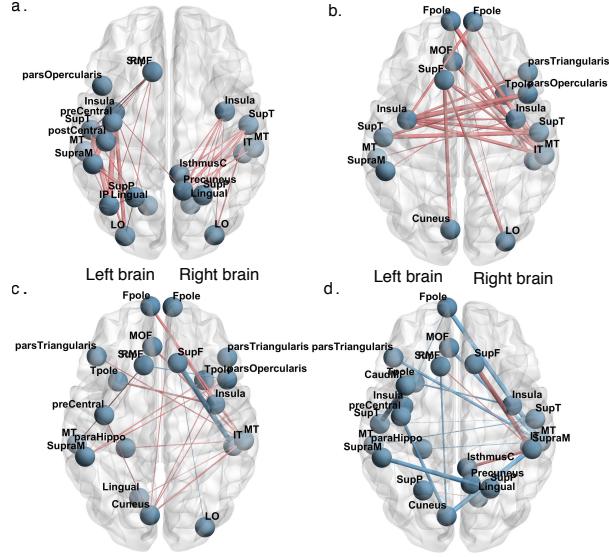


Figure 6: Top edges with large effects. Red edges represent relatively strong connections and blue edges represent relatively weak connections. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+.

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 [26]. 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 [27]. 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 [28] which records 56 relations among 14 countries between 1950 and 1965. The multi-relational networks can be organized into a  $14 \times 14 \times 56$  binary tensor, with each entry indicating the presence or absence of a connection, such as “sending tourist to”, “export”, “import”, between countries. The 56 relations span the fields of politics, economics, military, religion, 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 ?? 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

393 accuracy may also be desirable, such as the spectral norm or the maximum norm of the deviation.  
394 Exploiting the properties and benefits of different error quantification warrants future research.

395 **References**

- 396 [1] Will Wei Sun and Lexin Li. STORE: sparse tensor response regression and neuroimaging  
397 analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.
- 398 [2] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging  
399 data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.
- 400 [3] Nicolai Baldin and Quentin Berthet. Optimal link prediction with matrix logistic regression.  
401 *arXiv preprint arXiv:1803.07054*, 2018.
- 402 [4] Peter D Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical  
403 Association*, 100(469):286–295, 2005.
- 404 [5] Guillaume Rabusseau and Hachem Kadri. Low-rank regression with tensor responses. In  
405 *Advances in Neural Information Processing Systems*, pages 1867–1875, 2016.
- 406 [6] Rose Yu and Yan Liu. Learning from multiway data: Simple and efficient tensor regression. In  
407 *International Conference on Machine Learning*, pages 373–381, 2016.
- 408 [7] Qibin Zhao, Cesar F Caiafa, Danilo P Mandic, Zenas C Chao, Yasuo Nagasaka, Naotaka Fujii,  
409 Liqing Zhang, and Andrzej Cichocki. Higher order partial least squares (HOPLS): a generalized  
410 multilinear regression method. *IEEE transactions on pattern analysis and machine intelligence*,  
411 35(7):1660–1673, 2012.
- 412 [8] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value  
413 decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.
- 414 [9] David Hong, Tamara G Kolda, and Jed A Duersch. Generalized canonical polyadic tensor  
415 decomposition. *SIAM Review, in press. arXiv:1808.07452*, 2019.
- 416 [10] Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions  
417 on Information Theory*, 2018.
- 418 [11] Han Chen, Garvesh Raskutti, and Ming Yuan. Non-convex projected gradient descent for  
419 generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–  
420 208, 2019.
- 421 [12] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American  
422 Statistical Association*, 112(519):1131–1146, 2017.
- 423 [13] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*,  
424 51(3):455–500, 2009.
- 425 [14] Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities  
426 between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520:44–  
427 66, 2017.
- 428 [15] K Ruben Gabriel. Generalised bilinear regression. *Biometrika*, 85(3):689–700, 1998.
- 429 [16] Richard F Potthoff and SN Roy. A generalized multivariate analysis of variance model useful  
430 especially for growth curve problems. *Biometrika*, 51(3-4):313–326, 1964.
- 431 [17] Muni S Srivastava, Tatjana von Rosen, and Dietrich Von Rosen. Models with a kronecker  
432 product covariance structure: estimation and testing. *Mathematical Methods of Statistics*,  
433 17(4):357–370, 2008.
- 434 [18] Jingfei Zhang, Will Wei Sun, and Lexin Li. Network response regression for modeling popula-  
435 tion of networks with covariates. *arXiv preprint arXiv:1810.03192*, 2018.
- 436 [19] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of  
437 Mathematics and Physics*, 6(1-4):164–189, 1927.
- 438 [20] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*,  
439 33(5):2295–2317, 2011.

- 440 [21] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and  
441 Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.
- 442 [22] Lucy F Robinson, Lauren Y Atlas, and Tor D Wager. Dynamic functional connectivity using  
443 state-based dynamic community structure: Method and application to opioid analgesia.  
444 *NeuroImage*, 108:274–291, 2015.
- 445 [23] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The  
446 Journal of Machine Learning Research*, 18(1):6446–6531, 2017.
- 447 [24] Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. *Advances in  
448 Neural Information Processing Systems 32 (NeurIPS 2019)*. arXiv:1906.03807, 2019.
- 449 [25] Linda Geddes. Human brain mapped in unprecedented detail. *Nature*, 2016.
- 450 [26] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for  
451 human brain connectomics. *PloS one*, 8(7):e68910, 2013.
- 452 [27] Madhura Ingallikar, Alex Smith, Drew Parker, Theodore D Satterthwaite, Mark A Elliott,  
453 Kosha Ruparel, Hakon Hakonarson, Raquel E Gur, Ruben C Gur, and Ragini Verma. Sex  
454 differences in the structural connectome of the human brain. *Proceedings of the National  
455 Academy of Sciences*, 111(2):823–828, 2014.
- 456 [28] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective  
457 learning on multi-relational data. In *International Conference on Machine Learning*, volume 11,  
458 pages 809–816, 2011.