
Exponential Family Tensor Regression with Covariates on Multiple Modes

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.

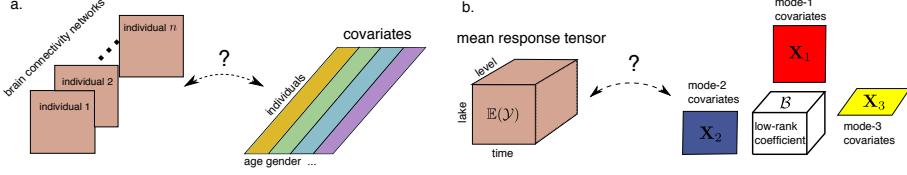


Figure 1: Examples of tensor response regression model with covariates on multiple modes. (a) Network population model. (b) Spatial-temporal growth model.

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 $\mathcal{Y} = \llbracket y_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$
 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 $\mathbf{X}_k = \llbracket x_{i_k, j_k}^{(k)} \rrbracket \in \mathbb{R}^{p_k \times d_k}$ is defined as

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

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 $\mathcal{Y} = \llbracket y_{i_1, \dots, i_K} \rrbracket$, $\mathcal{Y}' = \llbracket y'_{i_1, \dots, i_K} \rrbracket$ 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 $\|\mathcal{Y}\|_F = \langle \mathcal{Y}, \mathcal{Y} \rangle^{1/2}$.
 56 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 \mathbf{I}_d denote the $d \times d$ identity matrix, $[d]$ denote the d -set $\{1, \dots, d\}$, and allow an $\mathbb{R} \rightarrow \mathbb{R}$
 63 function to be applied to tensors in an element-wise manner.

64 3 Motivation and model

65 Let $\mathcal{Y} = \llbracket y_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$ 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,

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

68 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
 69 parameter tensor of interest, and \times denotes the tensor Tucker product. The choice of link function
 70 depends on the distribution of the response data. Some common choices are identity link for Gaussian
 71 tensor, logistic link for binary tensor, and $\exp(\cdot)$ link for Poisson tensor (see Table 1).

72 We give three 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 $f(\theta)$	θ	$\exp(\theta)$	$(1 + \exp(-\theta))^{-1}$

Table 1: Canonical links for common distributions.

73 **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

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

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

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

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

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

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

91 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
92 tensor-response model, with covariates on the last mode of the tensor. Specifically, stacking $\{\mathbf{Y}_i\}$
93 together yields an order-3 response tensor $\mathcal{Y} \in \{0, 1\}^{d \times d \times n}$, along with covariate matrix $\mathbf{X} =$
94 $[\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$. Then, the model (3) can be written as

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

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

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

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

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

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

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

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

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

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

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

120 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
121 coefficient tensor of interest. We comment that other forms of tensor low-rankness are also possible,
122 and here we choose Tucker rank just for parsimony. Similar models can be derived using various
123 notions of low-rankness based on CP decomposition [19] and train decomposition [20].

124 4 Rank-constrained likelihood-based estimation

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

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

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

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

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

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

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

140 In the following theoretical analysis, we assume the rank $\mathbf{r} = (r_1, \dots, r_K)$ is known and fixed. The
141 adaptation of unknown \mathbf{r} will be addressed in Section 5.2.

142 4.1 Statistical properties

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

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

145 In modern applications, the response tensor and covariates are often large-scale. We are particularly
146 interested in the high-dimensional region in which both d_k and p_k diverge; i.e. $d_k \rightarrow \infty$ and

147 $p_k \rightarrow \infty$, while $\frac{p_k}{d_k} \rightarrow \gamma_k \in [0, 1]$. As the size of problem grows, and so does the number of
 148 unknown parameters. As such, the classical MLE theory does not directly apply. We leverage the
 149 recent development in random tensor theory and high-dimensional statistics to establish the error
 150 bounds of the estimation.

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

152 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
 153 $k \in [K]$. Here $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denotes the smallest and largest singular values, respectively.*

154 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
 155 $|\theta_{i_1, \dots, i_K}| \leq \alpha$.*

156 A2'. *Equivalently, there exists two positive constants $L, U > 0$ such that $L \leq b''(\theta) \leq U$ for all
 157 $|\theta| \leq \alpha$, where α is the upper bound of the linear predictor.*

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

162 **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
 163 of an exponential family distribution, and $\mathbb{E}(\mathcal{Y} | \mathcal{X})$ follows the low-rank tensor regression model (5).
 164 Under Assumption 1, there exist two constants $C_1, C_2 > 0$, such that, with probability at least
 165 $1 - \exp(-C_1 \sum_k p_k)$,*

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

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

167 To gain further insight on the bound (7), we consider a special case when tensor dimensions are equal
 168 at every mode, 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
 169 design matrices with i.i.d. $N(0, 1)$ entries. To put the context in the framework of Theorem 4.1, we
 170 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 $1 \pm \sqrt{\gamma}$.
 171 The result in (7) implies that the estimated coefficient has a convergence rate $\mathcal{O}(\frac{p}{d^K})$ in the scale of
 172 the original covariates $\{\mathbf{X}_k\}$. Therefore, our estimation is consistent as the dimension grows, and the
 173 convergence becomes especially favorably as the order of tensor data increases.
 174

175 As immediate applications, we obtain the convergence rate for the three examples mentioned in
 176 Section 3. Without loss of generality, we assume that the singular values of the d_k -by- p_k covariate
 177 matrix \mathbf{X}_k are bounded by $\sqrt{d_k}$. In Network population model, the estimated node-by-node-by-
 178 covariate tensor converges at the rate $\mathcal{O}((2d + p)/d^2 n)$ where $p \leq n$. In Dyadic data with node
 179 attributes model, the estimated covariate-by-covariate matrix converges at the rate $\mathcal{O}(p/d^2)$ where
 180 $p \leq d$. Both of the estimations achieve consistency as long as the dimension grows.

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

183 **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
 184 the distributions of \mathcal{Y} given the true parameter $\mathcal{B}_{\text{true}}$ and estimated parameter $\hat{\mathcal{B}}$, respectively. Then,
 185 we have, with probability at least $1 - \exp(C_1 \sum_k p_k)$,*

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

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

187 5 Numerical implementation

188 5.1 Alternating optimization

189 In this section, we introduce an efficient algorithm to solve (6). The optimization (6) is a non-convex
 190 problem because the feasible set \mathcal{P} is non-convex. We utilize a Tucker factor representation of the
 191 coefficient tensor \mathcal{B} and turn the optimization into a block-wise convex problem.

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

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

193 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
194 columns. The optimization (6) can be written as $(\hat{\mathcal{C}}, \{\hat{\mathbf{M}}_k\}) = \arg \max \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K)$ where

$$195 \quad \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K) = \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1, \dots, i_K} b(\theta_{i_1, \dots, i_K}) \text{with } \Theta = \mathcal{C} \times \{\mathbf{M}_1 \mathbf{X}_1, \dots, \mathbf{M}_K \mathbf{X}_K\}.$$

196 We notice that, if any K out of the $K + 1$ blocks of variables are known, then the optimization with
197 respect to the last block of variables reduced to a simple GLM. We therefore choose to iteratively
198 update one block at a time while keeping others fixed. Although a non-convex optimization of
199 this type usually has no guarantee on global optimality, our numerical experiments have suggested
200 high-quality solutions (see Section 6). The full algorithm is provided in supplement.

202 5.2 Rank selection

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

$$206 \quad \hat{r} = \arg \min_{\mathbf{r}=(r_1, \dots, r_K)} \text{BIC}(\mathbf{r}) = \arg \min_{\mathbf{r}=(r_1, \dots, r_K)} [-2\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) + p_e(\mathbf{r}) \log(\prod_k d_k)], \quad (9)$$

207 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
208 choose \hat{r} that minimizes $\text{BIC}(\mathbf{r})$ via grid search. Our choice of BIC aims to balance between the
209 goodness-of-fit for the data and the degree of freedom in the population model. We test its empirical
performance in Section 6.

210 6 Simulation

211 We evaluate the empirical performance of our generalized tensor regression through simulations. We
212 consider order-3 tensors with a range of distribution types. The coefficient tensor \mathcal{B} is generated using
213 the factorization form (8) where both the core and factor matrices are drawn i.i.d. from Uniform[-1,1].
214 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
215 matrix (i.e. no covariate available) or Gaussian random matrix with i.i.d. entries from $N(0, \sigma_k^2)$. We
216 set $\sigma_k = \sqrt{d_k}$ to ensure the singular values of \mathbf{X}_k are bounded as d_k increases. The \mathcal{U} is scaled
217 such that $\|\mathcal{U}\|_\infty = 1$. Conditional on the linear predictor $\mathcal{U} = [\![y_{ijk}]\!]$, the entries in the tensor
218 $\mathcal{Y} = [\![y_{ijk}]\!]$ are drawn independently according to one of the following three probabilistic models:
219 (a) Gaussian entries $y_{ijk} \sim N(\alpha u_{ijk}, 1)$; (b) Poisson entries $y_{ijk} \sim \text{Poi}(e^{\alpha u_{ijk}})$; (c) binary entries
220 $y_{ijk} \sim \text{Ber}(e^{\alpha u_{ijk}} / (1 + e^{\alpha u_{ijk}}))$. Here $\alpha > 0$ is a scalar controlling the magnitude of the effect size.
221 In each simulation study, we report the mean squared error (MSE) for the coefficient tensor averaged
222 across $n_{\text{sim}} = 30$ replications.

223 6.1 Finite-sample performance

224 The experiment I evaluates the accuracy when covariates are available on all modes. We set $\alpha =$
225 $10, d_k = d, p_k = 0.4d_k, r_k = r \in \{2, 4, 6\}$ and increase d from 25 to 50. Our theoretical analysis
226 suggests that $\hat{\mathcal{B}}$ has a convergence rate $\mathcal{O}(d^{-2})$ in this setting. Figure 2a plots the estimation error
227 versus the ‘‘effective sample size’’, d^2 , under three different distribution models. We found that
228 the empirical MSE decreases roughly at the rate of $1/d^2$, which is consistent with our theoretical
229 ascertainment. We also observed that, tensors with higher ranks tend to yield higher estimation errors,
230 as reflected by the upward shift of the curves as r increases. Indeed, a larger r implies a higher model
231 complexity and thus greater difficulty in the estimation. Similar behaviors can be observed in the
232 non-Gaussian data in Figures 2b-c.

233 The experiment II investigates the capability of our model in handling correlation among coefficients.
234 We mimic the scenario of brain imaging analysis. A sample of $d_3 = 50$ networks are simulated, one
235 for each individual. Each network measures the connections between $d_1 = d_2 = 20$ brain nodes. We
236 simulate $p = 5$ covariates for each of the 50 individuals. These covariates may represent, for
237 example, age, cognitive score, etc. Recent study [22] has suggested that brain connectivity networks
238 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 [23] 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 [24].

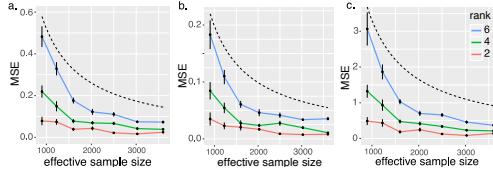


Figure 2: Mean squared error (MSE) against effective sample size. Responses are generated from Gaussian, Poisson and Bernoulli models. The dashed curves correspond to $\mathcal{O}(1/d^2)$.

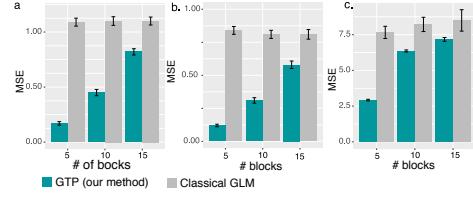


Figure 3: MSE when the networks have block structure. Responses are generated from Gaussian, Poisson and Bernoulli models. The x -axis represents the number of blocks in the networks.

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

The experiment assesses the selection accuracy of our BIC criterion (9) is relegated to supplement.

6.2 Comparison with alternative

We compare our generalized tensor regression (**GTR**) with three other supervised tensor methods: Higher-order low-rank regression (**HOLRR** [5]), Higher-order partial least square (**HOPLS** [7]) and Subsampled tensor projected gradient (**TPG** [6]). These three methods are the closest algorithms to ours. All the three methods allow only Gaussian data, whereas ours is applicable to any exponential family distributions. 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 method. 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 100 for modes with covariates, $d = 20$ for modes without covariates.

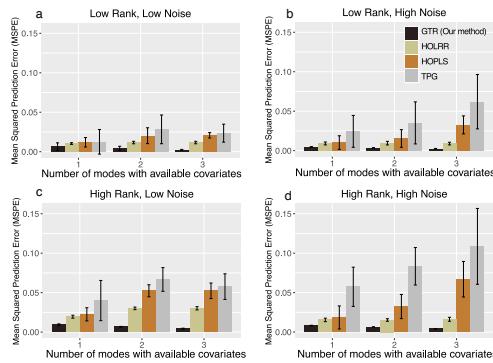


Figure 4: Comparison of MSPE versus the number of modes with covariates. We consider rank $r = (3, 3, 3)$ (low), $r = (4, 5, 6)$ (high), and noise $\sigma = 1/2$ (high), $\sigma = 1/4$ (low).

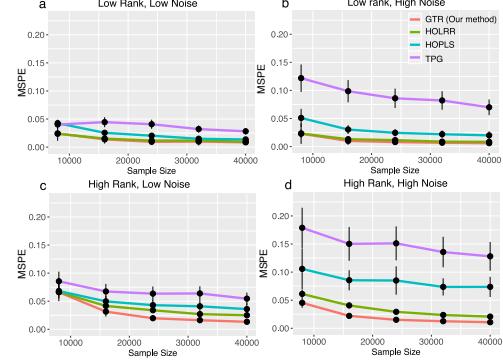


Figure 5: Comparison of MSPE versus sample size. We consider rank $r = (3, 3, 3)$ (low), $r = (4, 5, 6)$ (high), and noise $\sigma = 1/2$ (high), $\sigma = 1/4$ (low).

265 Figure 4 shows the averaged prediction error across 30 replicates. We see that our **GTR** outperforms
 266 others, especially in the high-rank high-noise setting. As the number of informative modes (i.e.
 267 modes with available covariates) increases, the **GTR** exhibits a reduction in error whereas others
 268 have increased errors. This showcases the benefit toward prediction via incorporation of multiple
 269 covariates. The accuracy gain in Figure 4 demonstrates the benefit of alternating algorithm – having
 270 informative modes also improves the estimation along non-informative modes.

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

277 7 Data analysis

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

281 We fit the tensor regression model to the Human connectome project (HCP, [25]) data. The HCP aims
 282 to build a network map that characterizes the anatomical and functional connectivity within healthy
 283 human brains. We take 136 individuals' brain structural networks. Each brain network is represented
 284 as a binary matrix, where the entries encode the presence or absence of fiber connections between 68
 285 brain regions. We consider four individual-covariates: gender, age 22-25, age 26-30, and age 31+.
 286 The BIC suggests a rank $r = (10, 10, 4)$. Figure 6 shows the top edges with high effect size, overlaid
 287 on the Desikan atlas brain template [26]. We depict only the top 3% edges whose connections
 288 are non-constant across samples. Figure 6a show that the global connection exhibits clear spatial
 289 separation, and that the nodes within each hemisphere are more densely connected with each other. In
 290 particular, the superior-temporal (*SupT*), middle-temporal (*MT*) and Insula are the top three popular
 291 nodes in the network. Interestingly, female brains display higher inter-hemispheric connectivity,
 292 especially in the frontal, parietal, and temporal lobes (Figure 6b). This is in agreement with a recent
 293 study showing that female brains are optimized for inter-hemispheric communication [27].

294

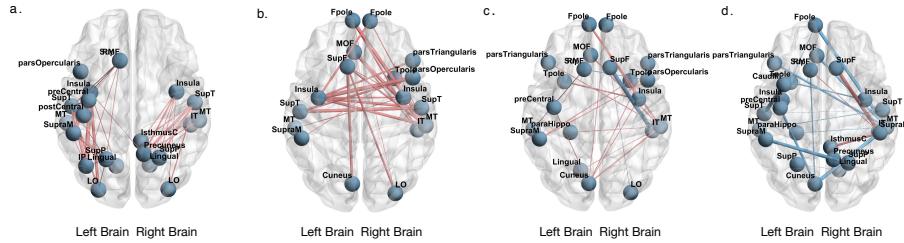


Figure 6: Top edges with large effects. Red edges refer relatively strong connections and blue edges refer relatively weak connections. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+.

295 The second application examines the multi-relational network analysis with node-level attributes.
 296 We consider *Nations* dataset [28] which records 56 relations among 14 countries between 1950
 297 and 1965. The multi-relational networks can be organized into a $14 \times 14 \times 56$ binary tensor. Our
 298 tensor regression results show that the relations reflecting the similar aspects of international affairs
 299 are grouped together. In particular, cluster I consists of political relations such as *officialvisits*,
 300 *intergovorgs*, and *militaryactions*; clusters II and III capture the economical relations; and Cluster IV
 301 represents the Cold War alliance blocs. Detailed results and analysis are in supplement.

302 8 Conclusion

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

309 **References**

- 310 [1] Will Wei Sun and Lexin Li. STORE: sparse tensor response regression and neuroimaging
311 analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.
- 312 [2] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging
313 data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.
- 314 [3] Nicolai Baldin and Quentin Berthet. Optimal link prediction with matrix logistic regression.
315 *arXiv preprint arXiv:1803.07054*, 2018.
- 316 [4] Peter D Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical
317 Association*, 100(469):286–295, 2005.
- 318 [5] Guillaume Rabusseau and Hachem Kadri. Low-rank regression with tensor responses. In
319 *Advances in Neural Information Processing Systems*, pages 1867–1875, 2016.
- 320 [6] Rose Yu and Yan Liu. Learning from multiway data: Simple and efficient tensor regression. In
321 *International Conference on Machine Learning*, pages 373–381, 2016.
- 322 [7] Qibin Zhao, Cesar F Caiafa, Danilo P Mandic, Zenas C Chao, Yasuo Nagasaka, Naotaka Fujii,
323 Liqing Zhang, and Andrzej Cichocki. Higher order partial least squares (HOPLS): a generalized
324 multilinear regression method. *IEEE transactions on pattern analysis and machine intelligence*,
325 35(7):1660–1673, 2012.
- 326 [8] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value
327 decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.
- 328 [9] David Hong, Tamara G Kolda, and Jed A Duersch. Generalized canonical polyadic tensor
329 decomposition. *SIAM Review, in press. arXiv:1808.07452*, 2019.
- 330 [10] Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions
331 on Information Theory*, 2018.
- 332 [11] Han Chen, Garvesh Raskutti, and Ming Yuan. Non-convex projected gradient descent for
333 generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–
334 208, 2019.
- 335 [12] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American
336 Statistical Association*, 112(519):1131–1146, 2017.
- 337 [13] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*,
338 51(3):455–500, 2009.
- 339 [14] Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities
340 between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520:44–
341 66, 2017.
- 342 [15] K Ruben Gabriel. Generalised bilinear regression. *Biometrika*, 85(3):689–700, 1998.
- 343 [16] Richard F Potthoff and SN Roy. A generalized multivariate analysis of variance model useful
344 especially for growth curve problems. *Biometrika*, 51(3-4):313–326, 1964.
- 345 [17] Muni S Srivastava, Tatjana von Rosen, and Dietrich Von Rosen. Models with a kronecker
346 product covariance structure: estimation and testing. *Mathematical Methods of Statistics*,
347 17(4):357–370, 2008.
- 348 [18] Jingfei Zhang, Will Wei Sun, and Lexin Li. Network response regression for modeling popula-
349 tion of networks with covariates. *arXiv preprint arXiv:1810.03192*, 2018.
- 350 [19] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of
351 Mathematics and Physics*, 6(1-4):164–189, 1927.
- 352 [20] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*,
353 33(5):2295–2317, 2011.

- 354 [21] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and
355 Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.
- 356 [22] Lucy F Robinson, Lauren Y Atlas, and Tor D Wager. Dynamic functional connectivity us-
357 ing state-based dynamic community structure: Method and application to opioid analgesia.
358 *NeuroImage*, 108:274–291, 2015.
- 359 [23] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The*
360 *Journal of Machine Learning Research*, 18(1):6446–6531, 2017.
- 361 [24] Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. *Advances in*
362 *Neural Information Processing Systems 32 (NeurIPS 2019)*. arXiv:1906.03807, 2019.
- 363 [25] Linda Geddes. Human brain mapped in unprecedented detail. *Nature*, 2016.
- 364 [26] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for
365 human brain connectomics. *PloS one*, 8(7):e68910, 2013.
- 366 [27] Madhura Ingallikar, Alex Smith, Drew Parker, Theodore D Satterthwaite, Mark A Elliott,
367 Kosha Ruparel, Hakon Hakonarson, Raquel E Gur, Ruben C Gur, and Ragini Verma. Sex
368 differences in the structural connectome of the human brain. *Proceedings of the National*
369 *Academy of Sciences*, 111(2):823–828, 2014.
- 370 [28] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective
371 learning on multi-relational data. In *International Conference on Machine Learning*, volume 11,
372 pages 809–816, 2011.