Exponential Family Tensor Regression with Covariates on Multiple Modes

Jiaxin Hu, Zhuoyan Xu, and Miaoyan Wang
Department of Statistics, University of Wisconsin-Madison

August 30, 2020

Abstract

We consider the problem of tensor-response regression given covariates on multiple modes. Such data problems arise frequently in applications such as neuroimaging, network analysis, and spatial-temporal modeling. We propose a new family of tensor response regression models that incorporate covariates, and establish the theoretical accuracy guarantees. Unlike earlier methods, our estimation allows high-dimensionality in both the tensor response and the covariate matrices on multiple modes. An efficient alternating updating algorithm is further developed. Our proposal handles a broad range of data types, including continuous, count, and binary observations. Through simulation and applications to two real datasets, we demonstrate the outperformance of our approach over the state-of-art.

Keywords: Tensor-response models, Multiple covariates, Dimension reduction, Reduced-rank regression, Generalized linear models

1 Introduction

Nowadays, multi-dimensional arrays, a.k.a. tensors, are collected with additional covariates on multiple modes in many modern scientific and engineering studies. One example is neuroimaging (**??**). The brain connectivity networks are collected from a sample of individuals, accompanied by individual characteristics such as age, gender, and diseases status (see Figure **??**a). Another example is in the field of network analysis (**??**). A typical social network consists of nodes that represent people and edges that represent the friendships. Other covariates such as peoples demographic information and the closeness of a friendship are often available. In both examples, it is of keen scientific interest to identify the variation in the response (e.g., brain connectivity, social community patterns) that may be affected by additional covariates. These seemingly different problems can be formulated as a regression problem in the tensor sense.

Emerging instances of the non-Gaussian datasets also require the effective tools applicable to a wide range of data, such as continuous, binary and count data. Examples include the Political Social science data (**?**) which is a count dataset, and the brain connectivity network data (**?**) which is usually encoded as binary datasets. Easily treating these datasets as Gaussian-valued data will yield unstable estimations and lead the predictions fall outside the valid range. A number of previous methods have been proposed (**????**) to address the tensor regression problem in various forms (e.g. scalar-to-tensor regression, tensor-response regression) from distinct perspectives. However, these methods assume the Gaussianity in the model, treating the input as Gaussian valued entries. Novel methods for exponential family tensor regression are necessary this time.

Related work. Our work is closely related to but also clearly distinctive from several lines of previous work.

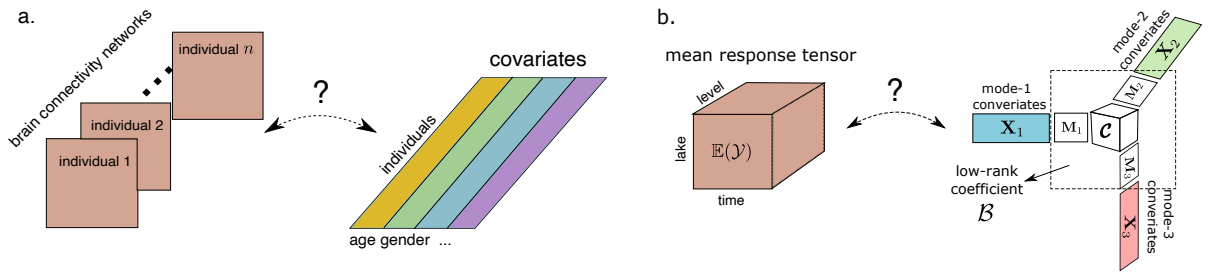

Figure 1: Examples of tensor response regression model with covariates on multiple modes. (a) Network population model. (b) Spatial-temporal growth model.

The first line is a class of *unsupervised* tensor decomposition such as Tucker and CP decomposition (**????**) that aims to find the best low-rank representation of data tensor. In contrast, our model can be viewed a *supervised* tensor learning, which aims to identify the association between a data tensor and covariates. The low-rank structure is determined jointly by the tensor response and matrix covariates.

The second line of work studies the scalar-to-tensor regression models, in which the response is a scalar and the *predictor* is a tensor (**??**). Our proposal is orthogonal to theirs because we treat the tensor as a *response*. The distinct model settings yield distinct focuses and interpretations. Take neuroimaging analysis as an example. The scalar-to-tensor regression aims to study the changes of the prediction outcomes when the brain connectivity network varies, while tensor-response model focuses on understanding the variation in the brain connectivity that may attribute to the individual features such as age, gender and disease status. Since scalar-to-tensor models do not utilize all the information in the response, they are not optimal when there exist correlations among response variables. In contrast, our model does not assume the independence among response entries, which allow the model to capture more structural information.

The third line of work studies the tensor-on-tensor regression models, in which both the response and predictor are tensors (????). Though our model actually tackles the similar problem as tensor-on-tensor regression, our work differs with previous work in several ways. First, previous methods mainly develop the model with Gaussian error. Their models are not suitable for general cases. Second, our study differs from previous work in the scope of results. (?) verifies the performance of estimation mainly via simulations, while our study provides a theoretical evidence on estimation. (??) perform solid inferences on parameter estimation, however, they do not assess the estimation and prediction errors, which are studied in our paper. Risk bounds are well studied in (?) and our convergence rate coincides with their results in the order-3 low-rank case. Whereas, the convexity assumption of the regularization in their model yields a convex optimization problem. In contrast, our model tackles a non-convex optimization problem, and employee distinct techniques.

Our contribution. In this article, we present a general model and the associated theory for exponential family tensor regression. Exponential family tensor observation serves as the response, and the node features and/or the interactions form the matrix predictor. Figure ??b illustrates the general set-up we consider. Due to the generalized framework, our models allows heteroscedasticity by relating the variance of tensor entry to its mean. This exibility is particularly important in practice. The regression approach allows the identication of variation in the data tensor that is explained by the covariates. We utilize a low-rank constraint in the regression coecient to encourage the sharing among tensor entries. The statistical convergence of our estimator is established, and we quantify the gain in predictive power.

2 Preliminaries

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

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

which results in an order-K $(p_1, \ldots, p_K)$-dimensional tensor. For ease of presentation, we use shorthand notion $\mathcal{Y} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$ to denote the tensor-by-matrix product. For any two tensors $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!]$, $\mathcal{Y}' = [\![y'_{i_1,\ldots,i_K}]\!]$ of identical order and dimensions, their inner product is defined as $\langle \mathcal{Y}, \mathcal{Y}' \rangle = \sum_{i_1,\ldots,i_K} y_{i_1,\ldots,i_K} y'_{i_1,\ldots,i_K}$. The Frobenius norm of tensor $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_F = \langle \mathcal{Y},\ \mathcal{Y} \rangle^{1/2}$. A higher-order tensor can be reshaped into a lower-order object (**?**). We use $\mathrm{vec}(\cdot)$ to denote the operation that reshapes the tensor into a vector, and $\mathrm{Unfold}_k(\cdot)$ the operation that reshapes the tensor along mode-k into a matrix of size d_k-by-$\prod_{i \neq k} d_i$. The Tucker rank of an order-K tensor $\mathcal{Y}$ is defined as a length-K vector $\boldsymbol{r} = (r_1, \ldots, r_K)$, where r_k is the rank of matrix $\mathrm{Unfold}_k(\mathcal{Y})$, $k = 1, \ldots, K$. We use lower-case letters (e.g., a, b, c) for scalars/vectors, upper-case boldface letters (e.g., $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$,) for matrices, and calligraphy letters (e.g., $\mathcal{A}, \mathcal{B}, \mathcal{C}$,) for tensors of order three or greater. We let $\boldsymbol{I}_d$ denote the $d \times d$ identity matrix, $[d]$ denote the d-set $\{1, \ldots, d\}$, and allow an $\mathbb{R} \to \mathbb{R}$ function to be applied to tensors in an element-wise manner.

3 Motivation and model

Let $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K data tensor. Suppose we observe covariates on some of the K modes. Let $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ denote the available covariates on the mode k, where $p_k \leq d_k$. We propose a multilinear structure on the conditional expectation of the tensor. Specifically,

$$\begin{aligned} &\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) = f(\Theta), \text{ with} \\ &\Theta = \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}, \end{aligned}$$

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

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.

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

Example 1 (Spatio-temporal growth model)**.** Let $\mathcal{Y} = [\![y_{ijk}]\!] \in \mathbb{R}^{d \times m \times n}$ denote the pH measurements of d lakes at m levels of depth and for n time points. Suppose the sampled lakes belong to q types, with p lakes in each type. Let $\{\ell_j\}_{j \in [m]}$ denote the sampled depth levels and $\{t_k\}_{k \in [n]}$ the time points. Assume that the expected pH trend in depth is a

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

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

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

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

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

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

network-response model (**??**) has the form

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

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

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

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

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

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

The above model has demonstrated its success in modeling transitivity, balance, and communities in the networks (**?**). We show that our tensor regression model (**??**) also incorporates the graph model as a special case. Let $\mathcal{Y} = [\![y_{ij}]\!]$ be a binary matrix where

$y_{ij} = \mathbb{1}_{(i,j)\in E}$. Define $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$. Then, the graph model (**??**) can be expressed as

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

In the above three examples and many other studies, researchers are interested in uncovering the variation in the data tensor that can be explained by the covariates. The regression coefficient $\mathcal{B}$ in our model model (**??**) serves this goal by collecting the effects of covariates and the interaction thereof. To encourage the sharing among effects, we assume that the coefficient tensor $\mathcal{B}$ lies in a low-dimensional parameter space:

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

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

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

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

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

4 Rank-constrained likelihood-based estimation

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

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

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

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

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

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

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

$$\hat{\mathcal{B}} = \underset{\hat{\mathcal{B}}:\ \mathrm{rank}(\mathcal{B}) \leq \boldsymbol{r}, \|\Theta(\mathcal{B})\|_\infty \leq \alpha}{\arg\max} \mathcal{L}_{\mathcal{Y}}(\mathcal{B}).$$

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

4.1 Statistical properties

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

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

In modern applications, the response tensor and covariates are often large-scale. We are particularly interested in the high-dimensional region in which both d_k and p_k diverge; i.e. $d_k \to \infty$ and $p_k \to \infty$, while $\frac{p_k}{d_k} \to \gamma_k \in [0, 1)$. As the size of problem grows, and so does

the number of unknown parameters. As such, the classical MLE theory does not directly apply. We leverage the recent development in random tensor theory and high-dimensional statistics to establish the error bounds of the estimation.

Assumption 1. *We make the following assumptions:*

A1. There exist two positive constants $c_1, c_2 > 0$ *such that* $c_1 \leq \sigma_{\min}(\boldsymbol{X}_k) \leq \sigma_{\max}(\boldsymbol{X}_k) \leq c_2$ *for all* $k \in [K]$. *Here* $\sigma_{min}(\cdot)$ *and* $\sigma_{max}(\cdot)$ *denotes the smallest and largest singular values, respectively.*

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

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

The assumptions are fairly mild. Assumption A1 guarantees the non-singularity of the covariates, and Assumption A2 ensures the log-likelihood $\mathcal{Y}(\Theta)$ is strictly concave in the linear predictor Θ. Assumption A2 and A2' are equivalent, because $\mathrm{Var}(y_{i_1,\ldots,i_K}|\mathcal{X},\mathcal{B}) = \phi b''(\theta_{i_1,\ldots,i_K})$ when $y_{i_1,\ldots,i_K}$ belongs to an exponential family (**?**).

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

$$Loss(\mathcal{B}_{true},\ \hat{\mathcal{B}}) \leq \frac{C_2 \prod_k r_k}{\max_k r_k} \sum_k p_k.$$

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

Theorem **??** establishes the statistical convergence for the estimator (**??**). Actually, the proof in section **??** shows that the statistically optimal rate holds, not only for the MLE (**??**), but also any local estimators $\hat{\mathcal{B}}$ in the level set $\{\hat{\mathcal{B}} \in \mathcal{P} \colon \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) \geq \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{true})\}$ satisfies the upper bound **??**. In Section **??**, we implement empirical studies to verify the algorithmic stability to find a local optimum.

To gain further insight on the bound (**??**), 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 $\boldsymbol{X}_k$ are Gaussian design matrices with i.i.d. $N(0, 1)$ entries. To put the context in the framework of Theorem **??**, we rescale the covariates into $\check{\boldsymbol{X}}_k = \frac{1}{\sqrt{d}} \boldsymbol{X}_k$ so that the singular values of $\check{\boldsymbol{X}}_k$ are bounded by $1 \pm \sqrt{\gamma}$. The result in (**??**) implies that the estimated coefficient has a convergence rate $\mathcal{O}(\frac{p}{d^K})$ in the scale of the original covariates $\{\boldsymbol{X}_k\}$. Therefore, our estimation is consistent as the dimension grows, and the convergence becomes especially favorably as the order of tensor data increases.

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

Corollary 4.1 (Spatio-temporal growth model)**.** *The estimated type-by-time-by-space coefficient tensor converges at the rate $\mathcal{O}\left(\frac{p+r+s}{dmn}\right)$ where $p \leq d$, $r \leq m$ and $s \leq n$. The estimation achieves consistency as long as the dimension grows in either of the three modes.*

Corollary 4.2 (Network population model)**.** *The estimated node-by-node-by-covariate tensor converges at the rate $\mathcal{O}\left(\frac{2d+p}{d^2 n}\right)$ where $p \leq n$. The estimation achieves consistency as the number of individuals or the number of nodes grows.*

Corollary 4.3 (Dyadic data with node attributes)**.** *The estimated covariate-by-covariate matrix converges at the rate* $\mathcal{O}\left(\frac{p}{d^2}\right)$ *where* $p \leq d$*. Again, our estimation achieves consistency as the number of nodes grows.*

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

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

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

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

5 Numerical implementation

5.1 Alternating optimization

In this section, we introduce an efficient algorithm to solve (**??**). The objective function $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})$ is concave in $\mathcal{B}$ when the link f is the canonical link function. We utilize a Tucker factor representation of the coefficient tensor $\mathcal{B}$ and turn the optimization into a block-wise convex problem.

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

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

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

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

The decision variables in the above objective function consist of $K+1$ blocks of variables, one for the core tensor $\mathcal{C}$ and K for the factor matrices $\boldsymbol{M}_k$'s. We notice that, if any K out of the $K+1$ blocks of variables are known, then the optimization with respect to the last block of variables reduced to a simple GLM. This observation suggests that we can iteratively update one block at a time while keeping others fixed. After each iteration, we rescale the core tensor $\mathcal{C}^{(t+1)}$ subject to the maximum norm constraint. This post-processing in principle may not guarantee the monotonic increase of the objective, but we found that in our experiment this simple step appears robust for obtaining a desirable solution. The full algorithm is described in Algorithm **??**.

Note that the feasible set $\mathcal{P}$ is a non-convex set. Therefore, the optimization (**??**) is a non-convex problem, and the Algorithm **??** usually not theoretically possesses the global optimality. However, as mentioned in Section **??**, the desired convergence rate (**??**)

Algorithm 1 Generalized tensor response regression with covariates on multiple modes

Input: Response tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, covariate matrices $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ for $k = 1, \ldots, K$, target Tucker rank $\boldsymbol{r} = (r_1, \ldots, r_K)$, link function f, maximum norm bound α

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

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

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

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

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

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

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

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

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

9: **end for**

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

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

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

13: **end while**

holds for the valid estimators satisfying $\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) \geq \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{true})$, which indicates the global optimality is not necessarily a serious concern in our context, as long as the convergent objective of local optimums are large enough. Fortunately, we find the Algorithm **??** often gives a satisfactory convergence point $\hat{\mathcal{B}}$, when we initialize the parameters via continuous-valued Tucker decomposition. Figure **??** shows the trajectory of the objective function for order-3 tensors under the balanced setting, where $\alpha = 10, d_k = d, p_k = 0.4d, r_k = r$ for $k = 1, 2, 3, p \in \{25, 30\}$ and $r \in \{3, 6\}$. We consider the inputs with Gaussian, Bernoulli, and Poisson entries. Under all combinations of the dimension d, rank r, and type of the entries, Algorithm **??** converges quickly in a few iterations, and the objective value at

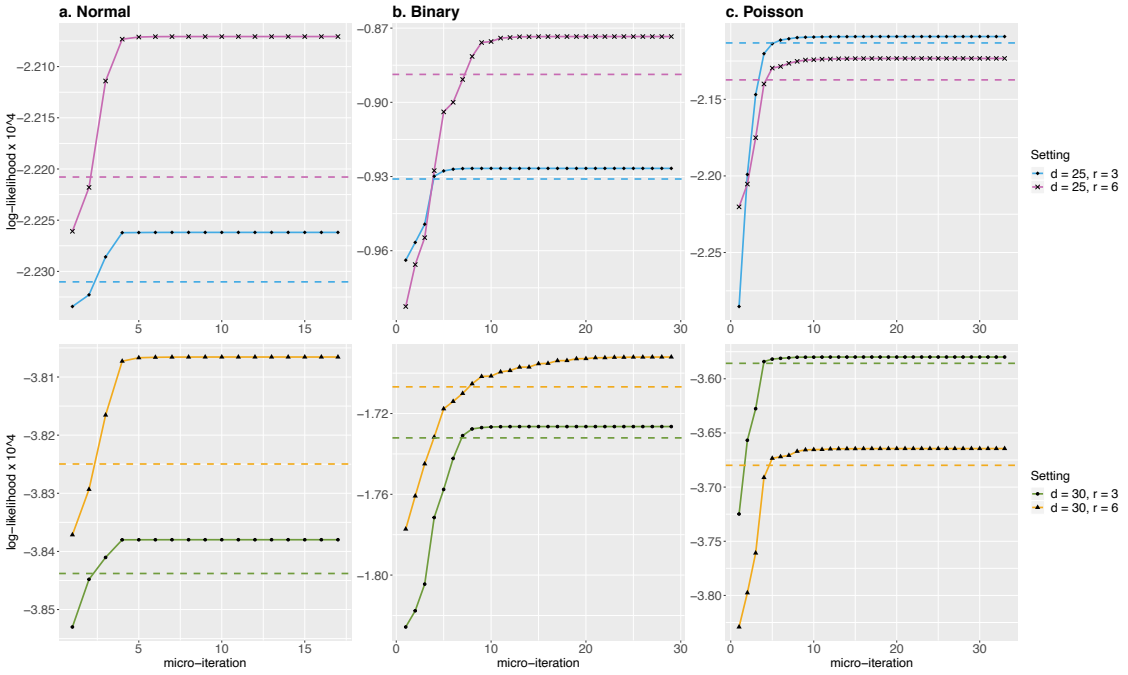

Figure 2: Trajectory of the objective function with various dimension d and rank r under (a) Gaussian (b) Bernoulli (c)Poisson models. The dashed line represents the objective value at true parameter $\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{true})$. A micro-iteration refers to the update for one of the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

convergent points are close to or larger than the value at true parameters.

5.2 Rank selection

Algorithm **??** takes the rank $\boldsymbol{r}$ as an input. Estimating an appropriate rank given the data is of practical importance. We propose to use Bayesian information criterion (BIC) and

choose the rank that minimizes BIC; i.e.

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

where $p_e(\boldsymbol{r}) \stackrel{\text{def}}{=} \sum_k (p_k - r_k - 1)r_k + \prod_k r_k$ is the effective number of parameters in the model. We choose $\hat{\boldsymbol{r}}$ that minimizes $\mathrm{BIC}(\boldsymbol{r})$ via grid search. Our choice of BIC aims to balance between the goodness-of-fit for the data and the degree of freedom in the population model. We test its empirical performance in Section **??**.

5.3 Time complexity

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

6 Simulation

We evaluate the empirical performance of our generalized tensor regression through simulations. We consider order-3 tensors with a range of distribution types. The coefficient tensor $\mathcal{B}$ is generated using the factorization form (**??**) where both the core and factor

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

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

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

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

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

6.1 Finite-sample Performance

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

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

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

rank, and the accuracy improves immediately when the dimension increases to $d = 40$. This agrees with our expectation, as in tensor regression, the sample size is related to the number of entries. A larger d implies a larger sample size, so the BIC selection becomes more accurate.

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

The third experiment investigates our model's ability in handling correlation among coefficients. We mimic the scenario of brain imaging analysis. A sample of $d_3 = 50$ networks are simulated, one for each individual. Each network measures the connections between $d_1 = d_2 = 20$ brain nodes. We simulate $p = 5$ covariates for the each of the 50

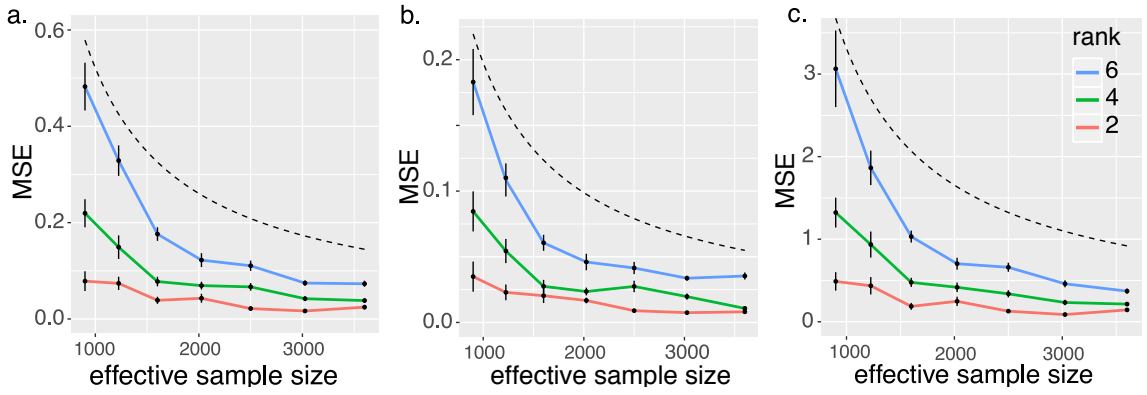

Figure 3: Estimation error against effective sample size. The three panels depict the MSE when the response tensors are generated form (a) Gaussian (b) Poisson and (c) Bernoulli models. The dashed curves correspond to $\mathcal{O}(1/d^2)$.

individuals. These covariates may represent, for example, age, gender, cognitive score, etc. Recent study (**?**) has suggested that brain connectivity networks often exhibit community structure represented as a collection of subnetworks, and each subnetwork is comprised of a set of spatially distributed brain nodes. To accommodate this structure, we utilize the stochastic block model (**?**) to generate the effect size. Specifically, we partition the nodes into r blocks by assigning each node to a block with uniform probability. Edges within a same block are assumed to share the same covariate effects, where the effects are drawn i.i.d. from $N(0, 1)$. We then apply our tensor regression model to the network data using the BIC-selected rank. Note that in this case, the true model rank is unknown; the rank of a r-block matrix is not necessarily equal to r (**?**).

Figure **??** 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 **??**,

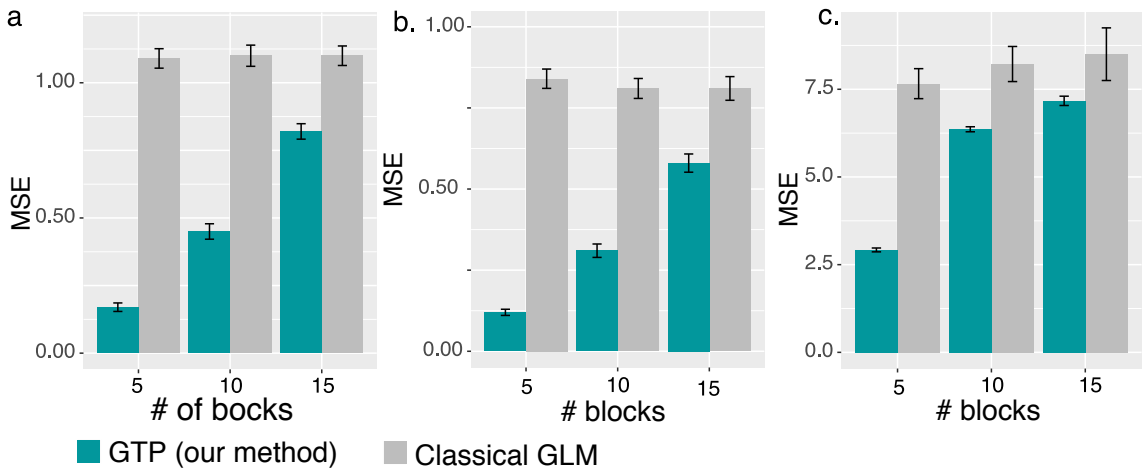

Figure 4: Performance comparison when the networks have block structure. The three panels depict the MSE when the response tensors are generated form (a) Gaussian (b) Poisson and (c) Bernoulli models. The x-axis represents the number of blocks in the networks.

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

6.2 Comparison with alternative methods

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

- Higher-order low-rank regression (**HOLRR**, (**?**)) is a least-square based tensor regression that allows covariates on a single mode.

- Higher-order partial least square (**HOPLS**, (**?**)) is a dimension-reduction method that jointly models a tensor response and a tensor covariate.
- Subsampled tensor projected gradient (**TPG**, (**?**)) tackles the same question as **HOLRR** but instead uses a different algorithm to solve the problem.

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

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

Figure **??** shows the averaged prediction error across 30 replicates. We see that our **GTR** outperforms others, especially in the high-rank high-noise setting. As the number of informative modes (i.e. modes with available covariates) increases, the **GTR** exhibits a reduction in error whereas others have increased errors. This showcases the benefit toward prediction via incorporation of multiple covariates. Note that our method **GTR** is most comparable to **HOLRR** when there is only a single informative mode. In such a case, both methods share a same cost function but have different algorithms. **GTR** alternates between informative and non-informative modes, whereas **HOLRR** approximates

the non-informative modes via unfolded response alone. The accuracy gain in Figure **??** demonstrates the benefit of alternating algorithm – having informative modes also improves the estimation along non-informative modes.

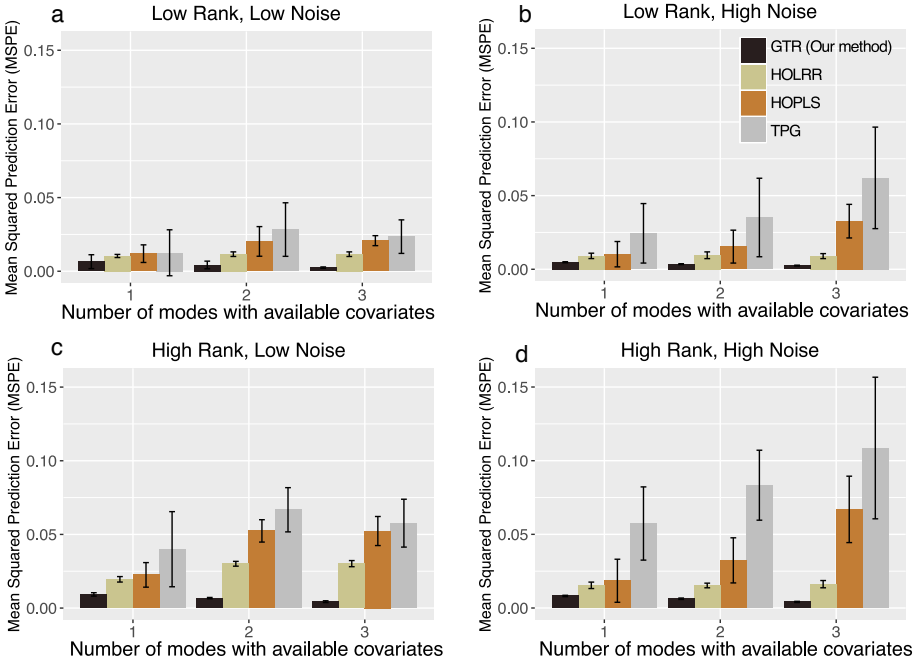

Figure 5: Comparison of MSPE versus the number of modes with covariates. We consider rank $\boldsymbol{r} = (3, 3, 3)$ (low), $\boldsymbol{r} = (4, 5, 6)$ (high), and noise $\sigma = 1/2$ (high), $\sigma = 1/4$ (low).

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

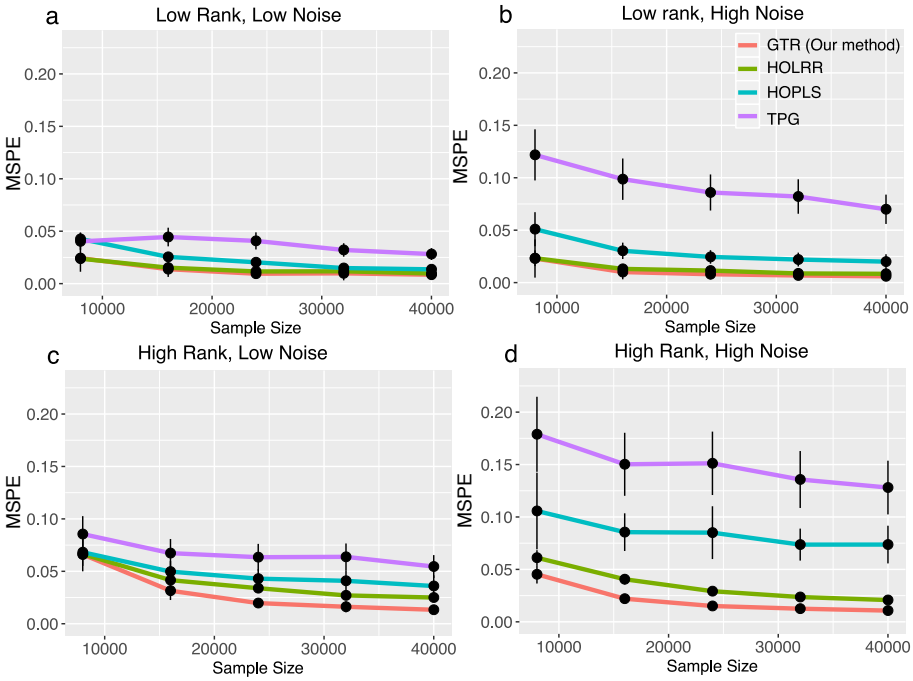

Figure 6: Comparison of MSPE versus sample size. We consider rank $\boldsymbol{r} = (3, 3, 3)$ (low), $\boldsymbol{r} = (4, 5, 6)$ (high), and noise $\sigma = 1/2$ (high), $\sigma = 1/4$ (low).

7 Data analysis

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

7.1 Human Connectome Project (HCP)

The Human connectome project (HCP (**?**)) aims to build a "network map" that characterizes the anatomical and functional connectivity within healthy human brains. We take a subset of HCP data that consists of 136 brain structural networks, one for each individual. Each brain network is represented as a 68-by-68 binary matrix, where the entries encode the presence or absence of fiber connections between the 68 brain regions. We consider four individual-covariates: gender (65 females vs. 71 males), age 22-25 ($n = 35$), age 26-30 ($n = 58$), and age 31+ ($n = 43$). The goal is to identify the connection edges that are affected by the individual covariates. A key challenge in brain network is that the edges are correlated; for example, two edges may stem out from a same brain region, and it is of importance to take into account the within-dyad dependence.

We fit the tensor regression model to the HCP data. The response is a binary tensor $\mathcal{Y} \in \{0,1\}^{68\times 68\times 136}$ and the covariates are of dimension 4 along the 3^{rd} mode. The BIC selection suggests a rank $\boldsymbol{r} = (10, 10, 4)$ with log-likelihood $\mathcal{L}_{\mathcal{Y}} = -174654.7$. Figure **??** shows the top edges with high effect size, overlaid on the Desikan atlas brain template (**??**). We utilize the sum-to-zero contrasts in the effects coding and depicted only the top 3% edges whose connections are non-constant across the sample. It is observed that the global connection exhibits clear spatial separation, and that the nodes within each hemisphere are more densely connected with each other (Figure **??**a). In particular, the superior-temproal ($SupT$), middle-temporal (MT) and Insula are the top three popular nodes in the network. Interestingly, female brains display higher inter-hemispheric connectivity, especially in the frontal, parental and temporal lobes (Figure **??**b). This is in agreement with a recent study showing that female brains are optimized for inter-hemispheric communication (**?**). We also found several edges with declined connection in the group Age 31+. Notably, those edges

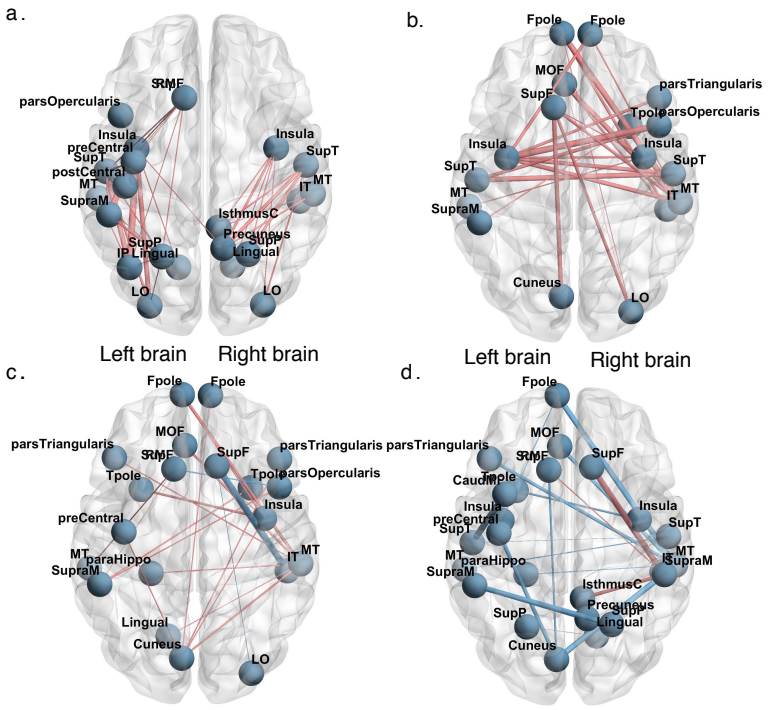

Figure 7: Top edges with large effects. Red edges represent relatively strong connections and blue edges represent relatively weak connections. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+.

involve Frontal-pole (*Fploe*), superior-frontal (*SupF*) and Cuneus nodes. The Frontal-pole region has long been known for its importance in memory and cognition, and the detected decline with age further highlights its biological importance.

7.2 Nations data

The second application concerns the multi-relational network analysis with node-level attributes. We consider *Nations* dataset (**?**) which records 56 relations among 14 countries between 1950 and 1965. The multi-relational networks can be organized into a $14 \times 14 \times 56$

binary tensor, with each entry indicating the presence or absence of a connection, such as "sending tourist to", "export", "import", between countries. The 56 relations span the fields of politics, economics, military, religion, and so on. In addition, country-level attributes are also available, and we focus on the following six covariates: *constitutional, catholics, lawngos, politicalleadership, geographyx*, and *medicinengo*. The goal is to identify the variation in connections due to country-level attributes and interactions thereof. One of the key features is that the 56 relations are correlated, and we would like to take that into account in assessing the covariate effects.

We applied our tensor regression model to the *Nations* data. The multi-relational network $\mathcal{Y} \in \{0,1\}^{14\times14\times56}$ was treated as the response tensor, and the country attributes $\boldsymbol{M} \in \mathbb{R}^{14\times6}$ were treated as covariates on both the 1st and 2nd modes. The BIC criterion suggests a rank $\boldsymbol{r} = (4,4,4)$ for the coefficient tensor $\mathcal{B} \in \mathbb{R}^{6\times6\times56}$. Table Section **??** shows the K-mean clustering of the 56 relations based on the 3^{rd} mode factor $\boldsymbol{M}_3 \in \mathbb{R}^{56\times4}$. We found that the relations reflecting the similar aspects of international affairs are grouped together. In particular, Cluster I consists of political relations such as *officialvisits, intergovorgs*, and *militaryactions*; Clusters II and III capture the economical relations such as *economicaid, booktranslations, tourism*; and Cluster IV represents the Cold War alliance blocs. The similarity among entities in each cluster suggests the plausibility of our dimension reduction.

To investigate the effects of dyadic attributes towards connections, we depicted the estimated coefficients $\hat{\mathcal{B}} = [\![\hat{b}_{ijk}]\!]$ for several relation types (Figure **??**). Note that entries $\hat{b}_{ijk}$ can be interpreted as the contribution, at the logit scale, of covariate pair (i,j) (ith covariate for the "sender" country and jth covariate for the "receiver" country) towards the connection of relation k. Several interesting findings emerge from the observation. We

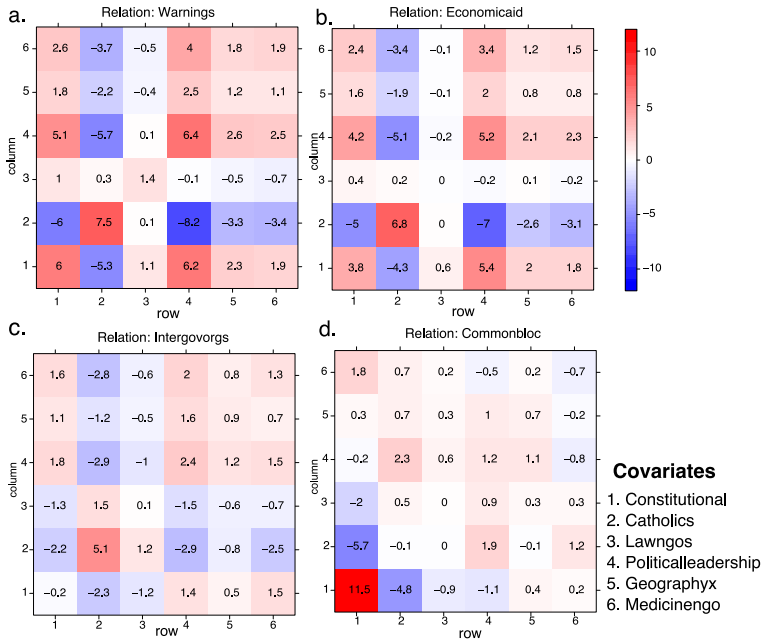

Figure 8: Effect estimation in the *Nations* data. Panels (a)-(d) represent the estimated effects of country-level attributes towards the connection probability, for relations *warnning*, *economicaid*,*intergovorg*, and *commonblock*, respectively.

found that relations belonging to a same cluster tend to have similar covariate effects. For example, the relations *warnings* and *ecnomicaid* are classified into Cluster II, and both exhibit similar covariate pattern (Figure **??**a-b). Moreover, the majority of the diagonal entries $\hat{\mathcal{B}}(i, i, k)$ positively contribute to the connection. This suggests that countries with coherent attributes tend to interact more often than others. We also found that the *constitutional* attribute is an important predictor for the *commonbloc* relation, whereas the effect is weaker for other relations (Figure **??**d). This is not surprising, as the block partition during Cold War is associated with the *constitutional* attribute.

8 Conclusion

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

SUPPLEMENTARY MATERIAL

A Proofs

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

equation (**??**) implies that

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

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

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

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

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

where the second line follows from (**??**). Therefore,

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

Combining (**??**) with C1 yields

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

Therefore, the final conclusion follows by noting that

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

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

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

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

Note that,

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

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

Proposition 2 (sub-Gaussian residuals)**.** *Define the residual tensor* $\mathcal{E} = [\![\varepsilon_{i_1,\ldots,i_K}]\!] = \mathcal{Y} - b'(\Theta) \in \mathbb{R}^{d_1 \times \cdots \times d_K}$*. Under the Assumption A2,* $\varepsilon_{i_1,\ldots,i_K}$ *is a sub-Gaussian random variable with sub-Gaussian parameter bounded by* ϕU*, for all* $(i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]$*.*

Proof. The proof is similar to Lemma 3 in (**?**). For ease of presentation, we drop the subscript $(i_1, \ldots, i_K)$ and simply write ε $(= y - b'(\theta))$. For any given $t \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}(\exp(t\varepsilon|\theta) &= \int c(x) \exp\left(\frac{\theta x - b(\theta)}{\phi}\right) \exp\left(t(x - b'(\theta))\right) dx \\
&= \int c(x) \exp\left(\frac{(\theta + \phi t)x - b(\theta + \phi t) + b(\theta + \phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right) dx \\
&= \exp\left(\frac{b(\theta + \phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right) \\
&\leq \exp\left(\frac{\phi U t^2}{2}\right),
\end{aligned}$$

where $c(\cdot)$ and $b(\cdot)$ are known functions in the exponential family corresponding to y. Therefore, ε is sub-Gaussian-(ϕU). □

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

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

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

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

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

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

B Additional results for real data analysis

B.1 HCP data analysis

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

B.2 Nations data analysis

Table **??** summarizes the K-means clustering of the 56 relations based on the 3^{rd} mode factor $\boldsymbol{M}_3 \in \mathbb{R}^{56\times 4}$ in the tensor regression model.

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

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

Acknowledgements

This research was supported by NSF grant DMS-1915978 and the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. The authors would like to thank Chen Zhang from Wisconsin State Lab of Hygiene for help with figure 4.

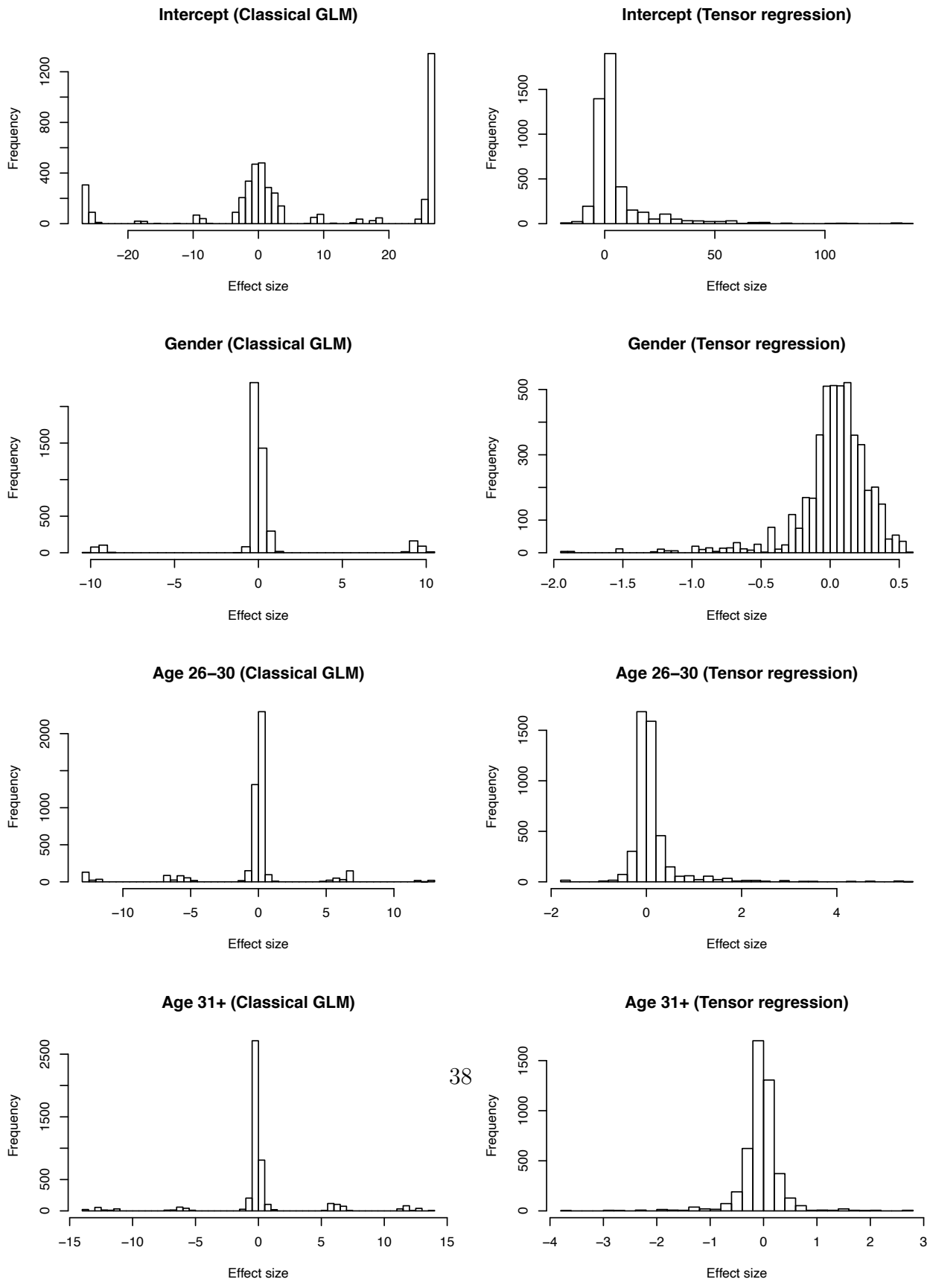

Figure 9: Comparison of coefficient estimation in the HCP data.