Generalized tensor response regression with multi-sided covariates

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

Abstract

We consider the problem of tensor-valued regression given covariates on a set of modes. Such data problems arise frequently arise in applications such as neuroimaging, network analysis, and spatial-temporal modeling. We propose a new family of tensor response regression models that incorporate covariate information, and obtain the theoretical accuracy guarantees. 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 application to two real datasets, we demonstrate the outperformance of our approach over the state-of-art.

1 Introduction

Many contemporary scientific and engineering studies collect multi-way array data, a.k.a. tensor, accompanied by additional covariates. For example, in neuro-imaging analysis, researchers measure brain connections from a sample of individuals with the goal to identify the brain edges affected by age and gender. In social network analysis, how to explain the connection (e.g. community, transitive, etc.) by attributable of both nodes. In this article, we provide a general treatment to these seemingly different problems.

2 Preliminaries

We begin by reviewing basic properties on tensors [1]. 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 $\boldsymbol{X}_k = [\![x^{(k)}_{i_k,j_k}]\!] \in \mathbb{R}^{s_k\times d_k}$ is defined as

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

which results in an order-K tensor $(s_1,\ldots,s_K)$-dimensional tensor. For ease of notation, we use shorthand notion $\mathcal{Y}\times\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K\}$ to denote the above product. A higher-order tensor can be reshaped into a lower-order representation. We let $\text{vec}(\cdot)$ to denote the operation that reshapes a tensor into a vector, and $\text{Unfold}_k(\cdot)$ the operation that reshapes a tensor along mode-k into a matrix of size d_k-by-$\prod_{i\neq k} d_i$. The Tucker rank of an order-K tensor $\mathcal{Y}$ is defined as a length-K vector $\boldsymbol{r}=(r_1,\ldots,r_K)$, where r_k is the rank of matrix $\text{Unfold}_k(\mathcal{Y})$.

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 $\|\mathcal{Y}\|_F = \langle \mathcal{Y},\ \mathcal{Y}\rangle^{1/2}$; it is the Euclidean norm of $\mathcal{Y}$ regarded as an $\prod_k d_k$-dimensional vector. We use lower-case letters, e.g., a,b,c, for scalars and vectors, upper-case boldface letters, e.g., $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$, for matrices, and calligraphy letter, e.g. $\mathcal{A},\mathcal{B},\mathcal{C}$, for tensors of order 3 or greater. We denote by $\boldsymbol{I}_n$ the identity the $d\times d$ identity matrix.

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 of interest. 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 the following multilinear structure for the mean tensor. Specifically,

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

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

Preliminary work. Under review by AISTATS 2020. Do not distribute.

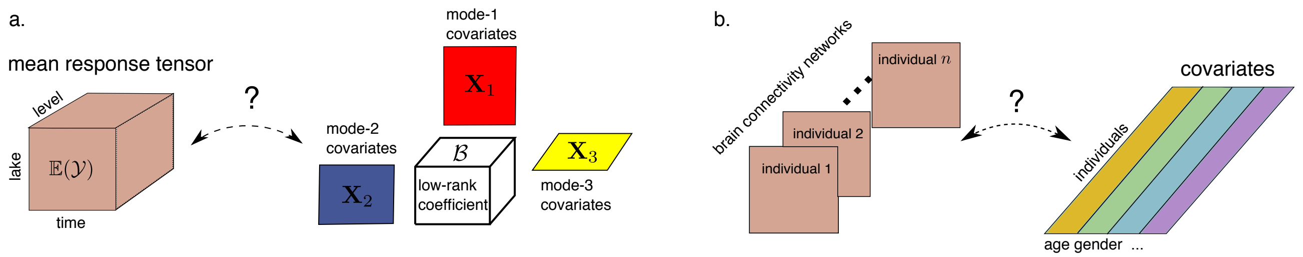

Figure 1: Examples of tensor regression with multi-sided covariates. (a) Spatio-temproal growth model. (b) Network population model.

are identity link for Gaussian tensor, logistic link for binary tensor, and log link for Poisson tensor. We give three concrete examples of tensor regression model that arises 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 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

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

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

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

are the design matrices for spatial and temporal effects, respectively. Clearly, the spatial-temporal model is a special case of our tensor regression model, with covariates on each of the three modes. The model (2) can be viewed as a higher-order extension of the classical "growth curve" models for matrix data in the context of factorial design [2, 3, 4].

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 $\boldsymbol{Y}_i \in \{0,1\}^{d\times n}$ 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 [5, 6] has the form

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

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

The model (3) is a special case of our tensor-response model, with covariate on one of the three modes. Specifically, let $\mathcal{Y} \in \{0,1\}^{d\times d\times n}$ denote the response tensor by stacking $\{\boldsymbol{Y}_i\}$ together along the 3rd mode and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p\times n}$, then model (3) can be expressed 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 data consist 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 $V = [n]$ is the node set of the graph, and $E \subset V \times V$ is the edge set. We also observe covariate $x_i \in \mathbb{R}^p$ associated to each $i \in V$. The network $\mathcal{G} = (V, E)$ is described by the following matrix model. The edge connects the two vertices i and j independently of the others, and the probability of connection is modeled as

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

The above model has demonstrated its success in explaining higher-order dependence, such as transitivity, balance, and clusterability in the network data [7]. We show that our tensor regression model incorporates the dyadic model as a special case. Let $\mathcal{Y} = [\![y_{ij}]\!]$ where $y_{ij} = \mathbb{1}_{(i,j)\in E}$. Define $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p\times n}$. Then, the dyadic model (4) can be expressed as

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

In the above three examples and many other studies, researchers are interested in uncovering the variation in the data tensor that are explained by the covariates. The regression coefficient $\mathcal{B}$ in our model model (5) serves this goal by revealing the effects that are attributable to covariates and the interaction thereof. Furthermore, we assume that the coefficient tensor $\mathcal{B}$ admits a low-rank Tucker decomposition,

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

where $r_k(\mathcal{B}) \leq p_k$ is the Tucker rank of the tensor at mode k. The low-rank assumption is plausible in many scientific applications. In brain imaging analysis, for instance, it is often believed that the nodes can be classified into fewer communities, and the numbers of communities are much smaller than the dimension. Moreover, the low-rank structure encourages the shared information across tensor entries, thereby greatly improving the estimation stability.

Our tensor regression model is able to incorporate covariates on some or all 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:

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

where $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the low-rank coefficient tensor of interest, and the entries of $\mathcal{Y}$ are independent conditional on $\mathcal{X}$. We comment that other forms of tensor low-rankness is also possible, and here we choose Tucker rank for parsimony. Similar models can be derived using other notions of low-rankness based on CP decomposition [8] and train decomposition [9].

4 Rank-constrained likelihood-based estimation

We develop a likelihood-based procedure to estimate the tensor. The exponential family is a flexible framework for different data types. In a classical GLM with a scalar response y and covariate $\boldsymbol{x}$, the density is expressed as:

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

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 for continuous data is $\mathbb{Y} = \mathbb{R}$, for count data $\mathbb{Y} = \mathbb{N}$, and for binary data, $\mathbb{Y} = \{0, 1\}$. 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.

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 link functions for various distribution types.

In our context, the entries in the response tensor y_{ijk} conditional on θ_{ijk} are independent drawn from exponential family. The quasi log-likelihood of (5) is equal (ignoring constant) to Bregman distance between $\mathcal{Y}$ and $b'(\Theta)$:

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

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 imaging where fiber connections are bounded. We propose a constrained maximum likelihood estimator for the coefficient tensor:

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

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

4.1 Statistical properties

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

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

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 maximum likelihood estimation 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 tensor 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 $L \leq Var(y_{i_1, \ldots, i_K}|\mathcal{X}, \mathcal{B}) \leq U$ uniformly over the parameter space $\mathcal{P}$.

Algorithm 1 Generalized tensor response regression with multi-sided covariates

```
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 $(r_1, \ldots, r_K)$, link function $f$, entrywise bound $\alpha$
Output: Estimated low-rank coefficient tensor $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$.
Calculate $\check{\mathcal{B}} = \mathcal{Y} \times_1 \left[(\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T\right] \times_2 \cdots \times_K \left[(\boldsymbol{X}_K^T \boldsymbol{X}_K)^{-1} \boldsymbol{X}_K^T\right]$.
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-$(r_1, \ldots, r_K)$ Tucker approximation of $\check{\mathcal{B}}$, in the least-square sense.
while the relative increase in objective function $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})$ is less than the tolerance do
    Update iteration index $t \leftarrow t + 1$.
    for  $k = 1$ to $K$ do
        Obtain the factor matrix $\boldsymbol{M}_k^{(t+1)} \in \mathbb{R}^{p_k \times r_k}$ by solving $d_k$ separate GLMs with link function $f$.
        Update the columns of $\boldsymbol{M}_k^{(t+1)}$ by Gram-Schmidt orthogonalization.
    end for
    Obtain the core tensor $\mathcal{C}^{(t+1)} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ by solving a GLM with $\text{vec}(\mathcal{Y})$ as response, $\odot_{k=1}^{K}[\boldsymbol{X}_k \boldsymbol{M}_k^{(t)}]$ as covariates, and $f$ as link function.
    Rescale the core tensor subject to the entrywise bound constraint.
    Update $\mathcal{B}^{(t+1)} \leftarrow \mathcal{C}^{(t+1)} \times_1 \boldsymbol{M}_1^{(t+1)} \times_2 \cdots \times_K \boldsymbol{M}_K^{(t+1)}$.
end while
```

Both 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. Note that, for exponential family, $\text{Var}(y_{i_1,\ldots,i_K}|\mathcal{X}, \mathcal{B}) = b''(\theta_{i_1,\ldots,i_K})$ [10], so the Assumption A2 can also be replaced by $L \leq b''(\theta) \leq U$ for all $|\theta| \leq \alpha$, where α is the upper bound of the linear predictor θ.

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

$$Loss(\mathcal{B}_{true},\ \hat{\mathcal{B}}) \leq C_3 \sum_k p_k, \tag{7}$$

where $C_3 = C_3(\boldsymbol{r}) = \frac{1}{C_2^{2K} U} \frac{\prod_k r_k}{\max_k r_k} > 0$ is a constant that does not depend on the dimensions $\{d_k\}$ and $\{p_k\}$.

To gain further insight on the bound (7), we consider a special case when 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. $N(0, \sigma^2)$ entries. To put the result in the context of Theorem 4.1, we define rescaled covariates $\check{\boldsymbol{X}}_k = \frac{1}{\sqrt{d}} \boldsymbol{X}_k$, and note that the singular values of $\check{\boldsymbol{X}}_k$ are bounded in $1 \pm \sqrt{\gamma}$. The bound (7) then reduces to $\mathcal{O}(\frac{p}{d^K})$ for the estimated coefficient under the original, unscaled covariates $\{\boldsymbol{X}_k\}$. Therefore, our estimation is consistent as the dimension grows, and the convergence becomes especially favorably as we have more modes in the tensor data.

As immediate applications, we obtain the convergence rate for the three examples mentioned in Section 3. Without loss of generality, we assume that the singular values of the (informative) covariates $\boldsymbol{X}_k$ are scaled as $\sqrt{d_k}$.

Corollary 1 (Spatio-temporal growth model)**.** The estimated type-by-time-by-level 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 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 3 (Link model 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 for the response tensor.

Theorem 4.2 (Prediction error)**.** *Assume the same set-up as in Theorem 4.1. Let $\mathbb{P}_{\mathcal{Y}_{true}}$ and $\mathbb{P}_{\hat{\mathcal{Y}}}$ denote the distribution of $\mathcal{Y}$ given the $\mathcal{B}_{true}$ and $\hat{\mathcal{B}}$, respectively. Similarly, let $\mathbb{E}(\mathcal{Y}|\mathcal{X})$ and $\widehat{\mathbb{E}(\mathcal{Y}|\mathcal{X})}$ denote, respectively, the true and estimated mean. We have, with probability* *at least* $1-\exp(C_1\sum_k p_k)$,

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

$$\textit{Loss}\left(\mathbb{E}(\mathcal{Y}|\mathcal{X}),\ \widehat{\mathbb{E}(\mathcal{Y}|\mathcal{X})}\right) \leq C_5 \sum_k p_k,$$

where $C_4, C_5 > 0$ *are two constants that do not depend on the dimension* $\{d_k\}$ *and* $\{p_k\}$.

4.2 Comparison with existing work

In this section, we discuss the comparison between our method and previous work. A fundamental feature of tensor-valued data is the statistical interdependence among entries, and one of our goals is to quantify this interdependence.

5 Numerical implementation

5.1 Alternating optimization

In this section, we introduce an efficient algorithm to solve (6). The objective function $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})$ is concave in $\mathcal{B}$ when the link f is canonical link function. However, the feasible set $\mathcal{P}$ is non-convex, and thus the optimization (6) is a non-convex problem. We utilize a Tucker factor representation of 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

$$\mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \dots, \boldsymbol{M}_K\}, \tag{8}$$

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 whose columns are orthogonal. Estimating $\mathcal{B}$ amounts to finding both the core tensor $\mathcal{C}$ and the factor matrices $\boldsymbol{M}_k$'s. Then, the optimization (6) can be written as $(\hat{\mathcal{C}}, \{\hat{\boldsymbol{M}}_k\}) = \arg\max \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, \dots, \boldsymbol{M}_K)$, where

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, \dots, \boldsymbol{M}_K) = \langle \mathcal{Y}, \Theta\rangle - \sum_{i_1,\dots,i_K} b(\theta_{i_1,\dots,i_K}), \tag{9}$$
$$\text{with } \Theta = \mathcal{C} \times \{\boldsymbol{M}_1\boldsymbol{X}_1, \dots, \boldsymbol{M}_K\boldsymbol{X}_K\}.$$

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. Specifically, suppose the core tensor and the factor matrix $\boldsymbol{M}_k$ are known for $k = 1, \dots, K-1$. It turns out that last factor matrix $\boldsymbol{M}_K$ can be solved in a row-by-row fashion via d_k separate GLMs. To see this, let $\mathcal{C}^{(t)}$ denote the core tensor at the t-th iteration, $\boldsymbol{M}_k^{(t)}$ the kth factor matrix for $k \in [K-1]$ at the t-th iteration, and define

$$\boldsymbol{X}_{-K}^{(t)} = \mathcal{C}^{(t)} \times \{\boldsymbol{M}_1^{(t)}\boldsymbol{X}_1,\ \dots,\ \boldsymbol{M}_{K-1}^{(t)}\boldsymbol{X}_{K-1}\},$$

Then, the objective (9) implies that the i-th row of $\boldsymbol{M}_K^{(t)}$ is the "regression coefficient" for a GLM whose response vector is $\text{vec}(\mathcal{Y}(:,\ :,\ i)) \in \mathbb{R}^{d_{-K}}$ and covariate matrix is $\text{Unfold}_K(\boldsymbol{X}_{-K}^{(t)}) \in \mathbb{R}^{d_{-K}\times r_K}$, where $d_{-K} \overset{\text{def}}{=} \prod_{k\in[K-1]} d_k$. The property of separation by row allows us to leverage state-of-art GLM solvers and parallel processing to achieve computational efficiency. After each iteration, we rescale the core tensor $\mathcal{C}^{(t+1)}$ subject to the infinity 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 post-processing appears to be good enough for a desirable solution. The full algorithm is described in Algorithm 1.

5.2 Rank selection, missing data handling

Before concluding this section, we briefly comment on two implementation details. First, Algorithm 1 takes the rank $\boldsymbol{r}$ as an input. Estimating an appropriate rank given the data is of practical importance. We propose to use Bayesian information criterion (BIC) and choose the rank that minimizes BIC; i.e.

$$\begin{aligned}\hat{\boldsymbol{r}} &= \underset{\boldsymbol{r}=(r_1,\dots,r_K)}{\arg\min}\ \text{BIC}(\boldsymbol{r}) \qquad\qquad (10)\\ &= \underset{\boldsymbol{r}=(r_1,\dots,r_K)}{\arg\min}\ \left[-2\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) + p_e(\boldsymbol{r})\log\left(\prod_k d_k\right)\right],\end{aligned}$$

where $p_e(\boldsymbol{r}) \overset{\text{def}}{=} \sum_k (p_k - r_k - 1)r_k + \prod_k r_k$ is the effective number of parameters in the model. We choose $\hat{\boldsymbol{r}}$ that minimizes $\text{BIC}(\boldsymbol{r})$ via grid search. Our choice of BIC aims to balance between the goodness-of-fit for the data and the degree of freedom in the population model. We test its empirical performance in Section 6.

Second, when some response entries $y_{i_1,\dots,i_K}$ are missing, we replace the objective function by $\mathcal{L}_{\mathcal{Y}} = \sum_{(i_1,\dots,i_K)\in\Omega} (y_{i_1,\dots,i_K}\theta_{i_1,\dots,i_K} - b(\theta_{i_1,\dots,i_K}))$, where $\Omega \subset [d_1]\times\cdots\times[d_K]$ is the index set of non-missing entries. That is, we model the observed response entries and exclude the missing entries in the fitting. Similar strategy has been used for classical (unsupervised) Tucker and CP tensor decomposition with missing data [11, 12, 13]. In the presence of missing response, we modify line 7 in Algorithm 1 by fitting GLMs to the data for which $y_{i_1,\dots,i_K}$ are observed. This approach requires there are no entirely missing slice, e.g. in the form of $\mathcal{Y}(:,\ :,\ k)$ for order-3 tensors. We regard this as a fairly mild condition akin to the coherence condition as in the completion literature [14, 15].

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 (8) where both the core and factor matrices are drawn i.i.d. from Uniform[0,1]. The linear predictor is then simulated as $\mathcal{U} = \mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$, where $\boldsymbol{X}_k$ is either an identity matrix (i.e. no covariate available) or Gaussian random matrix with entries i.i.d. from $N(0, \sigma^2)$. Without loss of generality, we set $\sigma^2 = d_k^{-1/2}$ to ensure the singular values of $\boldsymbol{X}_k$ are bounded as d_k increases. We rescale $\mathcal{U}$ such that $\|\mathcal{U}\|_\infty = 1$. Conditional on the linear predictor $\mathcal{U} = [\![u_{ijk}]\!]$, the entries in 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 linear predictor. In each simulation study, we report the mean squared error for the coefficient tensor averaged across $n_{\text{sim}} = 30$ replications.

The first experiment assesses the selection accuracy of our BIC criterion (10). We consider the balanced situation where $d_k = d$, $p_k = 0.4d_k$ for $k = 1, 2, 3$. We set $\alpha = 10$, $\alpha = 4$, 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 over $\boldsymbol{r}$ using a grid search over three dimensions. We set the hyper-parameter α to infinity in the fitting, which essentially poses no prior on the coefficient magnitude. Table 2 reports the selected rank averaged over $n_{\text{sim}} = 30$ replicates for Gaussian and Poisson models. (The results for Bernoulli model is shown in the Supplements.) We found that when $d = 20$ the selected rank is slightly smaller than the true rank, and the accuracy increases 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.

True Rank	Dimension (Gaussian tensors)		Dimension (Poisson tensors)	
$\boldsymbol{r}$	$d = 20$	$d = 40$	$d = 20$	$d = 40$
(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: Performance for 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.

The second experiment evaluates the accuracy when covariates are available on all modes. We set $\alpha = 10, d_k = d, r_k = r \in \{2, 4, 6\}$ and increase d from 25 to 50. Our theoretical analysis suggests $\hat{\mathcal{B}}$ 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 RMSE decreases roughly at the rate of $1/d^2$, which is consistent with our theoretical ascertainment. We also observed that coefficients 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 higher model complexity, thus increasing the difficulty of the estimation. Similar behaviors can be observed in the non-Gaussian data in Figure 2b-c.

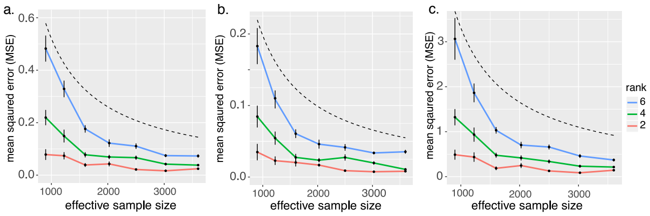

Figure 2: Estimation error against effective sample size. The three panels depicts the MSE when the response tensor is generated form (a) Gaussian (b) Poisson and (c) Bernoulli models. Each solid curve corresponds to a fixed rank. The dashed curve corresponds to $\mathcal{O}(1/d^2)$.

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 50 individuals. These covariates may represent, for example, age, gender, cognitive score, etc. Recent study [16] 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 popular stochastic block model to generate the effect size. Specifically, we created r blocks among the nodes by randomly assigning each node to a cluster 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 applied 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 bounded by but not necessarily equal to r [17].

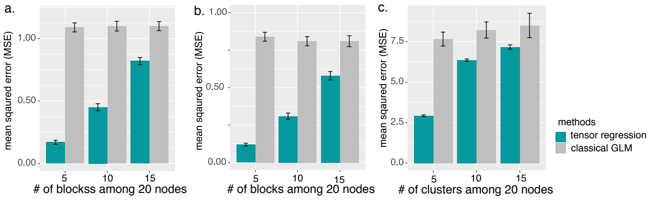

Figure 3: Performance comparison when the population network admit block structure. The three panels depicts the MSE when the response tensor is generated form (a) Gaussian (b) Poisson and (c) Bernoulli models. The x-axis represents the number of blocks in the networks.

Figure 3 compares the MSE of our model 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 approach, however, does not account for the correlation structure among the edges. As we can see in Figure 3, out tensor regression method achieves significant error reduction in all three models considered. The outer-performance is more apparent 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 a data-driven rank selection, our method is able to achieve accurate estimation with improved interpretability.

7 Data analysis

We apply our tensor regression model to two real datasets. The first application concerns the modeling of brain network population 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 took 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. Individual covariates are also available. For simplicity of presentation, we consider four 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.

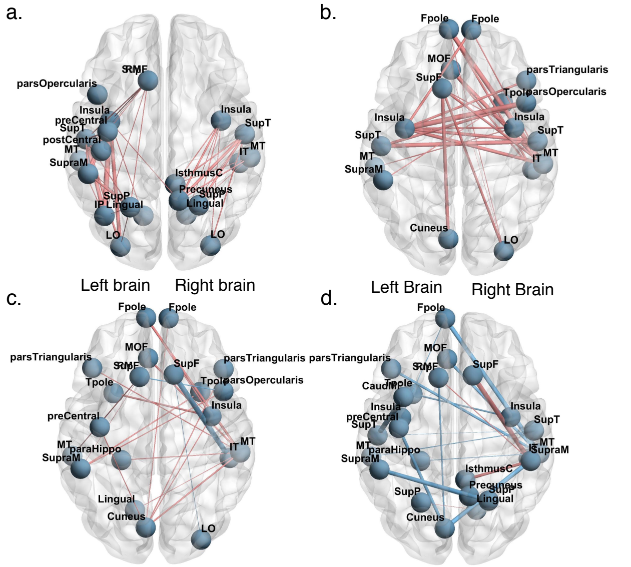

Figure 4: Top edges with high effect size. (a) Global effect; (b) Gender effect; (c) Age 22-25; (c) Age 31+

We fitted 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 3rd mode. The BIC selection suggests a rank $\boldsymbol{r} = (10, 10, 4)$ with log-likelihood $\mathcal{L}_{\mathcal{Y}} = -174654.7$. Figure 4 shows the top edges with high effect size, overlaid on the Desikan atlas brain template [18, 19]. For better interpretation, we utilized the sum-to-zero contrasts in the effects coding and depicted only the top 3% edges with non-constant connections in the study sample. It is observed that the global connection exhibit clear spatial separation, and that the nodes within each hemisphere are more densely connected with each other (Figure 4a). In particular, the superiortemproal ($SupT$), middletemporal (MT) and Insula are the top three popular nodes in the network. Interestingly, female brains displays higher inter-hemispheric connectivity, especially in the frontal, parental and temporal lobes (Figure 4b). This is in agreement with a recent study showing that female brains are optimized for inter-hemispheric communication [20]. We also found several edges with declined connection in the group Age 31+. Notably, those edges involves Frontalpole ($Fploe$), superiorfrontal ($SupF$) and Cuneus nodes. The Frontalpole region has long been known for its importance in memory and cognition, and the declined fiber connections in Frontalpole node may suggests its anatomical change with age.

7.2 Nations data

The second application concerns the multi-relational network analysis with node-level attributes. We consider *Nations* dataset [21] 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 with two-sided covariates to the *Nations* data. The multi-relational network $\mathcal{Y} \in \{0,1\}^{14\times 14\times 56}$ was treated as the response tensor, and the country attributes $\boldsymbol{M} \in \mathbb{R}^{14\times 6}$ were treated as covariates in both the 1st and 2nd modes. The BIC criterion suggests a rank $\boldsymbol{r} = (4,4,4)$ for the coefficient tensor $\mathcal{B} \in \mathbb{R}^{6\times 6\times 56}$. Supplementary Table S1 shows the K-mean clustering of the 56 relations based on the 3rd mode factor $\boldsymbol{M}_3 \in \mathbb{R}^{56\times 4}$. 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 are dominated by economical relations including *economicaid, booktranslations, tourism, etc*; 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 how the dyadic attributes affect the connection, we depicted the estimated coefficients $\hat{\mathcal{B}} = [\![\hat{b}_{ijk}]\!]$ for a few relation types (Figure 5). 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 estimation. We found that relations belonging to a same cluster tend to have similar covariate effects. For example, the relations *warnings* and *ecnomicaid* were classified into Cluster II, and both exhibit similar covariate pattern (Figure 5a-b). Moreover, the diagonal entries $\hat{\mathcal{B}}(i,i,k)$ tend to positively contribute to the connection. This is probably explained by the fact that countries with coherent attributes tend to interact more often than others. We also found that the *constitutional* attributes are greatly associated with the *commonbloc* relation, whereas such association is weaker for other relations (Figure 5d). This is not surprising, as the common block partition during Cold War is determined by capitalism vs. communism, which confounds with the *constitutional* attributes.

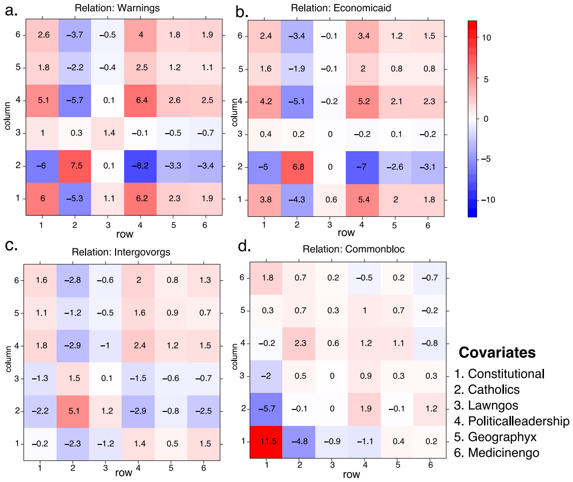

Figure 5: Estimated coefficient of country attributes towards the connection probability of relation type k.

8 Conclusions

References

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

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

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

[4] Muni S Srivastava, Tatjana Nahtman, and Dietrich Von Rosen. Estimation in general multivariate linear models with kronecker product covariance structure. Technical report.

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

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

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

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

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

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

[11] Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. Scalable tensor factorizations with missing data. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 701–712. SIAM, 2010.

[12] Miaoyan Wang and Lexin Li. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *arXiv preprint arXiv:1811.05076*, 2018.

[13] Miaoyan Wang and Yun Song. Tensor decompositions via two-mode higher-order SVD (HOSVD). In *Artificial Intelligence and Statistics*, pages 614–622, 2017.

[14] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[15] Navid Ghadermarzy, Yaniv Plan, and Ozgur Yilmaz. Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67(1):29–40, 2018.

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

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

[18] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

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

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

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