Generalized tensor-response model with multi-sided covariates

Abstract

We consider the problem of learning higher-order tensors with side information on a set of modes. Such data problems arise frequently arise in applications such as neuroimaging, network analysis, and ... 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 applications, including modeling brain connection in populations, link prediction in networks, and spatial-temporal growth model. The efficacy of our method is demonstrated through both simulations and analyses of two real-word datasets.

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. ... see demonstration plot.

In this article, we provide a general treatment to these seemingly different problems.

Comparison with other models

Our model is related to, but fundamentally different from, several lines of existing work.

Unsupervised tensor. supervised learning.

Tensor-predictor regression multilinear in coefficients.

Tensor-response regression multilinear in coefficient vs. multilinear in covariates.

Generalized linear model. In the high-dimensions both p and n increase while $p \leq d$. This is the case we consider. Classical GLM fixes p. Compared to Gaussian model, the log-likelihood is not strictly convex in the linear predictor. We allow various types of dependent variable.

2 Preliminaries

We begin by reviewing a few basic factors 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 $\boldsymbol{M}_k = [\![m^{(k)}_{i_k,j_k}]\!] \in \mathbb{R}^{s_k \times d_k}$ is defined as

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

which results in an order-K tensor $(s_1, \ldots, s_K)$-dimensional tensor. For ease of notatio, we also write the above Tucker product $\mathcal{Y} \times \{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K\}$ for short. 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 $(a, b, c, \ldots)$ for scalars and vectors, upper-case boldface letters $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots)$ for matrices, and calligraphy letter $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots)$ for tensors of order 3 or greater.

3 Motivation and models

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. In addition, suppose we observe covariates on a subset of modes. Let $\boldsymbol{X}_k \in \mathbb{R}^{d_k\times p_k}$ be the available covariates on the mode-k, where $p_k \le d_k$. We propose the following multilinear structure in the mean of the tensor. Specifically,

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

where $f(\cdot)$ is a known link function, $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is called the linear predictor tensor, $\mathcal{B} \in \mathbb{R}^{p_1\times\cdots p_K}$ is the parameter tensor of interest, and $\times$ denotes the tensor Tucker product. The link function depends on the distribution family of the response. Some common choices are identity link for Gaussian tensor, logistic link for binary tensor, and log link for Poisson tensor. We give three examples of multi-covariates tensor regression model that arises in practice.

[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, a classical 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, $\boldsymbol{X}_1 = \text{blockdiag}\{\mathbf{1}_p,\ldots,\mathbf{1}_p\} \in \{0,1\}^{d\times p}$ is the design matrix for lake types,

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

are the design matrices for spatial and temporal effects, respectively. [Network population model] Network response model is recently developed in the context of neuroimanig analysis. The goal is to study the relationship between the network-valued response with 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 subject covariate such as age, gender. The network-response model has the form

$$\text{logit}(\mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{x}_i)) = \mathcal{B}\times_3\boldsymbol{x}_i, \quad \text{for } i = 1,\ldots,n \tag{2}$$

where $\mathcal{B} \in \mathbb{R}^{d\times d\times p}$ is the coefficient tensor of interest. In fact, the model (2) is a special case of our multilinear tensor-response model. To see this, let $\mathcal{Y} \in \{0,1\}^{d\times d\times n}$ denote the response tensor by stacking $\{\boldsymbol{Y}_i\}$ together along the 3^{rd} mode and $\boldsymbol{X} = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n] \in \mathbb{R}^{p\times n}$, then model (2) can be expressed as

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

where $\boldsymbol{I}_d$ denotes the identity matrix of dimension d.

[Link model with node attributes] Let $V = [n]$ be a set of vertices and explanatory variable $x_i \in \mathbb{R}^p$ associated to each $i \in V$. The network $G = (V,E)$ is described by the following matrix

model. The edge connects the two vertices i and j independently of the others. The probability of connection is modeled as

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

Again, we show that this model is a special case of our tensor regression model. 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 above model can be expressed as

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

In the above three example and many other studies, researchers are interested in uncovering the variation in the data tensor that are explained by the covariates.

Without any structure on the coefficient tensor $\mathcal{B}$: *A naive approach is to regress the tensor entry, one at a time, on the covariates, and this model is repeatedly fitted for each tensor element.* Though this approach is scalable, it suffers from two drawbacks: (1) ignore the multilinear structure in the tensor (?) and (2) suffers from the multiplicity issue. To allow the structure among ..we further impose a multilinear low-rank structure on the coefficient tensor $\mathcal{B}$

$$\mathcal{P} = \{\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K} : r_k(\mathcal{B}) \le r_k \text{ for } k \in [K]\}, \tag{3}$$

where $r_k(\mathcal{B})$ is the Tucker rank of the tensor at mode k. We assume that r_k is known and $r_k \le p_k$. Other low-rankness such as CP rank is also possible. We choose the Tucker decomposition due to the following observation.

The low-rank structure in (3) implies that the coefficient tensor can be expressed as $\mathcal{B} = \mathcal{C} \times_1 \boldsymbol{M}_1 \times \cdots \boldsymbol{M}_K$. Then, our tensor regression model (4) is equivalent to

$$f(\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)) = \mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1, \ldots, \boldsymbol{X}_k\boldsymbol{M}_k\}.$$

The goal is to find a joint dimension reduction of $\mathcal{Y}$ and $\boldsymbol{X}_K$ such that the unexplained variation in the mean tensor. The factorization is restricted to the space spanned by $\boldsymbol{X}_k$. Here $\boldsymbol{X}_1\boldsymbol{M}_1$ can be interpreted as the latent covariates that explains the variation in the response tensor. The core tensor $\mathcal{C}$ collects the interaction effects of latent covariates across the K modes.

(Question: the columns of $\boldsymbol{X}$ should be normalized??)

Question: any connection to tensor completion?? If $\boldsymbol{X}_K$ is a random design matrix?

4 Rank-constrained likelihood-based estimation

Note that 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 $\boldsymbol{X}_k = \boldsymbol{I}_{d_k}$ when the mode-k has no (informative) covariate. Our general tensor regression model can be written as

$$\begin{aligned} &\mathbb{E}(\mathcal{Y}|\mathcal{X}) = f(\Theta), \quad \Theta = \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}, \\ &\text{where rank}(\mathcal{B}) = (r_1, \ldots, r_K), \end{aligned} \tag{4}$$

where $\mathcal{B} \in \mathbb{R}^{p_1\times\cdots\times p_K}$ is a low-rank coefficient tensor of interest. In the following theoretical analysis, we assume the Tucker rank $\boldsymbol{r} = (r_1, \ldots, r_K)$ is known. The adaptation of unknown $\boldsymbol{r}$ will be addressed in Section 5.2.

We develop a likelihood-based procedure to estimate $\mathcal{B}$. 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

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

where $c(\cdot)$ and $b(\cdot)$ are known functions and θ is the linear predictor. Note that the canonical link function f is chosen to be $f(\cdot) = b'(\cdot)$. Table 1 summarizes the canonical link function for common distributions.

In our context, the log-likelihood of (4) is the (?) divergence between the conditional distribution of $\mathcal{Y}|\Theta$ and the exponential family

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

Assume that we have an additional information on an upper bound $a > 0$ such that $\|\Theta\|_\infty \le \alpha$. (more comments? ill-condition in the bernoulli model?) We propose the following constrained maximum likelihood estimation for the tensor coefficient

$$\hat{\mathcal{B}} = \underset{\text{rank}(\mathcal{B})=\boldsymbol{r}, \|\Theta(\mathcal{B})\|_\infty\le\alpha}{\arg\max} \mathcal{L}_{\mathcal{Y}}(\mathcal{B}). \tag{5}$$

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

$$\text{Loss}(\mathcal{B}_{\text{true}}, \hat{\mathcal{B}}) = \frac{1}{\prod_k p_k}\|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}\|_F^2.$$

We focus on the high-dimensional region in which both $d_k \to \infty$ and $p_k \to \infty$ while $\frac{p_k}{d_k} \to \gamma_k \in [0,1]$. We make the following assumptions:

1. There exists two positive constants $c_1,\ c_2 > 0$ such that $c_1 \le \sigma_{\min}(\boldsymbol{X}_k) \le \sigma_{\max}(\boldsymbol{X}_k) \le c_2$ for all $k \in [K]$.

2. There exist two positive constants $L,\ U > 0$ such that $L \le \text{Var}(y_{i_1,\ldots,i_K}|\mathcal{X}) \le U$ uniformly over the parameter space $\mathcal{P}$.

2'. Equivalently, $L \le b''(x) \le U$ for all $x \le \alpha$, where $b(\cdot)$ is the known function in the exponential family distribution and α is the upper bound of the linear predictor.

[Convergence] Consider a generalized tensor regression model with multi-sided covariates. Let $\mathcal{Y} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ be the tensor response and $\mathcal{X} = \{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K\}$ the covariates, where $\boldsymbol{X}_k \in \mathbb{R}^{p_k\times d_k}$ is the covariate matrix on mode-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 (4). Under Assumption 4.1, there exist two constants $C_1, C_2 > 0$ such that, with probability at least $1 - \exp(-C_1\sum_k p_k)$,

$$\text{Loss}(\mathcal{B}_{\text{true}}, \hat{\mathcal{B}}) \le \frac{2}{c_1^{2K}}\min\left\{\frac{C(\boldsymbol{r},\alpha)\sum_k p_k}{\prod_k p_k},\ 2\alpha^2\right\},$$

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 α

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 $\mathrm{vec}(\mathcal{Y})$ 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

where $C(\alpha, \boldsymbol{r}) = \frac{1}{b''(\alpha)} \frac{\sum_k r_k}{r_{\max}} > 0$ is a constant that does not depend on dimension $\{d_k\}$ and $\{p_k\}$.

To gain further insight on the bound we consider the special case when dimensions are equal at each of the modes. 1. binary case; 2. large dimension region. $\mathcal{O}\left(\frac{p}{d^k}\right) \leq \mathcal{O}(d^{-(k-1)})$.

Corollary 1 (Spatio-temporal growth model)**.** *Our method yields the convergence rate* $\mathcal{O}\left(\frac{p+r+s}{dmn}\right)$. *Note that* $p \leq d$, $r \leq m$ *and* $s \leq m$, *so consistent estimator.*

Corollary 2 (Network population model)**.** *Our method yields the convergence rate* $\mathcal{O}\left(\frac{2d+p}{d^2 n}\right)$. *Note that* $p \leq m$, *so this is a consistent estimator. In contrast, a naive repeated glm will give* $\mathcal{O}(\frac{p}{n})$.

Corollary 3 (Link model with node attributes)**.** *Our method yields the convergence rate is* $\mathcal{O}\left(\frac{p}{d^2}\right)$. *Note that* $p \leq m$, *so again a consistent estimator. In contrast, a naive repeated glm will give* $\mathcal{O}(\frac{p}{n})$.

We provide the prediction accuracy for the response tensor. [Prediction error] Assume the same set-up as in Theorem (4.1). Let $\mathbb{P}_{\mathcal{Y}_{\text{true}}}$ the distribution of $\mathcal{Y}$ given the true $\mathcal{B}_{\text{true}}$ and $\mathbb{E}(\mathcal{Y}|\mathcal{X})$ the true mean. Let $\mathbb{P}_{\hat{\mathcal{Y}}}$ denote the distribution given the estimated $\hat{\mathcal{B}}$ and $\widehat{\mathbb{E}(\mathcal{Y}|\mathcal{X})}$ the predicted mean. We have, with probability at least $1 - \exp(C_1 \sum_k d_k)$,

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

$$\mathrm{Loss}\left(\mathbb{E}(\mathcal{Y}|\mathcal{X}),\ \widehat{\mathbb{E}(\mathcal{Y}|\mathcal{X})}\right) \leq b''(\alpha) C(...) \frac{\sum_k p_k}{\prod_k p_k}.$$

4.1.1 Comparison between Gaussian and non-Gaussian models

Our earlier result gives the general bound with a constant factor C. The factor depends on the tensor order, rank, the distribution family, and the infinity norm bound. We now discuss its

specific form for various distribution types of the response. We consider the special case when $d_1 = \ldots = d_K = d$, $p_1 = \ldots = p_K = p$, $r_1 = \ldots = r_K = r$. The constant $C = \frac{2}{b''(\alpha)} \prod_k \sigma_{\min}^{-1}(\boldsymbol{X}_k)$. For Bernoulli model with logistic link $b''(\alpha) = \frac{e^\alpha}{(1+e^\alpha)^2}$. The bound

$$\text{Loss}(\mathcal{B}_{\text{true}}, \hat{B}) \leq \frac{2}{c_1^2 K} \min \left\{ \frac{r^{K-1} e^\alpha}{(1+e^\alpha)^2} \frac{p}{d^K},\ 2\alpha^2 \right\}.$$

The α can be interpreted as the signal level. As $p \to \infty, d \to \infty, \frac{p}{d} \to \gamma \in [0, 1]$. For consistent estimation, we require $\mathcal{O}(d^{-(K-1)}) \ll \alpha \ll \mathcal{O}((K-1)\log d)$.

Type	$\alpha \lesssim d^{-(K-1)/2}$	$d^{-(K-1)/2} \ll \alpha \ll \mathcal{O}((K-1)\log d)$	$\alpha \gg (K-1)\log d$
Bernoulli	$2\alpha^2$		$e^{\alpha-(K-1)\log d}$

5 Numerical implementation

5.1 Alternating optimization

In this section, we introduce an efficient algorithm to solve (5). The objective function $\mathcal{L}(\mathcal{B})$ is concave in $\mathcal{B}$ when the link f is canonical link function. However, the feasible set $\mathcal{P}$ is nonconvex, and thus the optimization (5) is a non-convex problem. We utilize a factor representation of the Tucker decomposition, 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 whose columns are orthogonal. Estimating $\mathcal{B}$ amounts to finding both the core tensor $\mathcal{C}$ and the factor matrices $\boldsymbol{M}_k$'s. The optimization (5) 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

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

The decision variables 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 suggest 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 $k = 1, \ldots, K-1$. The last factor matrix $\boldsymbol{M}_K$ can be solved row-by-row via d_k separate GLMs. To see this, let $\boldsymbol{M}_k^{(t)}$ denote the kth factor matrix at the tth iteration, and

$$\boldsymbol{M}_{-K}^{(t)} = [\boldsymbol{M}_1^{(t)} \boldsymbol{X}_1] \odot [\boldsymbol{M}_2^{(t)} \boldsymbol{X}_2] \odot \cdots \odot [\boldsymbol{M}_{K-1}^{(t)} \boldsymbol{X}_{K-1}],$$

where $\odot$ denotes the Khatri-Rao product of matrices. Then, the j-th row of $\boldsymbol{A}_K$ is the "regression coefficient" for a GLM with $\mathrm{vec}(\mathcal{Y}(:, \ldots, j)) \in \mathbb{R}^{d_{-K}}$ as the "response", $\boldsymbol{X}_{-K} \in \mathbb{R}^{d_{-K} \times r_K}$ as the "predictors", where $\mathrm{vec}(\cdot)$ is the operator that reshapes a tensor into a vector, $d_{-K} \stackrel{\text{def}}{=} \prod_{k \in [K-1]} d_k$, and $j = 1, \ldots, d_K$. The separability by row allows us to leverage state-of-art GLM solvers and parallel processing to speed up the computation. After each iteration, we post-process the core tensor $\mathcal{C}^{(t+1)}$ via scaling, subject to the maximum constraints. The last step may not guarantee the monotonic increase of the objective, but we find in our experiment that the simple post-processing appears to be good enough for a desirable solution. The full algorithm is described in Algorithm 1.

5.2 Missing data, rank selection

When some tensor entries $y_{i_1, \ldots, i_K}$ are missing, we replace the objective function $\mathcal{L}_{\mathcal{Y}} = \sum_{(i_1, \ldots, i_K) \in} (y_{i_1, \ldots, i_K} \theta_{i_1, \ldots, i_K} -$ 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 only and exclude the missing entries in the model fitting. The same strategy has been used for classical Tucker and CP tensor decomposition. In the presence of missing data, we modify line 7 in Algorithm 1 by fitting GLMs to the data for which $y_{i_1, \ldots, i_K}$ are observed. This approach requires no completely missing sub-tensors, for example, $\mathcal{Y}(:, \ldots, K)$. We regard this as a fairly mild condition akin to the coherence condition in matrix completion literature; for instance, if an entire row or a column of a tensor is missing, the recovery of the true decomposition becomes impossible.

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 $p_e(\boldsymbol{r}) \overset{\text{def}}{=} \sum_k (d_k - 1)r_k + \prod_k r_k$ 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 6.

6 Simulation

We evaluate the empirical performance of our generalized tensor regression through simulations. We consider order-3 tensors with a range of data types. The linear predictors are 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 $N(0, d_k^{-1/2})$. The choice of $d_k^{-1/2}$ is to guarantee the boundedness of singular values of $\boldsymbol{X}_k$ as d_k increases (Assumption 4.1). Without loss of generality, we scale the linear predictor 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 tensor $y_{ijk} \sim N(\alpha u_{ijk}, 1)$.

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

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

Here $\alpha > 0$ is a scalar controlling the infinity norm 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 studies the selection accuracy. For the purpose of illustration, we consider the balanced situation where $d_i = d$, $p_i = 0.4d_i$ for $i = 1, 2, 3$. We set $\alpha = 5$ and consider various combinations of dimension d and rank r.

True rank	$d = 30$	$d = 60$	$d = 30$	$d = 60$
$\boldsymbol{r} = (2, 2, 4)$	(2,2,3)	(2,2,3.1)		
$\boldsymbol{r} = (4, 4, 6)$	(4,4,5)	(4,4,6)		
$\boldsymbol{r} = (4, 6, 8)$				

The second experiment evaluates the accuracy when covariates are available on all modes. 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 from Figure 2b-c.

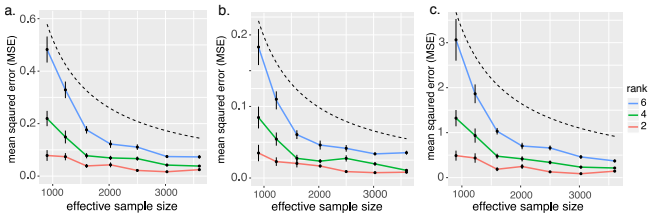


Figure 2: Estimation error against effective sample size. The three panels depicts the MSE when the response tensor is generated form (a) Gaussian (b) Poisson and (c) Bernoulli models. Each solid curve corresponds to a fixed rank. The dashed curve corresponds to $\mathcal{O}(1/d^2)$.

The third experiment investigate our model's ability in handling structured 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 fiber connection between the $d_1 = d_2 = 20$ brain nodes. There are $p = 5$ covariates for the 50 individuals. These covariates may represent, for example, age, gender, cognitive score, etc. Motivated by functional division of brains, we assume the 20 nodes can be divided r regions, 5 for each region. For each covariate, the effects over the brain nodes follow block-structure, with the region effect drawn i.i.d. from $N(0, 1)$. This is to reflect the spatial correlation among nodes within a same region.

Figure 2 compares our model with a naive approach. A naive approach is to regress the tensor entry, one at a time, on the covariates, and this model is repeatedly fitted for each tensor element. This approach does not account for the spatial similarity among the nodes. In contrast, our model automatically identifies the node clustering and encourages the sharing within a same region. The low-rankness implicitly encourages the spatial similarity among nodes. We note that the fewer

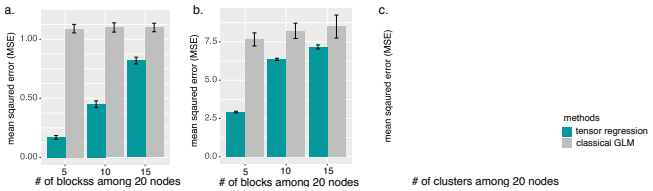


Figure 3: Performance comparison when the population network admit block structure. The three panels depicts the MSE when the response tensor is generated form (a) Gaussian (b) Poisson and (c) Bernoulli models. The x-axis represents the number of blocks in the networks.

The last experiment considers the correlation among multiple directions. Specifically, correlation within matrices, and correlation along the third modes. We simulate multi-relational network data.

7 Data analysis

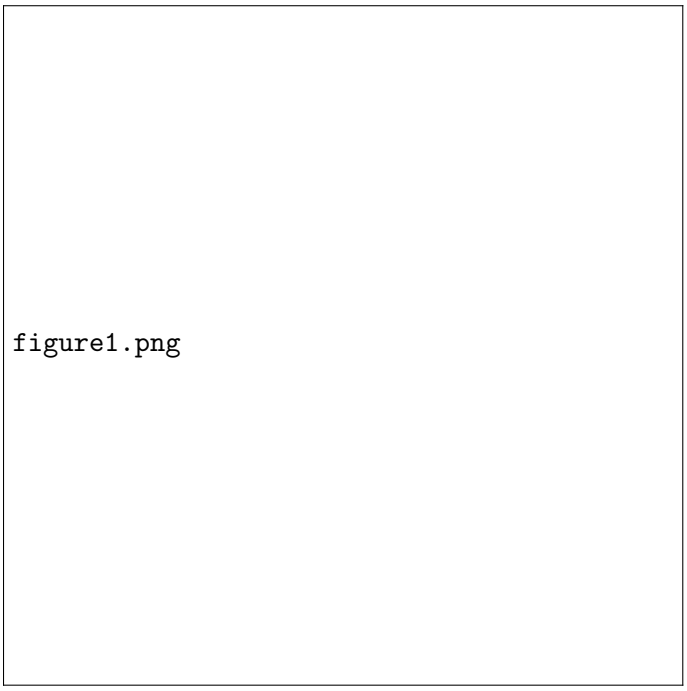
We apply our tensor regression model to the following two real datasets. The first dataset has covariates on two modes, and the second dataset has covariates on one mode.

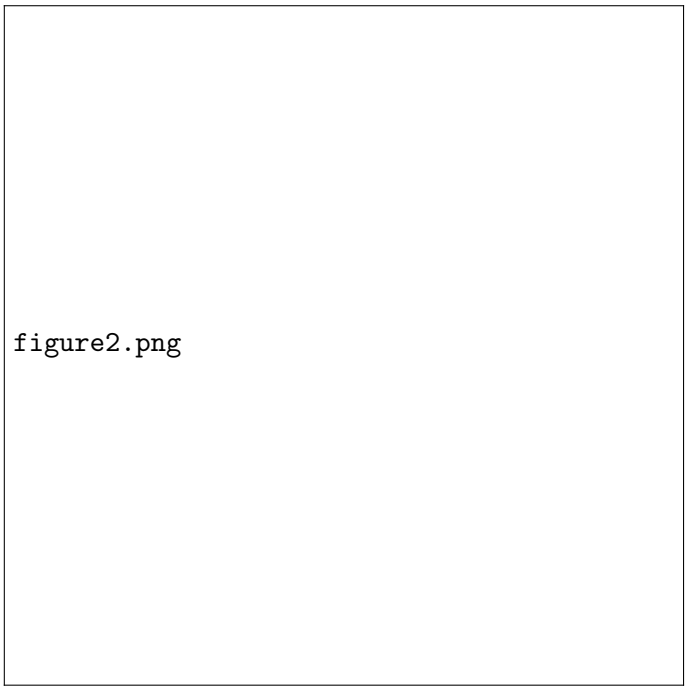
7.1 Human Connectome Project

The HCP data consists of 136 networks, each for one individual. Each network is of dimension 68-by-68. The network edges encode the presence or absence of fiber connections between 68 brain regions for each of the 138 individuals. The individual covariates, such as age and gender, are also

available. The goal is to identify the connections that are affected by the individual covariates. Presumably, the nodes are correlated with each other, and we aim to take that into account in assessing the covariate effect.

We fit the tensor regression model to the HCP data. The response is a $60 \times 68 \times 212$ tensor and the covariates is of dimension 5 along the third mode. The BIC selection suggests the rank $(10, 10, 4)$, and the corresponding loglikelihood -174654.7. We find that generally, the connection within each hemisphere are more significant than connections cross two hemispheres. Table





7.2 Nations data

The second data.. selected rank is $(3, 3, 5)$. The 56 relationships can be classified into 5 categories. We first identifies the key attribute pairs that affect the relationship.

Attribute	Attribute	Relationship	Coefficient
Political leadership	Constitution	NGO_Orgs 3	-28

8 SUPPLEMENTARY MATERIAL

If you need to include additional appendices during submission, you can include them in the supplementary material file.

9 INSTRUCTIONS FOR CAMERA-READY PAPERS

For the camera-ready paper, if you are using LaTeX, please make sure that you follow these instructions. (If you are not using LaTeX, please make sure to achieve the same effect using your chosen typesetting package.)

1. Download `fancyhdr.sty` – the `aistats2020.sty` file will make use of it.

2. Begin your document with

   ```
   \documentclass[twoside]{article}
   \usepackage[accepted]{aistats2020}
   ```

   The `twoside` option for the class article allows the package `fancyhdr.sty` to include headings for even and odd numbered pages. The option `accepted` for the package `aistats2020.sty` will write a copyright notice at the end of the first column of the first page. This option will also print headings for the paper. For the *even* pages, the title of the paper will be used as heading and for *odd* pages the author names will be used as heading. If the title of the paper is too long or the number of authors is too large, the style will print a warning message as heading. If this happens additional commands can be used to place as headings shorter versions of the title and the author names. This is explained in the next point.

3. If you get warning messages as described above, then immediately after `\begin{document}`, write

   ```
   \runningtitle{Provide here an alternative shorter version of the title of
   your paper}
   \runningauthor{Provide here the surnames of the authors of your paper, all
   separated by commas}
   ```

   Note that the text that appears as argument in `\runningtitle` will be printed as a heading in the *even* pages. The text that appears as argument in `\runningauthor` will be printed as a heading in the *odd* pages. If even the author surnames do not fit, it is acceptable to give a subset of author names followed by "et al."

4. Use the file sample_paper.tex as an example.

5. The camera-ready versions of the accepted papers are 8 pages, plus any additional pages needed for references.

6. If you need to include additional appendices, you can include them in the supplementary material file.

7. Please, don't change the layout given by the above instructions and by the style file.

Acknowledgements

Use the unnumbered third level heading for the acknowledgements. All acknowledgements go at the end of the paper.

References

References follow the acknowledgements. Use an unnumbered third level heading for the references section. Any choice of citation style is acceptable as long as you are consistent. Please use the same font size for references as for the body of the paper—remember that references do not count against your page length total.

References

J. Alspector, B. Gupta, and R. B. Allen (1989). Performance of a stochastic learning microchip. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 1*, 748–760. San Mateo, Calif.: Morgan Kaufmann.

F. Rosenblatt (1962). *Principles of Neurodynamics.* Washington, D.C.: Spartan Books.

G. Tesauro (1989). Neurogammon wins computer Olympiad. *Neural Computation* **1**(3):321–323.