

Supervised Tensor Decomposition with Interactive Side Information

Abstract

Higher-order tensors have received increased attention across science and engineering. While most tensor decomposition methods are developed for a single tensor observation, scientific studies often collect side information, in the form of node features and interactions thereof, together with the tensor data. Such data problems are common in neuroimaging, network analysis, and spatial-temporal modeling. Identifying the relationship between a high-dimensional tensor and side information is important yet challenging. Here, we develop a tensor decomposition method that incorporates multiple side information as interactive features. Unlike unsupervised tensor decomposition, our supervised decomposition captures the effective dimension reduction of the data tensor confined to feature space on each mode. An efficient alternating optimization algorithm is further developed. Our proposal handles a broad range of data types, including continuous, count, and binary observations. We apply the method to diffusion tensor imaging data from human connectome project and multi-relational political network data. We identify the key global connectivity pattern and pinpoint the local regions that are associated with available features. Our method will help the practitioners efficiently analyze tensor datasets in various areas.

Keywords: Applications and case studies, Tensor data analysis, Supervised dimension reduction, Exponential family distribution, Generalized multilinear model

1 Introduction

Multi-dimensional arrays, known as tensors, are often collected with side information on multiple modes in modern scientific and engineering studies. A popular example is in neuroimaging (Zhou et al., 2013). The brain connectivity networks are collected from a sample of individuals, accompanied by individual characteristics such as age, gender, and diseases status (see Figure 1a). Another example is in network analysis (Berthet and Baldwin, 2020; Hoff, 2005). A typical social network consists of nodes that represent people and edges that represent friendships. Side information such as people’s demographic information and friendship types are often available. In both examples, it is of keen scientific interest to identify the variation in the data tensor (e.g., brain connectivities, social community patterns) that is affected by available features. These seemingly different scenarios pose a common yet challenging problem for tensor data modeling.

In addition to the aforementioned challenges, many tensor datasets consist of non-Gaussian measurements. Examples include the political interaction dataset (Nickel et al., 2011) which measures action counts between countries under various relations, and the brain connectivity network dataset (Zhang et al., 2018) which is a collection of binary adjacency matrices. Classical tensor decomposition methods are based on minimizing the Frobenius norm of deviation, leading to suboptimal predictions for binary- or count-valued response variables. A number of supervised tensor methods have been proposed (Narita et al., 2012; Zhao et al., 2012; Yu and Liu, 2016; Lock and Li, 2018) to address the tensor regression problem in various forms, such as scalar-to-tensor regression and tensor-response regression. These methods often assume Gaussian distribution for the tensor entries, or impose random designs for the feature matrices, both of which are less suitable for applications of our interest. The gap between theory and practice means a great opportunity to modeling paradigms and better capture the complexity in tensor data.

We present a general model and associated method for decomposing a data tensor whose entries are from exponential family with interactive side information. We formulate the learning task as a structured regression problem, with tensor observation serving as the

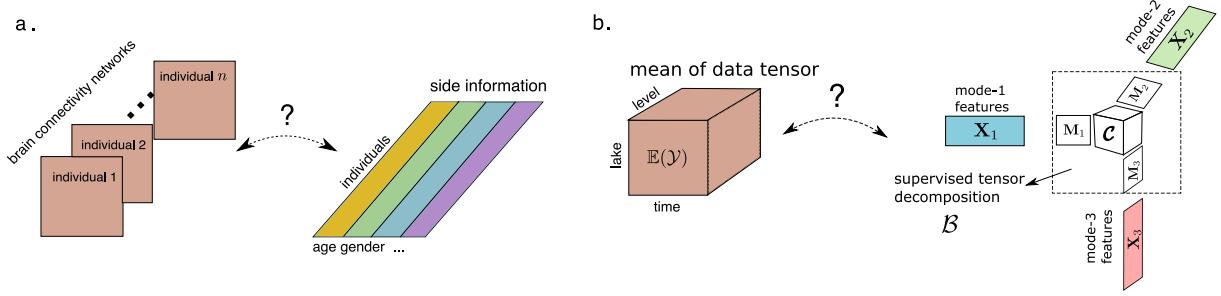


Figure 1: Examples of supervised tensor decomposition with interactive side information. (a) Network population model. (b) Spatio-temporal growth model.

response, and the multiple side information as interactive features. Figure 1b illustrates our model in the special case of order-3 tensors. A low-rank structure is imposed to the conditional mean of tensor observation, where unlike classical decomposition, the tensor factors $\mathbf{X}_k \mathbf{M}_k \in \mathbb{R}^{d_k \times r_k}$ belong to the space spanned by features $\mathbf{X}_k \in \mathbb{R}^{d_k \times p_k}$ for $k = 1, 2, 3$. The unknown matrices $\mathbf{M}_k \in \mathbb{R}^{p_k \times r_k}$ (referred to as “dimension reduction matrices”) link the conditional mean to the feature spaces, thereby allowing the identification of variations in the tensor data attributable to the side information.

Our proposal blends the modeling power of generalized linear model (GLM) and the exploratory capability of tensor dimension reduction in order to take the best out of both worlds. We leverage GLM to allow heteroscedacity due to the mean-variance relationship in the non-Gaussian data. This flexibility is important in practice. Furthermore, our low-rank model on the (transformed) conditional mean tensor effectively mitigates the curse of high dimensionality. In classical GLM, the sample size and feature dimension are well defined; however, in the tensor data analysis, we observe only one realization of an order- K tensor and up to K interactive feature matrices. Both the number of tensor entries and feature dimension grow exponentially in K . Dimension reduction is therefore crucial for prediction and interpretability. We establish the statistical convergence of our estimator, and we quantify the gain in prediction through simulations and case studies.

Our work is closely related to but also clearly distinctive from several lines of previous work. The first line is a class of *unsupervised* tensor decomposition such as Tucker and CP decomposition (De Lathauwer et al., 2000; Kolda and Bader, 2009; Hong et al., 2020) that

aims to find the best low-rank representation of a data tensor. In contrast, our model is a *supervised* tensor learning, which aims to identify the association between a data tensor and multiple features. The low-rank factorization is determined jointly by the tensor data and feature matrices in our model.

The second line of work studies the tensor-on-tensor regression (Raskutti et al., 2019; Lock, 2018; Gahrooei et al., 2020). Our model shares a common ground with earlier approaches, but we provide more efficient solutions to new settings that have more practical significance. Previous methods (Lock, 2018) mainly focus on Gaussian tensors. The Frobenius norm used in the objective function is statistical suboptimal for general exponential family tensors. Maximum likelihood estimator (MLE) is studied in (Raskutti et al., 2019) and a convex relaxation algorithm is proposed to solve for low-rank tensor coefficients. However, in the tensor case, convex MLE suffers from both computational intractability and statistical suboptimality. We advocate a non-convex approach and provide strong evidence for its success in our setting. Most previous tensor regression focuses on prediction (Lock, 2018; Raskutti et al., 2019), and we go step further by finding the sufficient dimension reduction (Adragni and Cook, 2009), $\text{Span}(\mathbf{M}_k)$, that facilitates the identification of interaction effects in features (see Figure 1b). The latter approach greatly improves the *interpretability* in prediction. In this regards, our method opens up new opportunities for tensor data analysis in a wider range of applications.

The third line of work uses side information for various tensor learning tasks, such as for completion (Narita et al., 2012) and for recommendation system (Ioannidis et al., 2019; Farias and Li, 2019). These methods also study tensors with side information, but they take regularization approaches to penalize predictions that are distant from side information Lock and Li (2018). One important difference is that their goal is prediction but not parameter estimation. The effects of features and their interactions are not estimated in these data-driven approaches. In contrast, our goal is interpretable prediction, and we estimate factor matrices \mathbf{M}_k using a model-based approach. Estimating \mathbf{M}_k allows us to identify sufficient features and the interactions thereof.

2 Preliminaries

We introduce the basic tensor properties (Kolda and Bader, 2009) used in the paper. We use lower-case letters (e.g., a, b, c) for scalars and vectors, upper-case boldface letters (e.g., $\mathbf{A}, \mathbf{B}, \mathbf{C}$) for matrices, and calligraphy letters (e.g., $\mathcal{A}, \mathcal{B}, \mathcal{C}$) for tensors of order three or greater. Let $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!] \in \mathbb{R}^{d_1 \times \dots \times d_K}$ denote an order- K (d_1, \dots, d_K)-dimensional tensor. The multilinear multiplication of a tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ by matrices $\mathbf{X}_k = [\![x_{i_k, j_k}^{(k)}]\!] \in \mathbb{R}^{p_k \times d_k}$ is defined as $\mathcal{Y} \times_1 \mathbf{X}_1 \times \dots \times_K \mathbf{X}_K = [\![\sum_{i_1, \dots, i_K} y_{i_1, \dots, i_K} x_{j_1, i_1}^{(1)} \dots x_{j_K, i_K}^{(K)}]\!]$, which results in an order- K (p_1, \dots, p_K)-dimensional tensor. For ease of presentation, we use the shorthand $\mathcal{Y} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\}$ to denote the tensor-by-matrix product. For any two tensors $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!]$, $\mathcal{Y}' = [\![y'_{i_1, \dots, i_K}]\!]$ of identical order and dimensions, their inner product is defined as $\langle \mathcal{Y}, \mathcal{Y}' \rangle = \sum_{i_1, \dots, i_K} y_{i_1, \dots, i_K} y'_{i_1, \dots, i_K}$. The tensor Frobenius norm is defined as $\|\mathcal{Y}\|_F = \langle \mathcal{Y}, \mathcal{Y} \rangle^{1/2}$, and the maximum norm is defined $\|\mathcal{Y}\|_\infty = \max_{i_1, \dots, i_K} y_{i_1, \dots, i_K}$. When a is a vector, we use $\|a\|_2 = \langle a, a \rangle^{1/2}$ to denote the vector 2-norm. A higher-order tensor can be reshaped into a lower-order object. We use $\text{vec}(\cdot)$ to denote the operation that reshapes the tensor into a vector, and $\text{Unfold}_k(\cdot)$ to denote the operation that reshapes the tensor along mode k into a matrix of size d_k -by- $\prod_{i \neq k} d_i$. The multilinear rank of an order- K tensor \mathcal{Y} is defined as a length- K vector $\mathbf{r} = (r_1, \dots, r_K)$, where r_k is the rank of matrix $\text{Unfold}_k(\mathcal{Y})$, $k = 1, \dots, K$. We use $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the smallest and largest matrix singular values. We let \mathbf{I}_d denote the $d \times d$ identity matrix and $[d]$ denote the d -set $\{1, \dots, d\}$. For ease of notation, we allow the basic arithmetic operators (e.g., $+, -, \geq$) and univariate functions $f: \mathbb{R} \rightarrow \mathbb{R}$ to be applied to tensors in an element-wise manner.

3 Motivation and model

3.1 General framework for supervised tensor decomposition

We begin with a general framework for supervised tensor decomposition and then discuss its implication in three concrete examples. Let $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!] \in \mathbb{R}^{d_1 \times \dots \times d_K}$ denote an order-

K data tensor. Suppose the side information is available on each of the K modes. Let $\mathbf{X}_k = \llbracket x_{ij} \rrbracket \in \mathbb{R}^{d_k \times p_k}$ denote the feature matrix on the mode $k \in [K]$, where x_{ij} denotes the j -th feature value for the i -th tensor entity, for $(i, j) \in [d_k] \times [p_k]$, $p_k \leq d_k$. We propose a multilinear conditional mean model between the data tensor and feature matrices. Assume that, conditional on the features \mathbf{X}_k , the entries of tensor \mathcal{Y} are independent realizations from an exponential family distribution, and the conditional mean tensor admits the form

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

where $\Theta \in \mathbb{R}^{d_1 \times \dots \times d_K}$ is the multilinear predictor, $\mathcal{B} \in \mathbb{R}^{p_1 \times \dots \times p_K}$ is the unknown parameter tensor, $f(\cdot)$ is a known link function whose form depending on the data type of \mathcal{Y} , and \times denotes the tensor-by-matrix product. The choice of link function is based on the assumed distribution family of tensor entries. Common choices of link functions include identity link for Gaussian distribution, logistic link for Bernoulli distribution, and $\exp(\cdot)$ link for Poisson distribution. In general, dispersion parameters can also be included in the model. Because our main focus is the tensor decomposition under the mean model, we omit the dispersion parameter in this section for ease of presentation.

In classical tensor decomposition, tensor factorization is performed on either data tensor \mathcal{Y} or mean tensor $\mathbb{E}(\mathcal{Y})$. In the context of supervised tensor decomposition, we propose to factorize the latent parameter tensor \mathcal{B} ,

$$\mathcal{B} = \mathcal{C} \times \{\mathbf{M}_1, \dots, \mathbf{M}_K\}, \quad (2)$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times \dots \times r_K}$ is a full-rank core tensor, and $\mathbf{M}_k \in \mathbb{R}^{p_k \times r_k}$ are factor matrices consisting of orthonormal columns, where $r_k \leq p_k$ for all $k \in [K]$. By the definition of multilinear rank, model equations (1) and (2) imply the low-rankness $\mathbf{r} = (r_1, \dots, r_K)$ of the conditional mean tensor under the link function. We now reach our final model for

supervised tensor decomposition,

$$\begin{aligned} \mathbb{E}(\mathcal{Y}|\mathbf{X}_1, \dots, \mathbf{X}_K) &= f(\mathcal{C} \times \{\mathbf{X}_1\mathbf{M}_1, \dots, \mathbf{X}_K\mathbf{M}_K\}), \\ \text{with } \mathbf{M}_k^T \mathbf{M}_k &= \mathbf{I}_{r_k}, \quad \mathbf{M}_k \in \mathbb{R}^{p_k \times r_k} \quad \text{for all } k = 1, \dots, K, \end{aligned} \quad (3)$$

where the parameters of interest are \mathbf{M}_k and \mathcal{C} . Note that model (3) assumes a fixed, known rank $\mathbf{r} = (r_1, \dots, r_K)$; the adaptation to unknown rank will be addressed in Section 4.2. Figure 1b provides a schematic illustration of our model. The features \mathbf{X}_k affect the distribution of tensor entries in \mathcal{Y} through the form $\mathbf{X}_k\mathbf{M}_k$, which are r_k linear combinations of features on mode k . We call $\mathbf{X}_k\mathbf{M}_k$ the “supervised tensor factors” or “sufficient features” (Adragni and Cook, 2009), and call \mathbf{M}_k the “dimension reduction matrix.” The core tensor \mathcal{C} collects the interaction effects between sufficient features across K modes. Our goal is to find \mathbf{M}_k and the corresponding \mathcal{C} . Note that \mathbf{M}_k and \mathcal{C} are identifiable only up to orthonormal transformations.

3.2 Two examples

We give two seemingly different examples that can be formulated as our supervised tensor decomposition model (3).

Example 1 (Spatio-temporal growth model). The growth curve model (Gabriel, 1998) was originally proposed as an example of bilinear model for matrix data, and we adopt its higher-order extension here. 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 at most r and that the expected trend in time is a polynomial of order s . Then, the conditional mean model for the spatio-temporal growth can be represented as

$$\mathbb{E}(\mathcal{Y}|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = \mathcal{C} \times \{\mathbf{X}_1\mathbf{M}_1, \mathbf{X}_2\mathbf{M}_2, \mathbf{X}_3\mathbf{M}_3\}, \quad (4)$$

where $\mathbf{X}_1 = \text{blockdiag}\{\mathbf{1}_p, \dots, \mathbf{1}_p\} \in \{0, 1\}^{d \times q}$ is the design matrix for lake types, and

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

are the design matrices for spatial and temporal effects, respectively, $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the unknown core tensor, and \mathbf{M}_k are unknown dimension reduction matrices on each mode. The factors $\mathbf{X}_k \mathbf{M}_k$ are sufficient features in the mean model (4). The spatial-temporal model is a special case of our supervised tensor decomposition model (3), with features available on each of the three modes.

Example 2 (Network population model). Network response model (Rabusseau and Kadri, 2016) is recently developed for neuroimaging analysis. The goal is to study the relationship between brain network connectivity pattern and features of individuals. Suppose we have a sample of n observations, $\{(\mathbf{Y}_i, \mathbf{x}_i) : i = 1, \dots, n\}$, where for each individual $i \in [n]$, $\mathbf{Y}_i \in \{0, 1\}^{d \times d}$ is the undirected adjacency matrix whose entries indicate presences/absences of connectivities between d brain nodes, and $\mathbf{x}_i \in \mathbb{R}^p$ is the individual's feature such as age, gender, cognition score, etc. The network-response model has the conditional mean

$$\text{logit}(\mathbb{E}(\mathbf{Y}_i | \mathbf{x}_i)) = \mathcal{B} \times_3 \mathbf{x}_i, \quad \text{for } i = 1, \dots, n, \quad (5)$$

where $\mathcal{B} \in \mathbb{R}^{d \times d \times p}$ is a rank- (r_1, r_1, r_2) coefficient tensor, and \mathcal{B} is assumed to be symmetric in the first two modes.

The model (5) is a special case of our supervised tensor decomposition, with feature matrix on the last mode of the tensor. Specifically, we stack the network observations $\{\mathbf{Y}_i\}$ together and obtain an order-3 response tensor $\mathcal{Y} \in \{0, 1\}^{d \times d \times n}$. Define a feature matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$. Then, the model (5) has the equivalent representation

of supervised tensor decomposition,

$$\text{logit}(\mathbb{E}(\mathcal{Y}|\mathbf{X})) = \mathcal{C} \times \{\mathbf{M}, \mathbf{M}, \mathbf{X}\mathbf{M}'\},$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times r_1 \times r_2}$ is the core tensor, $\mathbf{M} \in \mathbb{R}^{d \times r}$ is the dimension reduction matrix on the first two modes, and $\mathbf{M}' \in \mathbb{R}^{p \times r_2}$ is for the last mode.

In the above two examples and many other studies, researchers are interested in uncovering the variation in the data tensor that can be explained by features. Our supervised tensor decomposition (3) allows arbitrary numbers of feature matrices. When certain mode k has no side information, we set $\mathbf{X}_k = \mathbf{I}_{d_k}$ in the model (3). In particular, our model (3) reduces to classical unsupervised tensor decomposition (De Lathauwer et al., 2000; Hong et al., 2020) when no side information is available; i.e., $\mathbf{X}_k = \mathbf{I}_{d_k}$ for all $k \in [K]$.

4 Estimation

4.1 Rank-constrained M-estimator

We develop a likelihood-based procedure to estimate \mathcal{C} and \mathbf{M}_k in (3). We adopt the exponential family as a flexible framework for different data types. In a classical generalized linear model with a scalar response y and feature \mathbf{x} , the density is expressed as

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

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. The canonical link function f is chosen to be $f(\cdot) = b'(\cdot)$, the first-order derivative of $b(\cdot)$. 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 links for common distributions.

In our context, we model the tensor entries y_{i_1, \dots, i_K} , conditional on θ_{i_1, \dots, i_K} , as independent draws from an exponential family. Ignoring constants that do not depend on Θ , the quasi log-likelihood of (3) is equal to Bregman distance between \mathcal{Y} and $b'(\Theta)$:

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

We propose a constrained maximum quasi-likelihood estimator (M-estimator),

$$(\hat{\mathcal{C}}, \hat{\mathbf{M}}_1, \dots, \hat{\mathbf{M}}_k) = \arg \max_{(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K) \in \mathcal{P}} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K), \quad (6)$$

where the parameter space \mathcal{P} is

$$\mathcal{P} = \left\{ (\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K) \mid \mathbf{M}_k^T \mathbf{M}_k = \mathbf{I}_{r_k}, \|\Theta(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K)\|_\infty \leq \alpha \right\}.$$

The maximum norm constraint on the linear predictor Θ is a technical condition to avoid the divergence in the non-Gaussian variance.

4.2 Alternating optimization

We propose an alternating optimization algorithm to solve (6). The decision variables in the objective function (6) consist of $K+1$ blocks of variables, one for the core tensor \mathcal{C} and K for the factor matrices \mathbf{M}_k . We notice that, if any K out of the $K+1$ blocks of variables are known, then the optimization reduces to a simple GLM with respect to the last block of variables. This observation leads to an iterative updating scheme for 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. The full algorithm is described in Algorithm 1.

The optimization (6) is a non-convex problem due to the non-convexity in the feasible

Algorithm 1 Supervised Tensor Decomposition with Interactive Side Information

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

Output: Estimated core tensor $\hat{\mathcal{C}} \in \mathbb{R}^{r_1 \times \dots \times r_K}$ and factor matrices $\hat{\mathbf{M}}_k \in \mathbb{R}^{p_k \times r_k}$.

- 1: Random initialization of the core tensor \mathcal{C} and factor matrices \mathbf{M}_k .
- 2: **while** Do until convergence **do**
- 3: **for** $k = 1$ to K **do**
- 4: Obtain the factor matrix $\tilde{\mathbf{M}}_k \in \mathbb{R}^{p_k \times r_k}$ by a GLM with link function f .
- 5: Perform QR factorization $\tilde{\mathbf{M}}_k = \mathbf{Q}\mathbf{R}$.
- 6: Update $\mathbf{M}_k \leftarrow \mathbf{Q}$ and core tensor $\mathcal{C} \leftarrow \mathcal{C} \times_k \mathbf{R}$.
- 7: **end for**
- 8: Update the core tensor \mathcal{C} by solving a GLM with $\text{vec}(\mathcal{Y})$ as response, $\otimes_{k=1}^K [\mathbf{X}_k \mathbf{M}_k]$ as features, and f as link function. Here \otimes denotes the Kronecker product of matrices.
- 9: Rescale the core tensor \mathcal{C} subject to the maximum norm constraint.
- 10: **end while**

set \mathcal{P} . Under mild conditions, our algorithm enjoys global convergence; i.e. any sequence of iterates generated by the alternating algorithm converges to a stationary point of $\mathcal{L}_{\mathcal{Y}}(\cdot)$ module orthogonal transformation (see Supplementary Notes, Section ??). In theory, global optimality of non-convex optimization is often challenging to obtain. Fortunately, we will show in Section 4.3 that the desired statistical performance holds for all local optimizers satisfying $\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \hat{\mathbf{M}}_1, \dots, \hat{\mathbf{M}}_K) \geq \mathcal{L}_{\mathcal{Y}}(\mathcal{C}_{\text{true}}, \mathbf{M}_{1,\text{true}}, \dots, \mathbf{M}_{K,\text{true}})$, where the subscript “true” denotes the true parameter in model (3). This result indicates the global optimality is not necessarily a serious concern in our context, as long as the convergent objective is large enough. In the experiments we considered, we find that Algorithm 1 typically gives satisfactory convergent points upon random initialization. Figure 2 shows the trajectories of objective function for order-3 tensors based on model (3), where $d_k = d \in \{25, 30\}$, $p_k = 0.4d$, $r_k = r \in \{3, 6\}$ for $k = 1, 2, 3$. We consider input tensors with Gaussian, Bernoulli, and Poisson entries. Under all combinations of the dimension d , rank r , and type of the entries, Algorithm 1 converges quickly in a few iterations, and the objective values at convergent points are close to or larger than the value at true parameters. The time complexity is provided in Supplementary Notes.

Algorithm 1 assumes the rank \mathbf{r} is given. In practice, the rank is often unknown and

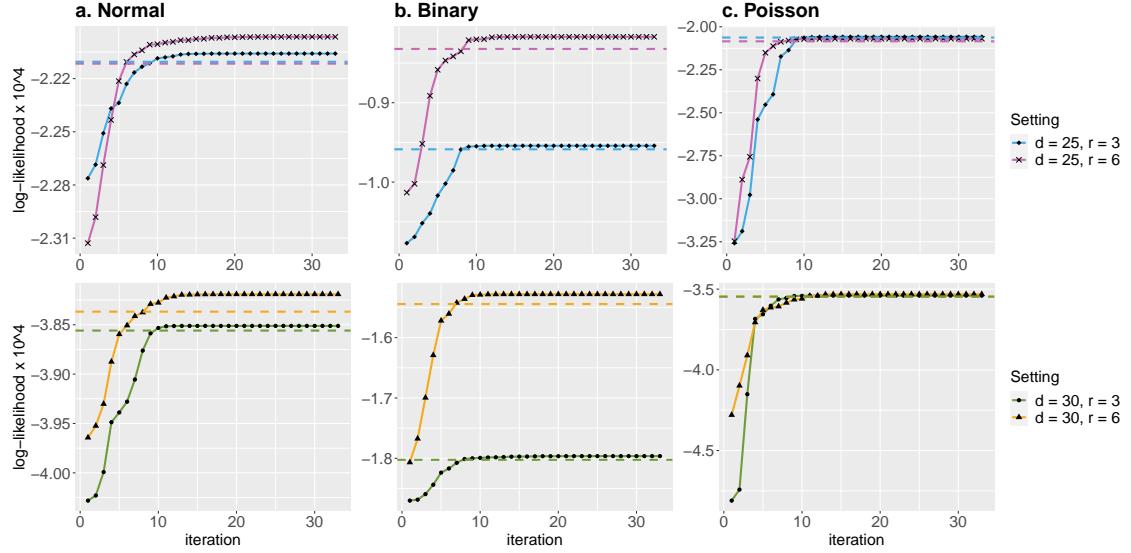


Figure 2: Trajectory of the objective function with various dimension d and rank r under (a) Gaussian (b) Bernoulli (c) Poisson models. The dashed line represents the objective value at true parameters.

must be determined from the data. We propose to use Bayesian information criterion (BIC) and choose the rank that minimizes BIC. The detailed formula of BIC and the simulation which evaluates the empirical performance of BIC are provided in Supplementary Notes.

4.3 Statistical properties

In this section, we provide the accuracy guarantee for the proposed M-estimator (6). Note that the factor matrices \mathbf{M}_k are identifiable only up to orthogonal transformation. Therefore, we use angle distance to assess the accuracy in estimating the column space, $\text{Span}(\mathbf{M}_k)$. For any two matrices \mathbf{A}, \mathbf{B} with orthonormal columns of same dimension, the angle distance is defined as

$$\sin \Theta(\mathbf{A}, \mathbf{B}) = \|\mathbf{A}^T \mathbf{B}^\perp\|_\sigma = \max \left\{ \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} : \mathbf{x} \in \text{Span}(\mathbf{A}), \mathbf{y} \in \text{Span}(\mathbf{B}^\perp) \right\}.$$

In modern applications, the tensor data and features are often large-scale. We are particularly interested in the high-dimensional regime in which both d_k and p_k diverge; i.e. $d_k \rightarrow \infty$ and $p_k \rightarrow \infty$, while $p_k/d_k \rightarrow \gamma_k \in [0, 1]$. As the size of problem grows, and so does

the number of unknown parameters. 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 $c_1 \leq \sigma_{\min}(\mathbf{X}_k) \leq \sigma_{\max}(\mathbf{X}_k) \leq c_2$ for all $k \in [K]$.
- A1'. The feature matrices \mathbf{X}_k are Gaussian designs with i.i.d. $N(0, 1)$ entries.
- A2. There exist two positive constants $L, U > 0$, such that $L\phi \leq \text{Var}(y_{i_1, \dots, i_K} | \theta_{i_1, \dots, i_K}) \leq U\phi$, or equivalently, $L \leq b''(\theta_{i_1, \dots, i_K}) \leq U$, for all $|\theta_{i_1, \dots, i_K}| \leq \alpha$. Here α is the upper bound of the linear predictor in (6), and $b''(\cdot)$ denotes the second-order derivative.

The assumptions are fairly mild. Assumptions A1 and A1' consider two separate scenarios about feature matrices. Assumption A1 is applicable when feature matrix is asymptotically non-singular and has bounded spectral norm, whereas Assumption A1' imposes the commonly-used Gaussian design (Raskutti et al., 2019). Assumption A2 ensures the log-likelihood $\mathcal{L}_{\mathcal{Y}}(\Theta)$ is strictly concave in the linear predictor Θ .

Theorem 4.1 (Statistical convergence). Consider a data tensor generated from model (3), where the entries are conditionally independent realizations from an exponential family. Let $(\hat{\mathcal{C}}, \hat{\mathbf{M}}_1, \dots, \hat{\mathbf{M}}_K)$ be the M-estimator in (6) and $\hat{\mathcal{B}} = \hat{\mathcal{C}} \times \hat{\mathbf{M}}_1 \times \dots \times \hat{\mathbf{M}}_K$. Define $r_{\text{total}} = \prod_k r_k$ and $r_{\max} = \max_k r_k$. Under Assumptions A1 and A2 with scaled feature matrices $\check{\mathbf{X}}_k = \sqrt{d_k} \mathbf{X}_k$, or under Assumptions A1' and A2 with original feature matrices, there exist two positive constants $C_1 = C_1(\alpha, K), C_2 = C_2(\alpha, K) > 0$ independent of dimensions $\{d_k\}$ and $\{p_k\}$, such that, with probability at least $1 - \exp(-C_1 \sum_k p_k)$,

$$\|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}\|_F^2 \leq \frac{C_2 r_{\text{total}}}{r_{\max}} \frac{\sum_k p_k}{\prod_k d_k}. \quad (7)$$

Furthermore, if the unfolded core tensor has non-degenerate singular values at mode $k \in$

$[K]$, i.e., $\sigma_{\min}(\text{Unfold}_k(\mathcal{C}_{\text{true}})) \geq c > 0$ for some constant c , then

$$\sin^2 \Theta(\mathbf{M}_{k,\text{true}}, \hat{\mathbf{M}}_k) \leq \frac{C_2 r_{\text{total}}}{r_{\max} \sigma_{\min}^2(\text{Unfold}_k(\mathcal{C}_{\text{true}}))} \frac{\sum_k p_k}{\prod_k d_k}.$$

Theorem 4.1 establishes the statistical convergence for the estimator (6). In fact, our proof shows that the desired convergence rate holds not only for the M-estimator, but also for any local optimizers satisfying $\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \hat{\mathbf{M}}_1, \dots, \hat{\mathbf{M}}_K) \geq \mathcal{L}_{\mathcal{Y}}(\mathcal{C}_{\text{true}}, \mathbf{M}_{1,\text{true}}, \dots, \mathbf{M}_{K,\text{true}})$. Consider a special case when tensor dimensions are equal on each of the modes, i.e., $d_k = d$ for all $k \in [K]$, and feature dimension grows with tensor dimension, $p_k = \gamma d$, $\gamma \in [0, 1)$, for $k \in [K]$. The result in (7) implies that the estimation has a convergence rate $\mathcal{O}(d^{-(K-1)})$. Therefore, our estimation is consistent in high dimensional regimes, and the convergence becomes especially favorable as the order of tensor data increases. As immediate applications, we obtain the convergence rate for the three examples mentioned in Section 3.

Example 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 along either of the three modes.

Example 2 (Network population model). The estimated node-by-node-by-feature 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.

5 Numerical experiments

We evaluate the empirical performance of our supervised tensor decomposition (STD) through simulations. We consider order-3 tensors with a range of distribution types. Unless otherwise specified, the conditional mean tensor is generated from model (3), where the core tensor entries are i.i.d. drawn from Uniform[-1,1], the factor matrix \mathbf{M}_k is uniformly sampled with respect to Haar measure from matrices with orthonormal columns. The feature matrix \mathbf{X}_k is either an identity matrix (i.e., no feature available) or Gaussian random matrix with i.i.d. entries from $N(0, 1)$. The linear predictor $\Theta = \mathcal{C} \times \{\mathbf{M}_1 \mathbf{X}_1, \mathbf{M}_2 \mathbf{X}_3, \mathbf{M}_3 \mathbf{X}_3\}$

is scaled such that $\|\Theta\|_\infty = 1$. Conditional on the linear predictor $\Theta = \llbracket \theta_{ijk} \rrbracket$, the entries in the tensor $\mathcal{Y} = \llbracket y_{ijk} \rrbracket$ are drawn independently according to three probabilistic models: (a) Gaussian model: continuous tensor entries $y_{ijk} \sim N(\alpha\theta_{ijk}, 1)$; (b) Poisson model: count tensor entries $y_{ijk} \sim \text{Poisson}(e^{\alpha\theta_{ijk}})$; (c) Bernoulli model: binary tensor entries $y_{ijk} \sim \text{Bernoulli}\left(\frac{e^{\alpha\theta_{ijk}}}{1+e^{\alpha\theta_{ijk}}}\right)$. Here $\alpha > 0$ is a scalar controlling the magnitude of the effect size. In each simulation study, we report the mean squared error (MSE), $\|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}\|_F^2$, averaged across 30 replications.

5.1 Comparison with GLMs under stochastic block models

The first experiment investigates the performance of our model under correlated feature effects. 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 $p = 5$ features for each of the 50 individuals. These features 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 (Abbe, 2017) 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 feature 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 network is not necessarily equal to matrix rank r (Wang and Zeng, 2019).

Figure 3 compares the MSE of our method with a multiple-response GLM approach. The multiple-response GLM is to regress the dyadic edges, one at a time, on the features, and this model is repeatedly fitted for each edge. As we find in Figure 3, our tensor regression method achieves significant error reduction in all three data types considered. The outperformance is substantial in the presence of large communities; even in the less

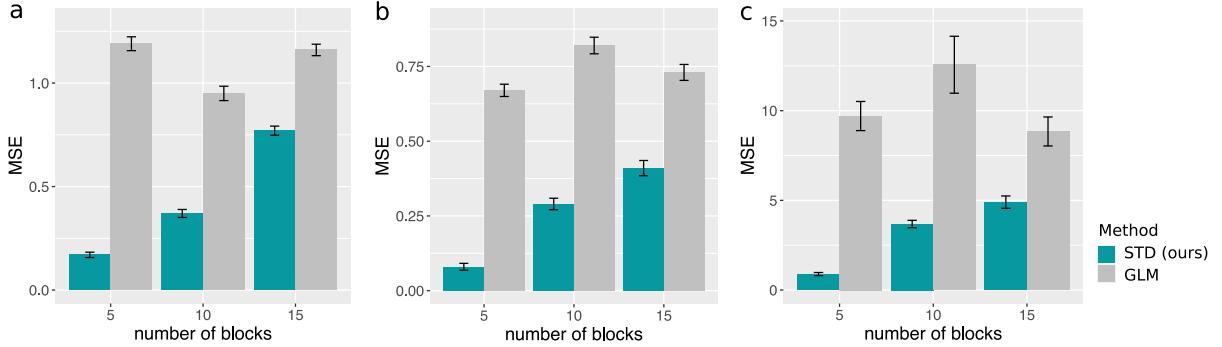


Figure 3: Performance comparison under stochastic block models. The three panels plot the MSE when the response tensors are generated from (a) Gaussian (b) Poisson and (c) Bernoulli models. The x -axis represents the number of blocks in the networks.

structured case ($\sim 20/15 = 1.33$ nodes per block), our method still outer-performs GLM. The possible reason is that the multiple-response GLM approach does not account for the correlation among the edges, and suffers from overfitting. In contrast, the low-rankness in our modeling incorporates the shared information across entries. By selecting the rank in a data-driven way, our method achieves accurate estimation in a wide range of settings.

5.2 Comparison with other tensor methods

We compare our supervised tensor decomposition with three other tensor methods: Higher-order low-rank regression (**HOLRR**, Rabusseau and Kadri (2016)), Higher-order partial least square (**HOPLS**, Zhao et al. (2012)), and Subsampled tensor projected gradient (**TPG**, Yu and Liu (2016)).

These three methods are the closest algorithms to ours, in that they all relate a data tensor to features using a low-rank structure. The three existing methods allow only Gaussian data, whereas ours is applicable to any exponential family distribution including Gaussian, Bernoulli, Poisson, etc. For fair comparison, we consider Gaussian tensors in the experiment. Because not every method returns the effect estimate $\hat{\mathcal{B}}$ as outputs, we measure the accuracy using mean squared prediction error, $\text{MSPE} = (\prod_k d_k)^{-1} \|\hat{\mathcal{Y}} - f(\Theta)\|_F^2$, where $f(\Theta)$ is the conditional mean of the tensor, and $\hat{\mathcal{Y}}$ is the fitted tensor from each method.

The comparison is assessed from three aspects: (a) benefit of incorporating features 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, $\mathbf{r} = (3, 3, 3)$ (low) vs. $(4, 5, 6)$ (high), signal $\alpha = 3$ (low) vs. 6 (high), dimension d ranging from 20 to 100 for modes with features, and $d = 20$ for modes without features. Two methods (**STD** and **HOLRR**) require the tensor rank as inputs. For fair comparison, we provide both algorithms the true rank. For algorithms (**HOPLS** and **TPG**) that use different notions of model rank, we use a grid search to set the hyperparameter that gives the best mean square prediction error.

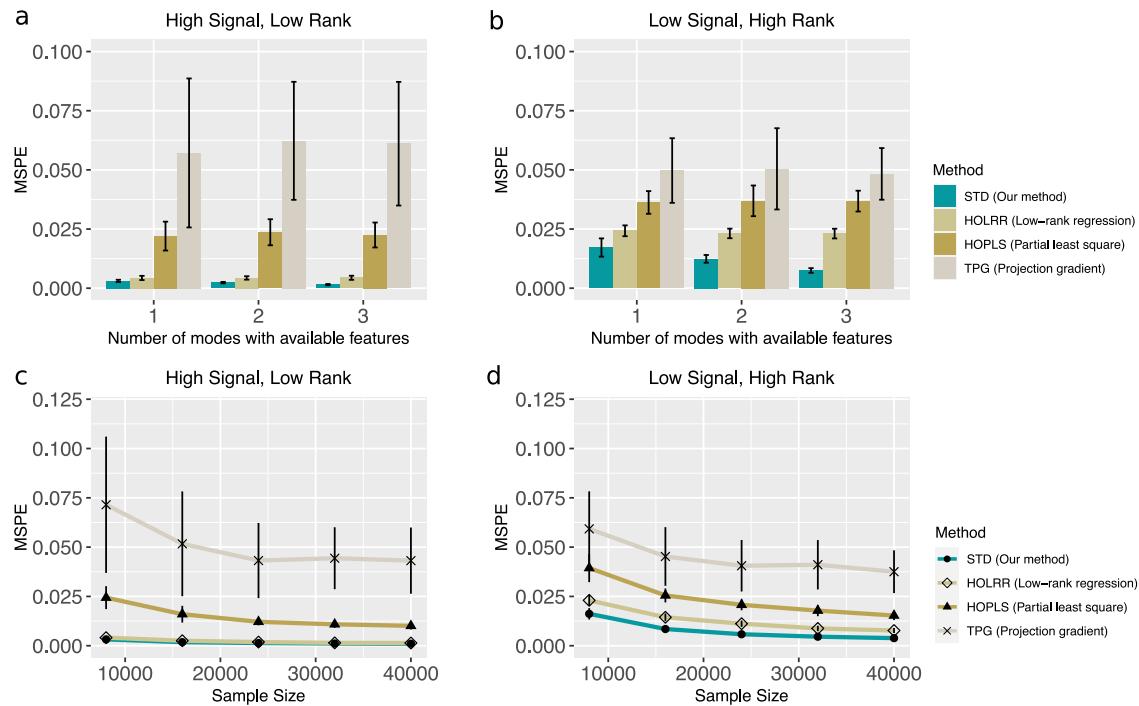


Figure 4: Comparison between different tensor methods. Panels (a) and (b) plot MSPE versus the number of modes with available features. Panels (c) and (d) plot MSPE versus the effective sample size d^2 .

Figures 4a-b show the averaged prediction error across 30 replicates. We see that our **STD** outperforms others, especially in the low-signal, high-rank setting. As the number of informative modes (i.e., modes with available features) increases, the **STD** exhibits a substantial reduction in error whereas others remain unchanged (Figure 4b). This showcases the benefit of incorporation of multiple features. Note that our method **STD** 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. **STD** alternates between informative and non-informative modes, whereas **HOLRR** approximates the non-informative modes without alternating. The accuracy gain in Figure 4 demonstrates the benefit of alternating algorithm – incorporation of informative modes also improves the estimation in the non-informative modes.

Figures 4c-d compare the prediction error with respect to effective sample size when only one mode has side information. In the high-signal low-rank setting, our method has similar performance as **HOLRR**, and the improvement becomes more pronounced in the low-signal high-rank setting. The latter setting is harder because of the higher inter-mode complexity, and our **STD** method shows the advantage in addressing this challenge.

6 Data analyses

6.1 Application to human brain connection data

The Human Connectome Project (HCP) aims to build a network map that characterizes the anatomical and functional connectivity within healthy human brains (Geddes, 2016). We follow the preprocessing procedure as in Zhang et al. (2018) and parcellate the brain into 68 regions of interest (Desikan et al., 2006). The dataset 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 features: 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 individual features. A key challenge in brain network is that the edges are correlated; for example, the nodes in edges may be from a same brain region, and it is of importance to take into account the within-dyad dependence.

We perform the supervised tensor decomposition to the HCP data. The BIC selection suggests a rank $\mathbf{r} = (10, 10, 4)$ with quasi log-likelihood $\mathcal{L}_Y = -174654.7$. We utilize the

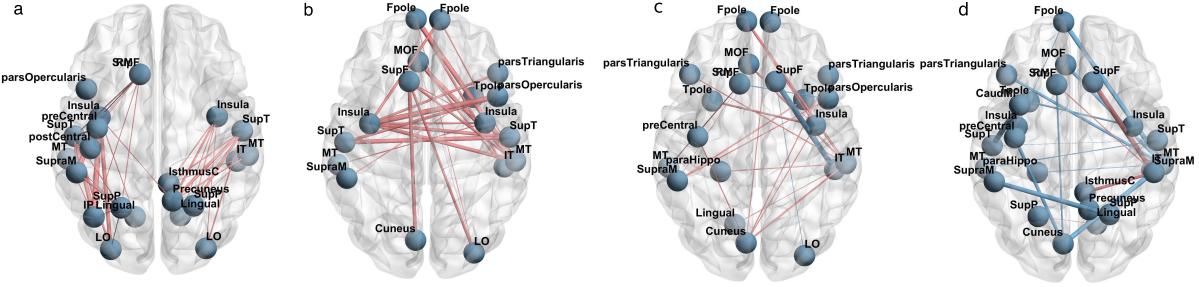


Figure 5: Top edges with large effects. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+. Red edges represent positive effects and blue edges represent negative effects. The edge-width is proportional to the magnitude of the effect size.

sum-to-zero contrasts in coding the feature effects, and depict only the top 3% edges whose connections are non-constant across the sample. Figure 5 shows the top edges with high effect size, overlaid on the Desikan atlas brain template (Desikan et al., 2006). We find that the global connection exhibits clear spatial separation, and that the nodes within each hemisphere are more densely connected with each other (Figure 5a). In particular, the superior-temporal (*SupT*), middle-temporal (*MT*) and Insula are the top three popular nodes in the network. Interestingly, female brains display higher inter-hemispheric connectivity, especially in the frontal, parental and temporal lobes (Figure 5b). This is in agreement with a recent study showing that female brains are optimized for inter-hemispheric communication (Ingalhalikar et al., 2014). We find several edges with declined connection in the group Age 31+. Those edges involve Frontal-pole (*Fpole*), superior-frontal (*SupF*) and Cuneus nodes. The Frontal-pole region is known for its importance in memory and cognition, and the detected decline with age further highlights its biological importance.

6.2 Application to political relation data

The second application studies the multi-relational networks with node-level attributes. We consider *Nations* dataset (Nickel et al., 2011) 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 an action, such as “sending tourist to”, “export”, “import”, between countries. The 56 relations span

the fields of politics, economics, military, religion, etc. In addition, country-level attributes are also available, and we focus on the following six features: *constitutional*, *catholics*, *law ngo*, *political leadership*, *geography*, and *medicine ngo*. The goal is to identify the variation in connections due to country-level attributes and their interactions.

We apply our tensor model to the *Nations* data. The multi-relational network \mathcal{Y} is a binary data tensor, and the country attributes $\mathbf{X} \in \mathbb{R}^{14 \times 6}$ are features on both the 1st and 2nd modes. The BIC criterion suggests a rank $\mathbf{r} = (4, 4, 4)$ for the coefficient tensor $\mathcal{B} \in \mathbb{R}^{6 \times 6 \times 56}$. We perform the supervised tensor decomposition and obtain the dimension reduction matrices $\hat{\mathbf{M}}_k$ from the model. Then we apply K -mean clustering to dimension reduction matrix on each of the modes. Table 2 shows the K-means clustering of the 56 relations based on the dimension reduction matrix on the 3rd mode. We find that the relations reflecting the similar aspects of actions are grouped together. In particular, Cluster I consists of military relations such as *violentactions*, *warnings* and *militaryactions*; Clusters II and III capture the economic relations such as *economicaid*, *booktranslations*, *tourism*; and Cluster IV represents the political alliance and territory relations. More analyses about the effects of dyadic attributes towards the connections are in the Supplementary Notes.

	Category	Relations
I	Military	warning, violentactions, militaryactions, duration, negativebehavior, protests, severdiplomatictimesincewar, commonbloc0, commonbloc1, blockpositionindex, expeldiplomats
II	Emigrant	emigrants, emigrants3, relemigrants, accusation, nonviolentbehavior, ngoorgs, commonbloc2, intergovorgs3, releconomicaid, relintergovorgs, relngo, students, relstudents, economicaid, negativecomm, militaryalliance
III	Economics	treaties, reltreaties, officialvisits, exportbooks, relexportbooks, booktranslations, relbooktranslations, boycottembargo, weightedunvote, unweightedunvote, reltourism, tourism, tourism3, exports, exports3, reexports, intergovorgs, ngo, embassy, reldiplomacy, timesinceally, independence, conferences, dependent
IV	Territory	aidenemy, lostterritory, unoffialacts, attackembassy

Table 2: K -means clustering of relations based on dimension reduction on the 3rd mode.

7 Conclusion

We have developed a supervised tensor decomposition method with side information on multiple modes. One important challenge of tensor data analysis is the complex interdependence among tensor entries and between multiple features. Our approach incorporates side information as feature matrices in the conditional mean tensor. The empirical results

demonstrate the improved interpretability and accuracy over previous approaches. Applications to the brain connection and political relationship datasets yield conclusions with sensible interpretations, suggesting the practical utility of the proposed approach.

References

- Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.
- Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405.
- Berthet, Q. and Baldin, N. (2020). Statistical and computational rates in graph logistic regression. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 108:2719–2730.
- De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.
- Farias, V. F. and Li, A. A. (2019). Learning preferences with side information. *Management Science*, 65(7):3131–3149.
- Gabriel, K. R. (1998). Generalised bilinear regression. *Biometrika*, 85(3):689–700.
- Gahrooei, M. R., Yan, H., Paynabar, K., and Shi, J. (2020). Multiple tensor-on-tensor regression: An approach for modeling processes with heterogeneous sources of data. *Technometrics*, pages 1–23.
- Geddes, L. (2016). Human brain mapped in unprecedented detail. *Nature*.
- Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295.

- Hong, D., Kolda, T. G., and Duersch, J. A. (2020). Generalized canonical polyadic tensor decomposition. *SIAM Review*, 62(1):133–163.
- Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Hakonarson, H., Gur, R. E., Gur, R. C., and Verma, R. (2014). Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828.
- Ioannidis, V. N., Zamzam, A. S., Giannakis, G. B., and Sidiropoulos, N. D. (2019). Coupled graph and tensor factorization for recommender systems and community detection. *IEEE Transactions on Knowledge and Data Engineering*.
- Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.
- Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics*, 27(3):638–647.
- Lock, E. F. and Li, G. (2018). Supervised multiway factorization. *Electronic journal of statistics*, 12(1):1150.
- Narita, A., Hayashi, K., Tomioka, R., and Kashima, H. (2012). Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery*, 25(2):298–324.
- Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*, volume 11, pages 809–816.
- Rabusseau, G. and Kadri, H. (2016). Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems*, pages 1867–1875.
- Raskutti, G., Yuan, M., Chen, H., et al. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584.
- Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. arXiv:1906.03807.
- Yu, R. and Liu, Y. (2016). Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381.

- Zhang, Z., Descoteaux, M., Zhang, J., Girard, G., Chamberland, M., Dunson, D., Srivastava, A., and Zhu, H. (2018). Mapping population-based structural connectomes. *NeuroImage*, 172:130–145.
- Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Nagasaka, Y., Fujii, N., Zhang, L., and Cichocki, A. (2012). Higher order partial least squares (HOPLS): a generalized multilinear regression method. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1660–1673.
- Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.