Learning Multiple Networks via Supervised Tensor Decomposition

Anonymous Author(s)
Affiliation
Address
email

Abstract

We develop a tensor decomposition method that incorporates consider the problem of tensor decomposition with multiple side information available as interactive features. Unlike classical tensor decomposition, our supervised decomposition captures the effective dimension reduction of the data tensor confined to feature space on each modeSuch problems are common in neuroimaging, network modeling, and spatial-temporal analysis. We develop a new family of exponential tensor decomposition models and establish the theoretical accuracy guarantees. An efficient alternating optimization algorithm is further developed. Our Unlike earlier methods, our proposal handles a broad range of data types, including continuous, count, and binary observations, along with available features. We apply the method to diffusion tensor imaging data from human connectome project . We and identify the key global brain connectivity pattern and pinpoint the local regions that are brain connectivity patterns associated with available features. Our method will help the practitioners efficiently analyze tensor datasets in various areas. Toward this end, all data and code has been made available to the public.

1 Introduction

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. 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 Baldin, 2020; Hoff, 2005). A typical social network consists of nodes that represent people and edges that represent friendships. (Berthet and Baldin, 2020). 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 scientists are interested in identifing 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 challenge of incorporating side information, many tensor datasets consist of non-Gaussian measurements. 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). These methods often assume Gaussian distribution for the tensor entries, or impose random designs for the feature matrices, both of which are less suitable

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

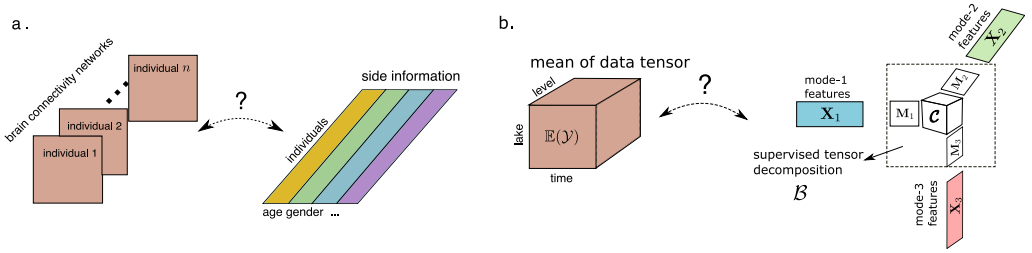

Figure 1: Examples of supervised tensor decomposition with interactive side information. (a) Network population model. (b) Spatio-temporal growth model.

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~~ **Our contribution.** This paper presents 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~~ low-rank tensor regression problem, with tensor observation serving as the response, and the multiple side information as interactive features. We ~~leverage~~ blend the modeling power of generalized linear model (GLM) ~~to allow heteroscedacity due to the mean-variance relationship in the non-Gaussian data. The low-rank structure on the conditional mean tensor effectively mitigates the curse of high dimensionality. Our proposal blends the modeling power of GLM~~ and the exploratory capability of tensor dimension reduction in order to take the best out of both worlds. Our methods greatly improves the classical tensor decomposition, and we quantify the gain in prediction through numerical experiments and data applications.

2 ~~Method~~

1.1 ~~Preliminary~~

~~We introduce the basic tensor properties used in the paper. 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 letters (e. g., $\mathcal{A}, \mathcal{B}, \mathcal{C}$) for tensors of order three or greater. Let $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K $(d_1, \ldots, d_K)$-dimensional tensor~~ **Notation.** We follow the tensor notation as in Kolda and Bader (2009). The multilinear multiplication of a tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ by matrices $\boldsymbol{X}_k = [\![x^{(k)}_{i_k, j_k}]\!] \in \mathbb{R}^{p_k \times d_k}$ is defined as $\mathcal{Y} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} = [\![\sum_{i_1,\ldots,i_K} y_{i_1,\ldots,i_K} x^{(1)}_{j_1,i_1} \cdots x^{(K)}_{j_K,i_K}]\!]$, which results in an order-K $(p_1, \ldots, p_K)$-dimensional tensor. ~~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~~ The inner product between two tensors of equal size 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 tensor Frobenius norm is defined as $\|\mathcal{Y}\|_F = \langle \mathcal{Y}, \mathcal{Y} \rangle^{1/2}$, and the maximum norm is defined $\|\mathcal{Y}\|_\infty = \max_{i_1,\ldots,i_K} y_{i_1,\ldots,i_K}$. We let $\boldsymbol{I}_d$ denote the $d \times d$ identity matrix and $[d]$ denote the d-set $\{1, \ldots, d\}$.~~ ~~$\langle \mathcal{Y}, \mathcal{Y}' \rangle = \sum_{i_1,\ldots,i_K} y_{i_1,\ldots,i_K} y'_{i_1,\ldots,i_K}$.~~ For ease of notation, we allow basic arithmetic operators (e.g., $+, -$) and univariate functions $f : \mathbb{R} \to \mathbb{R}$ to be applied to tensors in an element-wise manner.

1.1 ~~General Model~~

2 Proposed models and motivating examples

Let $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K data tensor. Suppose the side information is available on each of the K modes. Let $\boldsymbol{X}_k = [\![x_{ij}]\!] \in \mathbb{R}^{d_k \times p_k}$ 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 $(i, j) \in [d_k] \times [p_k]$, $p_k \leq d_k$. ~~Assume~~ We assume that, conditional on the features $\boldsymbol{X}_k$, the entries of tensor $\mathcal{Y}$ are independent realizations from an exponential family distribution, and the conditional mean tensor

admits the form

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K) = f(\Theta), \text{with} \Theta = \left(\mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1,\ldots,\boldsymbol{X}_K\boldsymbol{M}_K\}\right),$$
$$\text{with } \boldsymbol{M}_k^T\boldsymbol{M}_k = \boldsymbol{I}_{r_k},\ \boldsymbol{M}_k \in \mathbb{R}^{p_k\times r_k} \quad \text{for all } k = 1,\ldots,K. \tag{1}$$

where ~~$\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is the multilinear predictor, $\mathcal{B} \in \mathbb{R}^{p_1\times\cdots\times p_K}$ is the unknown parameter tensor, $f(\cdot)$ is a known link function whose form depending on the data type of $\mathcal{Y}$. The choice of link function is based on the assumed distribution family of tensor entries.~~

~~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 $\mathcal{B}$,~~

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

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

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K) = f\left(\mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1,\ldots,\boldsymbol{X}_K\boldsymbol{M}_K\}\right),$$
$$\text{with } \boldsymbol{M}_k^T\boldsymbol{M}_k = \boldsymbol{I}_{r_k},\ \boldsymbol{M}_k \in \mathbb{R}^{p_k\times r_k} \quad \text{for all } k = 1,\ldots,K,$$

~~where the parameters of interest are $\boldsymbol{M}_k$ and $\mathcal{C}$. Note that model assumes a fixed, known rank $\boldsymbol{r} = (r_1,\ldots,r_K)$.~~ , $f(\cdot)$ is a known link function whose form depending on the data type of $\mathcal{Y}$. Common choices of link functions include identity link for Gaussian distribution, logistic link for Bernoulli distribution, and $\exp(\cdot)$ link for Poisson distribution.

Figure 1b provides a schematic illustration of our model. The features $\boldsymbol{X}_k$ affect the distribution of tensor entries in $\mathcal{Y}$ through the form $\boldsymbol{X}_k\boldsymbol{M}_k$, which are r_k linear combinations of features on mode k. ~~We call $\boldsymbol{X}_k\boldsymbol{M}_k$ the "supervised tensor factors" or "sufficient features" (Adragni and Cook, 2009) , and call $\boldsymbol{M}_k$ the "dimension reduction matrix."~~ The core tensor $\mathcal{C}$ collects the interaction effects between sufficient features across K modes, which links the conditional mean to the feature spaces, and thereby allows the identification of variations in the tensor data attributable to the side information. Our goal is to find $\boldsymbol{M}_k$ and the corresponding $\mathcal{C}$, thereby allowing us to reveal the relationship between side information $\boldsymbol{X}_k$ and the observed tensor $\mathcal{Y}$. Note that $\boldsymbol{M}_k$ and $\mathcal{C}$ are identifiable only up to orthonormal transformations.

3 ~~Estimation~~

2.1 ~~Rank-constrained M-estimator~~

~~We adopt the exponential family as a flexible framework for different datatypes. In a classical generalized linear model with a scalar response y and feature $\boldsymbol{x}$, the density is expressed as~~

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

~~where $b(\cdot)$ is a known function depending on the data types, θ is~~ We give two examples of supervised tensor decomposition models (1) that arise in practice.

Example 1 (Spatio-temporal growth model)**.** The growth curve model (Srivastava et al., 2008) was originally proposed as an example of bilinear model for matrix data, and we extend it to higher-order cases. 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 is a special case of our model (1), where $\boldsymbol{X}_1 = \text{blockdiag}\{\mathbf{1}_q,\ldots,\mathbf{1}_q\} \in \{0,1\}^{d\times p}$ is the ~~linear~~

predictor, $\phi > 0$ is the dispersion parameter, and $c(\cdot)$ is a known normalizing function. In our context, we modelthe tensor entries $y_{i_1,\ldots,i_K}$, conditional on $\theta_{i_1,\ldots,i_K}$, as independent draws from an exponential family. 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. The spatial-temporal mode has covariates available on each of the three modes.

Example 2 (Network population model)**.** Network response model is recently developed in the context of neuroimanig analysis. The goal is to study the relationship between network-valued response and the individual covariates. Suppose we observe n i.i.d. observations $\{(\boldsymbol{Y}_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$, where $\boldsymbol{Y}_i \in \{0,1\}^{d\times d}$ is the brain connectivity network on the i-th individual, and $\boldsymbol{x}_i \in \mathbb{R}^p$ is the individual covariate such as age, gender, cognition, etc. The network-response model (Rabusseau and Kadri, 2016) 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. The model (2) is also a special case of our tensor-response model, with covariates on the last mode of the tensor.

3 Estimation algorithms

We develop a likelihood-based procedure to estimate $\mathcal{C}$ and $\boldsymbol{M}_k$ in (1). Ignoring constants that do not depend on Θ, the quasi log-likelihood of (1) is equal to Bregman distance between $\mathcal{Y}$ and $b'(\Theta)$:

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

where $b(\theta) = \theta^2/2$ for Gaussian response, $b(\theta) = \exp(\theta)$ for Poisson response, and $b(\theta) = \log(1 + \exp(\theta))$ for Bernoulli response. We propose a constrained maximum quasi-likelihood estimator (M-estimator),

$$(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_1, \ldots, \hat{\boldsymbol{M}}_k) = \underset{(\mathcal{C},\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K)\in\mathcal{P}}{\arg\max} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K), \tag{3}$$

where parameter space $\mathcal{P} = \left\{(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K) \;\middle|\; \boldsymbol{M}_k^T\boldsymbol{M}_k = \mathbf{I}_{r_k},\ \|\Theta(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K)\|_\infty \le \alpha\right\}$.
The maximum norm constraint on the linear predictor Θ is a technical condition to avoid the divergence in the non-Gaussian variance.

3.1 Alternating optimization

The decision variables in the objective function (3) consist of $K + 1$ blocks of variables, one for the core tensor $\mathcal{C}$ and K for the factor matrices $\boldsymbol{M}_k$. 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. A simplified version of the algorithm is described in Algorithm 1.

3.1 Statistical properties

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. $d_k \to \infty$ and $p_k \to \infty$, while $p_k/d_k \to \gamma_k \in [0, 1)$. 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 We provide the accuracy guarantee for the proposed M-estimator (3) by leveraging recent development in random tensor theory and high-dimensional statisticsto establish the error bounds.

Consider a data tensor generated from model .

Algorithm 1 Supervised Tensor Decomposition with Interactive Side Information (Simplified)

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

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

1: Random initialization of the core tensor $\mathcal{C}$ and factor matrices $\boldsymbol{M}_k$.
2: **while** Do until convergence **do**
3: Obtain $\tilde{\boldsymbol{M}}_k \in \mathbb{R}^{p_k \times r_k}$ by a GLM. Orthogonalize $\tilde{\boldsymbol{M}}_k$ by QR factorization, for $k \in [K]$.
4: Update the core tensor $\mathcal{C}$ by solving a GLM. Rescale the core tensor $\mathcal{C}$ such that $\|\mathcal{C}\|_{\max} \leq \alpha$.
5: **end while**

Theorem 3.1 (Convergence)**.** Let $(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_1, \ldots, \hat{\boldsymbol{M}}_K)$ be the M-estimator in (3) and $\hat{\mathcal{B}} = \hat{\mathcal{C}} \times \hat{\boldsymbol{M}}_1 \times \cdots \times \hat{\boldsymbol{M}}_K$. Define $r_{\text{total}} = \prod_k r_k$ and $r_{\max} = \max_k r_k$. Under mild technical assumptions, there exist two positive constants ~~$C_1 = C_1(\alpha, K), C_2 = C_2(\alpha, K) > 0$ independent of dimensions $\{d_k\}$ and $\{p_k\}$~~$C_1, C_2 > 0$, such that, with probability at least $1 - \exp(-C_1 \sum_k p_k)$,

$$\|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}\|_F^2 \leq \frac{C_2 r_{\text{total}}}{r_{\max}} \frac{\sum_k p_k}{\prod_k d_k}, \quad \text{and} \quad \sin^2 \Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k) \leq \frac{C_2 r_{\text{total}}}{r_{\max} \sigma_{\min}^2(\text{Unfold}_k(\mathcal{C}_{\text{true}}))} \frac{\sum_k p_k}{\prod_k d_k},$$

~~Furthermore, if the unfolded core tensor has non-degenerate singular values at mode $k \in [K]$~~where $\sin \Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k) = \|\boldsymbol{M}_{k,\text{true}}^T \hat{\boldsymbol{M}}_k^{\perp}\|_\sigma$ is the angle distance between column spaces.

Theorem 3.1 implies that the estimation has a convergence rate $\mathcal{O}(d^{-(K-1)})$ in the special case when tensor dimensions are equal on each of the modes, i.e., ~~$\sigma_{\min}(\text{Unfold}_k(\mathcal{C}_{\text{true}})) \geq c > 0$ for some constant c, then~~

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

~~where $\sin \Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k) = \|\boldsymbol{M}_{k,\text{true}}^T \hat{\boldsymbol{M}}_k^{\perp}\|_\sigma = \max \left\{ \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2} : \boldsymbol{x} \in \text{Span}(\boldsymbol{M}_{k,\text{true}}), \boldsymbol{y} \in \text{Span}(\hat{\boldsymbol{M}}_k^{\perp}) \right\}$ is the angle distance to assess the accuracy in estimating the column space, $\text{Span}(\boldsymbol{M}_k)$.~~ $d_k = d$ 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 convergence of our estimation becomes especially favorable as the order of tensor data increases.

~~4 Real data analysis~~

4 Numerical experiments

We compare our supervised tensor decomposition (**STD**) with three other supervised 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)). Figure 2 shows that **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 2b). This showcases the benefit of incorporation of multiple features. The accuracy gain in Figure 2 demonstrates the benefit of alternating algorithm – incorporation of informative modes also improves the estimation in the non-informative modes.

~~The~~

We then apply our method to brain structural connectivity networks from 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

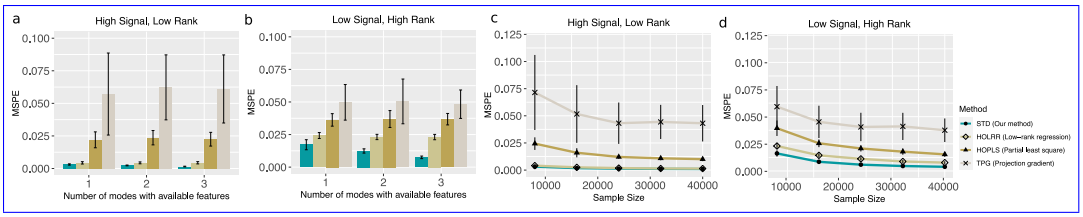

Figure 2: Comparison between different tensor methods. Panels (a) and (b) plot mean squared prediction error (MSPE) versus the number of modes with available features. Panels (c) and (d) plot MSPE versus the effective sample size d^2. We consider rank $\boldsymbol{r} = (3, 3, 3)$ (low) vs (4,5,6) (high), and signal $\alpha = 3$ (low) vs. 6 (high).

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.

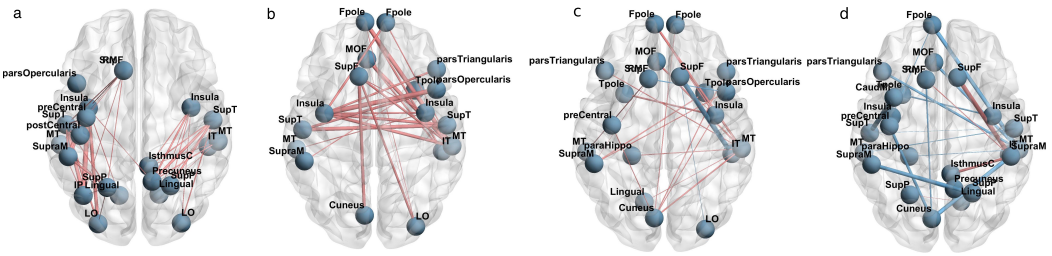

Figure 3: Top edges with large effects. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+. Red ~~edges represent positive effects and~~ (blue) edges represent positive (negative) effects. ~~The edge-width is~~ Edge-widths are proportional to the ~~magnitude~~ magnitudes of ~~the~~ effect ~~size~~sizes.

We perform the supervised tensor decomposition to the HCP data. The BIC selection suggests a rank $\boldsymbol{r} = (10, 10, 4)$ with quasi log-likelihood $\mathcal{L}_{\mathcal{Y}} = -174654.7$. ~~We utilize the 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 3 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 3a). In particular, the superior-temproal (*SupT*), middle-temporal (*MT*) and Insula are the top three popular nodes in the network. Interestingly, female brains display higher inter-hemispheric connectivity, especially in the frontal, parental and temporal lobes (Figure 3b). 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 (*Fploe*), superior-frontal (*SupF*) and Cuneus nodes. ~~The~~ Our results highlight the importance of Frontal-pole region~~is known for its importance in memory and cognition~~, and the detected decline ~~with age further highlights its biological importance~~further suggests the age effects to brain connections.

5 Conclusion

We have developed a supervised tensor decomposition method with side information on multiple modes. The empirical results demonstrate the improved interpretability and accuracy over previous approaches. Applications to the brain connection data yield conclusions with sensible interpretations, suggesting the practical utility of the proposed approach.

Broader Impact

Our supervised tensor decomposition method is widely applicable to network analysis, dyadic data analysis, spatial-temporal model, and recommendation systems. We have shown the improved predictive power and enhanced interpretability by incorporating the interactive side information in tensor decomposition method. The application to the brain connection dataset yields conclusions with sensible interpretations, suggesting the practical utility of the proposed approach. Tensor learning is a clear challenge for further research. We believe that our model enriches the research of tensor-based learning and is a powerful tool to boost scientific discoveries in various fields. We hope the work opens up new inquiry that allows more machine learning researchers to contribute to this field.

References

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.

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.

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.

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.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

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.

Rabusseau, G. and Kadri, H. (2016). Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems*, pages 1867–1875.

Srivastava, M. S., von Rosen, T., and Von Rosen, D. (2008). Models with a kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17(4):357–370.

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.