
Learning Multiple Networks via Supervised Tensor Decomposition

Anonymous Author(s)

Affiliation

Address

email

Abstract

We consider the problem of tensor decomposition with multiple side information available as interactive features. Such 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. 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 and identify the key 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 have 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 Baldwin, 2020). Side information such as people’s demographic information and friendship types are often available. In both examples, scientists are interested in identifying 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.

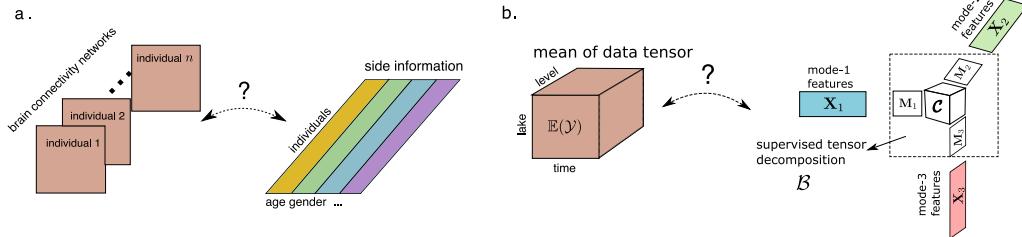


Figure 1: Examples of supervised tensor decomposition with interactive side information. (a) Network population model. (b) Spatio-temporal growth model.

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

26 Frobenius norm of deviation, leading to suboptimal predictions for binary- or count-valued response
 27 variables. A number of supervised tensor methods have been proposed (Narita et al., 2012; Zhao et al.,
 28 2012; Yu and Liu, 2016; Lock and Li, 2018). These methods often assume Gaussian distribution for
 29 the tensor entries, or impose random designs for the feature matrices, both of which are less suitable
 30 for applications of our interest. The gap between theory and practice means a great opportunity to
 31 modeling paradigms and better capture the complexity in tensor data.

32 **Our contribution.** This paper presents a general model and associated method for decomposing a
 33 data tensor whose entries are from exponential family with interactive side information. We formulate
 34 the learning task as a low-rank tensor regression problem, with tensor observation serving as the
 35 response, and the multiple side information as interactive features. We blend the modeling power of
 36 generalized linear model (GLM) and the exploratory capability of tensor dimension reduction in order
 37 to take the best out of both worlds. Our methods greatly improves the classical tensor decomposition,
 38 and we quantify the gain in prediction through numerical experiments and data applications.

39 **Notation.** We follow the tensor notation as in Kolda and Bader (2009). The multilinear mul-
 40 tiplication of a tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ by matrices $\mathbf{X}_k = [\mathbf{x}_{i_k, j_k}^{(k)}] \in \mathbb{R}^{p_k \times d_k}$ is defined
 41 as $\mathcal{Y} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\} = [\sum_{i_1, \dots, i_K} y_{i_1, \dots, i_K} x_{j_1, i_1}^{(1)} \cdots x_{j_K, i_K}^{(K)}]$, which results in an order- K
 42 (p_1, \dots, p_K)-dimensional tensor. The inner product between two tensors of equal size is defined as
 43 $\langle \mathcal{Y}, \mathcal{Y}' \rangle = \sum_{i_1, \dots, i_K} y_{i_1, \dots, i_K} y'_{i_1, \dots, i_K}$. For ease of notation, we allow basic arithmetic operators
 44 (e.g., $+$, $-$) and univariate functions $f: \mathbb{R} \rightarrow \mathbb{R}$ to be applied to tensors in an element-wise manner.
 45 Besides, let \otimes denote the Kronecker product of matrices.

46 2 Proposed models and motivating examples

47 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
 48 available on each of the K modes. Let $\mathbf{X}_k = [\mathbf{x}_{i,j}] \in \mathbb{R}^{d_k \times p_k}$ denote the feature matrix on the mode
 49 $k \in [K]$, where $x_{i,j}$ denotes the j -th feature value for the i -th tensor entity, for $(i, j) \in [d_k] \times [p_k]$,
 50 $p_k \leq d_k$. We assume that, conditional on the features \mathbf{X}_k , the entries of tensor \mathcal{Y} are independent
 51 realizations from an exponential family distribution, and the conditional mean tensor admits the form

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

52 where $\mathcal{C} \in \mathbb{R}^{r_1 \times \dots \times r_K}$ is a full-rank core tensor, $\mathbf{M}_k \in \mathbb{R}^{p_k \times r_k}$ are factor matrices consisting of
 53 orthonormal columns with $r_k \leq p_k$ for all $k \in [K]$, and $f(\cdot)$ is a known link function whose form
 54 depending on the data type of \mathcal{Y} . Common choices of link functions include identity link for Gaussian
 55 distribution, logistic link for Bernoulli distribution, and $\exp(\cdot)$ link for Poisson distribution.

56 Figure 1b provides a schematic illustration of our model. The features \mathbf{X}_k affect the distribution of
 57 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 .
 58 The core tensor \mathcal{C} collects the interaction effects between sufficient features across K modes, which
 59 links the conditional mean to the feature spaces, and thereby allows the identification of variations in
 60 the tensor data attributable to the side information. Our goal is to find \mathbf{M}_k and the corresponding \mathcal{C} ,
 61 thereby allowing us to reveal the relationship between side information \mathbf{X}_k and the observed tensor
 62 \mathcal{Y} . Note that \mathbf{M}_k and \mathcal{C} are identifiable only up to orthonormal transformations.

63 We give two examples of supervised tensor decomposition models (1) that arise in practice.

64 **Example 1** (Spatio-temporal growth model). The growth curve model (Srivastava et al., 2008) was
 65 originally proposed as an example of bilinear model for matrix data, and we extend it to higher-order
 66 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
 67 and for n time points. Suppose the sampled lakes belong to q types, with p lakes in each type. Let
 68 $\{\ell_j\}_{j \in [m]}$ denote the sampled depth levels and $\{t_k\}_{k \in [n]}$ the time points. Assume that the expected
 69 pH trend in depth is a polynomial of order at most r and that the expected trend in time is a polynomial
 70 of order s . Then, the conditional mean model for the spatio-temporal growth is a special case of our

71 model (1), where $\mathbf{X}_1 = \text{blockdiag}\{\mathbf{1}_q, \dots, \mathbf{1}_q\} \in \{0, 1\}^{d \times p}$ is the design matrix for lake types,

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

72 are the design matrices for spatial and temporal effects, respectively. The spatial-temporal mode has
73 covariates available on each of the three modes.

74 **Example 2** (Network population model). Network response model is recently developed in the
75 context of neuroimaging analysis. The goal is to study the relationship between network-valued
76 response and the individual covariates. Suppose we observe n i.i.d. observations $\{(\mathbf{Y}_i, \mathbf{x}_i) : i =$
77 $1, \dots, n\}$, where $\mathbf{Y}_i \in \{0, 1\}^{d \times d}$ is the brain connectivity network on the i -th individual, and
78 $\mathbf{x}_i \in \mathbb{R}^p$ is the individual covariate such as age, gender, cognition, etc. The network-response
79 model (Rabusseau and Kadri, 2016) has the form

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

80 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
81 tensor-response model, with covariates on the last mode of the tensor.

82 3 Estimation algorithms

83 We develop a likelihood-based procedure to estimate \mathcal{C} and \mathbf{M}_k in (1). Ignoring constants that do not
84 depend on Θ , the quasi log-likelihood of (1) is equal to

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

85 where $b(\theta) = \theta^2/2$ for Gaussian response, $b(\theta) = \exp(\theta)$ for Poisson response, and $b(\theta) =$
86 $\log(1 + \exp(\theta))$ for Bernoulli response. We propose a constrained maximum quasi-likelihood
87 estimator (M-estimator),

$$(\hat{\mathcal{C}}, \hat{\mathbf{M}}_1, \dots, \hat{\mathbf{M}}_K) = \arg \max_{(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K) \in \mathcal{P}} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K), \quad (3)$$

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

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

91 The decision variables in the objective function (3) consist of $K + 1$ blocks of variables, one for the
92 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
93 variables are known, then the optimization reduces to a simple GLM with respect to the last block
94 of variables. This observation leads to an iterative updating scheme for one block at a time while
95 keeping others fixed. A simplified version of the algorithm is described in Algorithm 1.

Change the algorithm to the complete version.

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

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: Obtain $\tilde{\mathbf{M}}_k \in \mathbb{R}^{p_k \times r_k}$ by a GLM. Orthogonalize $\tilde{\mathbf{M}}_k$ 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**

96

97 We provide the accuracy guarantee for the proposed M-estimator (3) by leveraging recent development
98 in random tensor theory and high-dimensional statistics.

99 **Theorem 3.1** (Convergence). Let $(\hat{\mathcal{C}}, \hat{M}_1, \dots, \hat{M}_K)$ be the M-estimator in (3) and $\hat{\mathcal{B}} = \hat{\mathcal{C}} \times \hat{M}_1 \times \dots \times \hat{M}_K$. Define $r_{\text{total}} = \prod_k r_k$ and $r_{\max} = \max_k r_k$. Under mild technical assumptions, there
100 exist two positive constants $C_1, C_2 > 0$, such that, with probability at least $1 - \exp(-C_1 \sum_k p_k)$,
101

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

102 where $\sin \Theta(\mathbf{M}_{k,\text{true}}, \hat{\mathbf{M}}_k) = \|\mathbf{M}_{k,\text{true}}^T \hat{\mathbf{M}}_k^\perp\|_\sigma$ is the angle distance between column spaces.

103 Theorem 3.1 implies that the estimation has a convergence rate $\mathcal{O}(d^{-(K-1)})$ in the special case when
104 tensor dimensions are equal on each of the modes, i.e., $d_k = d$ for all $k \in [K]$, and feature dimension
105 grows with tensor dimension, $p_k = \gamma d$, $\gamma \in [0, 1)$, for $k \in [K]$. The convergence of our estimation
106 becomes especially favorable as the order of tensor data increases.

107 4 Numerical experiments

108 We evaluate the empirical performance of our supervised tensor decomposition (STD) through
109 simulations. We consider order-3 tensors, where the conditional mean tensor is generated from
110 model (1). Given the generated linear predictor $\Theta = [\theta_{ijk}] = \mathcal{C} \times \{\mathbf{M}_1 \mathbf{X}_1, \mathbf{M}_2 \mathbf{X}_3, \mathbf{M}_3 \mathbf{X}_3\}$, the
111 entries in the tensor $\mathcal{Y} = [y_{ijk}]$ are drawn independently according to three probabilistic models:
112 (a) Gaussian model: $y_{ijk} \sim N(\theta_{ijk}, 1)$; (b) Poisson model: $y_{ijk} \sim \text{Poisson}(e^{\theta_{ijk}})$; (c) Bernoulli
113 model: $y_{ijk} \sim \text{Bernoulli}(e^{\theta_{ijk}} / (1 + e^{\theta_{ijk}}))$.

114 The experiment I evaluates the accuracy when covariates are available on all modes. We set
115 $\alpha = 10$, $d_k = d$, $p_k = 0.4d_k$, $r_k = r \in \{2, 4, 6\}$ and increase d from 25 to 50. Our theoretical
116 analysis suggests that $\hat{\mathcal{B}}$ has a convergence rate $\mathcal{O}(d^{-2})$ in this setting. Figure 2 plots the estimation
117 error versus the “effective sample size”, d^2 , under three different distribution models. We found that
118 the empirical MSE decreases roughly at the rate of $1/d^2$, which is consistent with our theoretical
119 ascertainment. Similar behaviors can be observed in the non-Gaussian data in Figures 2b-c.

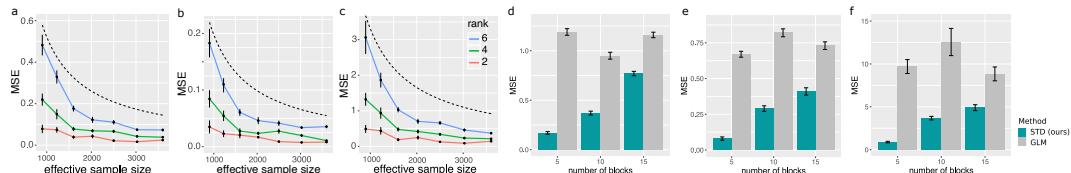


Figure 2: (a)-(c): Estimation error against effective sample size. The dashed curves correspond to $\mathcal{O}(1/d^2)$. (d)-(f): Performance comparison under stochastic block models. The x -axis represents the number of blocks in the networks. The response tensors are generated from Gaussian (a, d), Poisson (b, e) and Bernoulli (d, f) models .

120 The experiment II investigates the capability of our model in handling correlation among coefficients.
121 We mimic the scenario of brain imaging analysis. A sample of $d_3 = 50$ networks are simulated,
122 where each network measures the connections between $d_1 = d_2 = 20$ brain nodes. We simulate
123 $p = 5$ features for each of the 50 individuals. These features may represent, for example, age,
124 gender, cognitive score, etc. Recent study has suggested that brain connectivity networks often
125 exhibit community structure represented as a collection of subnetworks, and each subnetwork is
126 comprised of a set of spatially distributed brain nodes. To accommodate this structure, we utilize the
127 stochastic block model (Abbe, 2017) to generate the effect size. Specifically, we partition the nodes
128 into r blocks by assigning each node to a block with uniform probability. Edges within a same block
129 are assumed to share the same feature effects, where the effects are drawn i.i.d. from $N(0, 1)$.

130 Figure 2(d)-(f) compares the MSE of our method with a multiple-response GLM approach.
131 The multiple-response GLM is to regress the dyadic edges, one at a time, on the features,
132 and this model is repeatedly fitted for each edge. As we find in Figure 2(d)-(f), our tensor
133 regression method achieves significant error reduction in all three data types considered. The
134 outperformance is substantial in the presence of large communities; even in the less structured case
135 ($\approx 20/15 = 1.33$ nodes per block), our method still outperforms GLM. The possible reason is

136 that the multiple-response GLM approach does not account for the correlation among the edges,
 137 and suffers from overfitting. In contrast, the low-rankness in our modeling incorporates the shared
 138 information across entries.

139 **The experiment III compares STD with three other supervised tensor methods:** We compare our
 140 supervised tensor decomposition (STD) with three other supervised tensor methods: Higher-order
 141 low-rank regression (**HOLRR** Rabusseau and Kadri (2016)), Higher-order partial least square
 142 (**HOPLS** Zhao et al. (2012)) and Subsampled tensor projected gradient (**TPG** Yu and Liu (2016)).
 143 Figure 3 shows that **STD** outperforms others, especially in the low-signal, high-rank setting. As
 144 the number of informative modes (i.e., modes with available features) increases, the **STD** exhibits
 145 a substantial reduction in error whereas others remain unchanged (Figure 3b). This showcases the
 146 benefit of incorporation of multiple features. The accuracy gain in Figure 3 demonstrates the benefit
 147 of alternating algorithm – incorporation of informative modes also improves the estimation in the
 148 non-informative modes.

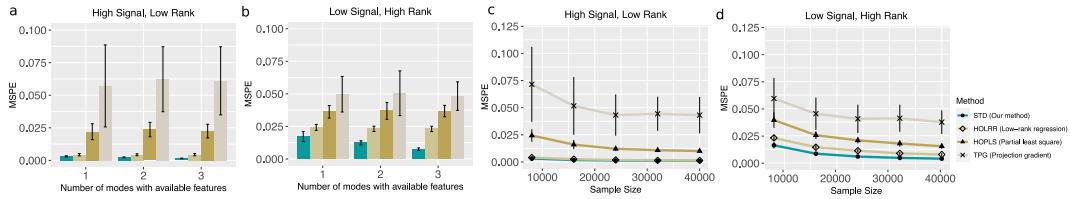


Figure 3: Comparison between different tensor methods. Panels (a) and (b) plot mean squared prediction error (MSPE) versus the number of modes with available features. Panels (c) and (d) plot MSPE versus the effective sample size d^2 . We consider rank $r = (3, 3, 3)$ (low) vs $(4, 5, 6)$ (high), and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

149 We then apply our method to brain structural connectivity networks from Human Connectome Project
 150 (HCP) (Geddes, 2016). The dataset consists of 136 brain structural networks, one for each individual.
 151 Each brain network is represented as a 68-by-68 binary matrix, where the entries encode the presence
 152 or absence of fiber connections between the 68 brain regions. We consider four individual features:
 153 gender (65 females vs. 71 males), age 22-25 ($n = 35$), age 26-30 ($n = 58$), and age 31+ ($n = 43$).
 154 The goal is to identify the connection edges that are affected by individual features.

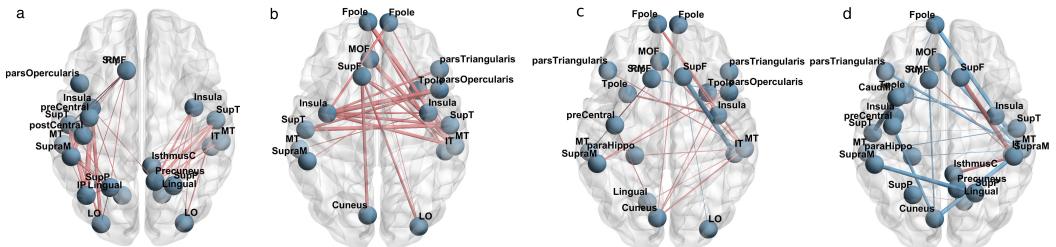


Figure 4: Top edges with large effects. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+. Red (blue) edges represent positive (negative) effects. Edge-widths are proportional to the magnitudes of effect sizes.

155 We perform the supervised tensor decomposition to the HCP data. The BIC selection suggests a
 156 rank $r = (10, 10, 4)$ with quasi log-likelihood $\mathcal{L}_y = -174654.7$. Figure 4 shows the top edges with
 157 high effect size, overlaid on the Desikan atlas brain template (Desikan et al., 2006). We find that the
 158 global connection exhibits clear spatial separation, and that the nodes within each hemisphere are
 159 more densely connected with each other (Figure 4a). In particular, the superior-temporal (*SupT*),
 160 middle-temporal (*MT*) and Insula are the top three popular nodes in the network. Interestingly, female
 161 brains display higher inter-hemispheric connectivity, especially in the frontal, parietal and temporal
 162 lobes (Figure 4b). This is in agreement with a recent study showing that female brains are optimized
 163 for inter-hemispheric communication (Ingalhalikar et al., 2014). We find several edges with declined
 164 connection in the group Age 31+. Those edges involve Frontal-pole (*Fpole*), superior-frontal (*SupF*)
 165 and Cuneus nodes. Our results highlight the importance of Frontal-pole region, and the detected
 166 decline further suggests the age effects to brain connections.

¹⁶⁷ **5 Conclusion**

¹⁶⁸ We have developed a supervised tensor decomposition method with side information on multiple
¹⁶⁹ modes. The empirical results demonstrate the improved interpretability and accuracy over previous
¹⁷⁰ approaches. Applications to the brain connection data yield conclusions with sensible interpretations,
¹⁷¹ suggesting the practical utility of the proposed approach.

172 **Broader Impact**

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

181 **References**

- 182 Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The
183 Journal of Machine Learning Research*, 18(1):6446–6531.
- 184 Berthet, Q. and Baldin, N. (2020). Statistical and computational rates in graph logistic regression.
185 *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*,
186 108:2719–2730.
- 187 Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner,
188 R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system
189 for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.
190 *Neuroimage*, 31(3):968–980.
- 191 Geddes, L. (2016). Human brain mapped in unprecedented detail. *Nature*.
- 192 Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Hakonarson,
193 H., Gur, R. E., Gur, R. C., and Verma, R. (2014). Sex differences in the structural connectome of
194 the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828.
- 195 Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*,
196 51(3):455–500.
- 197 Lock, E. F. and Li, G. (2018). Supervised multiway factorization. *Electronic journal of statistics*,
198 12(1):1150.
- 199 Narita, A., Hayashi, K., Tomioka, R., and Kashima, H. (2012). Tensor factorization using auxiliary
200 information. *Data Mining and Knowledge Discovery*, 25(2):298–324.
- 201 Rabusseau, G. and Kadri, H. (2016). Low-rank regression with tensor responses. In *Advances in
202 Neural Information Processing Systems*, pages 1867–1875.
- 203 Srivastava, M. S., von Rosen, T., and Von Rosen, D. (2008). Models with a kronecker product
204 covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17(4):357–370.
- 205 Yu, R. and Liu, Y. (2016). Learning from multiway data: Simple and efficient tensor regression. In
206 *International Conference on Machine Learning*, pages 373–381.
- 207 Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Nagasaka, Y., Fujii, N., Zhang, L., and Cichocki,
208 A. (2012). Higher order partial least squares (HOPLS): a generalized multilinear regression method.
209 *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1660–1673.
- 210 Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data
211 analysis. *Journal of the American Statistical Association*, 108(502):540–552.