
Learning Multiple Networks via Supervised Tensor Decomposition

Anonymous Author(s)

Affiliation

Address

email

Abstract

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. We identify the key global brain 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.

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; 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.

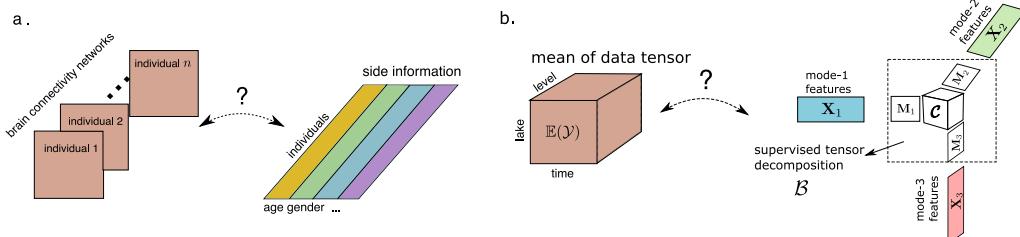


Figure 1: Examples of supervised tensor decomposition with interactive side information. (a) Network population model. (b) Spatio-temporal growth model.

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

33 We present a general model and associated method for decomposing a data tensor whose entries
 34 are from exponential family with interactive side information. We formulate the learning task as
 35 a structured regression problem, with tensor observation serving as the response, and the multiple
 36 side information as interactive features. We leverage generalized linear model (GLM) to allow
 37 heteroscedacity due to the mean-variance relationship in the non-Gaussian data. The low-rank
 38 structure on the conditional mean tensor effectively mitigates the curse of high dimensionality. Our
 39 proposal blends the modeling power of GLM and the exploratory capability of tensor dimension
 40 reduction in order to take the best out of both worlds.

41 2 Method

42 2.1 Preliminary

43 We introduce the basic tensor properties used in the paper. We use lower-case letters (e.g., a, b, c)
 44 for scalars and vectors, upper-case boldface letters (e.g., $\mathbf{A}, \mathbf{B}, \mathbf{C}$) for matrices, and calligraphy
 45 letters (e.g., $\mathcal{A}, \mathcal{B}, \mathcal{C}$) for tensors of order three or greater. Let $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!] \in \mathbb{R}^{d_1 \times \dots \times d_K}$
 46 denote an order- K (d_1, \dots, d_K)-dimensional tensor. The multilinear multiplication of a tensor
 47 $\mathcal{Y} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ by matrices $\mathbf{X}_k = [\![x_{i_k, j_k}^{(k)}]\!] \in \mathbb{R}^{p_k \times d_k}$ is defined as $\mathcal{Y} \times \{\mathbf{X}_1, \dots, \mathbf{X}_K\} =$
 48 $[\![\sum_{i_1, \dots, i_K} y_{i_1, \dots, i_K} x_{j_1, i_1}^{(1)} \cdots x_{j_K, i_K}^{(K)}]\!]$, which results in an order- K (p_1, \dots, p_K)-dimensional ten-
 49 sor. 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
 50 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
 51 defined as $\|\mathcal{Y}\|_F = \langle \mathcal{Y}, \mathcal{Y} \rangle^{1/2}$, and the maximum norm is defined $\|\mathcal{Y}\|_\infty = \max_{i_1, \dots, i_K} |y_{i_1, \dots, i_K}|$.
 52 We let \mathbf{I}_d denote the $d \times d$ identity matrix and $[d]$ denote the d -set $\{1, \dots, d\}$.

53 2.2 General Model

54 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
 55 available on each of the K modes. Let $\mathbf{X}_k = [\![x_{i,j}]\!] \in \mathbb{R}^{d_k \times p_k}$ denote the feature matrix on the mode
 56 $k \in [K]$, where $x_{i,j}$ denotes the j -th feature value for the i -th tensor entity, for $(i, j) \in [d_k] \times [p_k]$,
 57 $p_k \leq d_k$. Assume that, conditional on the features \mathbf{X}_k , the entries of tensor \mathcal{Y} are independent
 58 realizations from an exponential family distribution, and the conditional mean tensor admits the form

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

59 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,
 60 $f(\cdot)$ is a known link function whose form depending on the data type of \mathcal{Y} . The choice of link
 61 function is based on the assumed distribution family of tensor entries.

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

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

65 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
 66 orthonormal columns, where $r_k \leq p_k$ for all $k \in [K]$. By the definition of multilinear rank, model
 67 equations (1) and (2) imply the low-rankness $\mathbf{r} = (r_1, \dots, r_K)$ of the conditional mean tensor under
 68 the link function. We now reach our final model for supervised tensor decomposition,

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

69 where the parameters of interest are \mathbf{M}_k and \mathcal{C} . Note that model (3) assumes a fixed, known rank
70 $\mathbf{r} = (r_1, \dots, r_K)$. Figure 1b provides a schematic illustration of our model. The features \mathbf{X}_k affect
71 the distribution of tensor entries in \mathcal{Y} through the form $\mathbf{X}_k \mathbf{M}_k$, which are r_k linear combinations of
72 features on mode k . We call $\mathbf{X}_k \mathbf{M}_k$ the “supervised tensor factors” or “sufficient features” (Adragni
73 and Cook, 2009), and call \mathbf{M}_k the “dimension reduction matrix.” The core tensor \mathcal{C} collects the
74 interaction effects between sufficient features across K modes, which links the conditional mean to
75 the feature spaces, and thereby allows the identification of variations in the tensor data attributable
76 to the side information. Our goal is to find \mathbf{M}_k and the corresponding \mathcal{C} . Note that \mathbf{M}_k and \mathcal{C} are
77 identifiable only up to orthonormal transformations.

78 3 Estimation

79 3.1 Rank-constrained M-estimator

80 We adopt the exponential family as a flexible framework for different data types. In a classical
81 generalized linear model with a scalar response y and feature \mathbf{x} , the density is expressed as

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

82 where $b(\cdot)$ is a known function depending on the data types, θ is the linear predictor, $\phi > 0$ is the
83 dispersion parameter, and $c(\cdot)$ is a known normalizing function. In our context, we model the tensor
84 entries y_{i_1, \dots, i_K} , conditional on θ_{i_1, \dots, i_K} , as independent draws from an exponential family. Ignoring
85 constants that do not depend on Θ , the quasi log-likelihood of (3) is equal to Bregman distance
86 between \mathcal{Y} and $b'(\Theta)$:

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

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

$$(\hat{\mathcal{C}}, \hat{\mathbf{M}}_1, \dots, \hat{\mathbf{M}}_k) = \arg \max_{(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K) \in \mathcal{P}} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}_1, \dots, \mathbf{M}_K), \quad (4)$$

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

91 3.2 Alternating optimization

92 The decision variables in the objective function (4) consist of $K + 1$ blocks of variables, one for the
93 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
94 variables are known, then the optimization reduces to a simple GLM with respect to the last block
95 of variables. This observation leads to an iterative updating scheme for one block at a time while
96 keeping others fixed. A simplified version of the algorithm is described in Algorithm 1.

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**

97 3.3 Statistical properties

98 In modern applications, the tensor data and features are often large-scale. We are particularly
99 interested in the high-dimensional regime in which both d_k and p_k diverge; i.e. $d_k \rightarrow \infty$ and

100 $p_k \rightarrow \infty$, while $p_k/d_k \rightarrow \gamma_k \in [0, 1)$. As the size of problem grows, and so does the number of
101 unknown parameters. The classical MLE theory does not directly apply. We leverage the recent
102 development in random tensor theory and high-dimensional statistics to establish the error bounds.

103 **Theorem 3.1** (Statistical convergence). Consider a data tensor generated from model (3). Let
104 $(\hat{\mathcal{C}}, \hat{M}_1, \dots, \hat{M}_K)$ be the M-estimator in (4) and $\hat{\mathcal{B}} = \hat{\mathcal{C}} \times \hat{M}_1 \times \dots \times \hat{M}_K$. Define $r_{\text{total}} =$
105 $\prod_k r_k$ and $r_{\max} = \max_k r_k$. Under mild technical assumptions, there exist two positive constants
106 $C_1 = C_1(\alpha, K), C_2 = C_2(\alpha, K) > 0$ independent of dimensions $\{d_k\}$ and $\{p_k\}$, such that, with
107 probability at least $1 - \exp(-C_1 \sum_k p_k)$,

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

108 Furthermore, if the unfolded core tensor has non-degenerate singular values at mode $k \in [K]$, i.e.,
109 $\sigma_{\min}(\text{Unfold}_k(\mathcal{C}_{\text{true}})) \geq c > 0$ for some constant c , then

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

110 where $\sin \Theta(\mathcal{M}_{k,\text{true}}, \hat{\mathcal{M}}_k) = \|\mathcal{M}_{k,\text{true}}^T \hat{\mathcal{M}}_k^\perp\|_\sigma = \max \left\{ \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} : \mathbf{x} \in \text{Span}(\mathcal{M}_{k,\text{true}}), \mathbf{y} \in \text{Span}(\hat{\mathcal{M}}_k^\perp) \right\}$
111 is the angle distance to assess the accuracy in estimating the column space, $\text{Span}(\mathcal{M}_k)$.

112 4 Real data analysis

113 The Human Connectome Project (HCP) aims to build a network map that characterizes the anatomical
114 and functional connectivity within healthy human brains (Geddes, 2016). We follow the preprocessing
115 procedure as in Zhang et al. (2018) and parcellate the brain into 68 regions of interest (Desikan et al.,
116 2006). The dataset consists of 136 brain structural networks, one for each individual. Each brain
117 network is represented as a 68-by-68 binary matrix, where the entries encode the presence or absence
118 of fiber connections between the 68 brain regions. We consider four individual features: gender (65
119 females vs. 71 males), age 22-25 ($n = 35$), age 26-30 ($n = 58$), and age 31+ ($n = 43$). The goal is
120 to identify the connection edges that are affected by individual features.

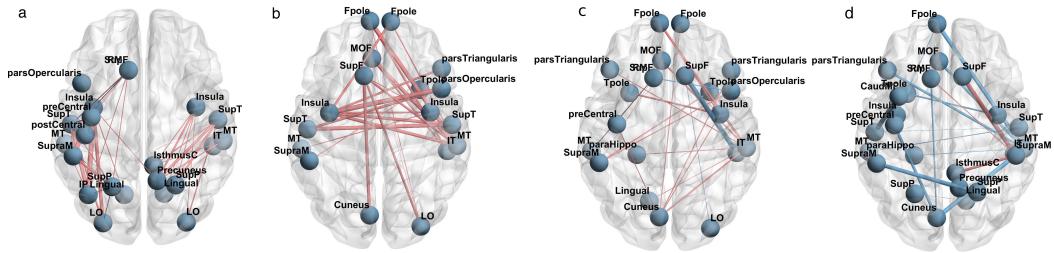


Figure 2: Top edges with large effects. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+. Red edges represent positive effects and blue edges represent negative effects. The edge-width is proportional to the magnitude of the effect size.

121 We perform the supervised tensor decomposition to the HCP data. The BIC selection suggests
122 a rank $r = (10, 10, 4)$ with quasi log-likelihood $\mathcal{L}_Y = -174654.7$. We utilize the sum-to-zero
123 contrasts in coding the feature effects, and depict only the top 3% edges whose connections are
124 non-constant across the sample. Figure 2 shows the top edges with high effect size, overlaid on
125 the Desikan atlas brain template (Desikan et al., 2006). We find that the global connection exhibits
126 clear spatial separation, and that the nodes within each hemisphere are more densely connected
127 with each other (Figure 2a). In particular, the superior-temporal ($SupT$), middle-temporal (MT) and
128 Insula are the top three popular nodes in the network. Interestingly, female brains display higher
129 inter-hemispheric connectivity, especially in the frontal, parietal and temporal lobes (Figure 2b). This
130 is in agreement with a recent study showing that female brains are optimized for inter-hemispheric
131 communication (Ingalhalikar et al., 2014). We find several edges with declined connection in the
132 group Age 31+. Those edges involve Frontal-pole ($Fpole$), superior-frontal ($SupF$) and Cuneus nodes.
133 The Frontal-pole region is known for its importance in memory and cognition, and the detected
134 decline with age further highlights its biological importance.

135 **Border Impact**

136 **References**

- 137 Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression.
138 *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering*
139 *Sciences*, 367(1906):4385–4405.
- 140 Berthet, Q. and Baldwin, N. (2020). Statistical and computational rates in graph logistic regression.
141 *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*,
142 108:2719–2730.
- 143 Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner,
144 R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system
145 for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.
146 *Neuroimage*, 31(3):968–980.
- 147 Geddes, L. (2016). Human brain mapped in unprecedented detail. *Nature*.
- 148 Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical*
149 *Association*, 100(469):286–295.
- 150 Ingallhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Hakonarson,
151 H., Gur, R. E., Gur, R. C., and Verma, R. (2014). Sex differences in the structural connectome of
152 the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828.
- 153 Lock, E. F. and Li, G. (2018). Supervised multiway factorization. *Electronic journal of statistics*,
154 12(1):1150.
- 155 Narita, A., Hayashi, K., Tomioka, R., and Kashima, H. (2012). Tensor factorization using auxiliary
156 information. *Data Mining and Knowledge Discovery*, 25(2):298–324.
- 157 Yu, R. and Liu, Y. (2016). Learning from multiway data: Simple and efficient tensor regression. In
158 *International Conference on Machine Learning*, pages 373–381.
- 159 Zhang, Z., Descoteaux, M., Zhang, J., Girard, G., Chamberland, M., Dunson, D., Srivastava, A., and
160 Zhu, H. (2018). Mapping population-based structural connectomes. *NeuroImage*, 172:130–145.
- 161 Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Nagasaka, Y., Fujii, N., Zhang, L., and Cichocki,
162 A. (2012). Higher order partial least squares (HOPLS): a generalized multilinear regression method.
163 *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1660–1673.
- 164 Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data
165 analysis. *Journal of the American Statistical Association*, 108(502):540–552.