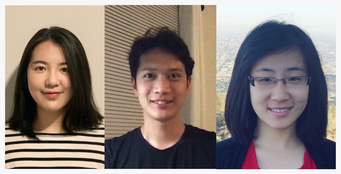

Learning Multiple Networks via Supervised Tensor Decomposition

Jiaxin Hu, Chanwoo Lee, and Miaoyan Wang

{jhu267, chanwoo.lee, miaoyan.wang}@wisc.edu

Department of Statistics, University of Wisconsin-Madison

MOTIVATION

- Scientific studies (e.g. neuroimaging, social network analysis) often collect the tensor observations with side information.
- Tensor observations may consist of non-Gaussian measurements (e.g. binary, count data).

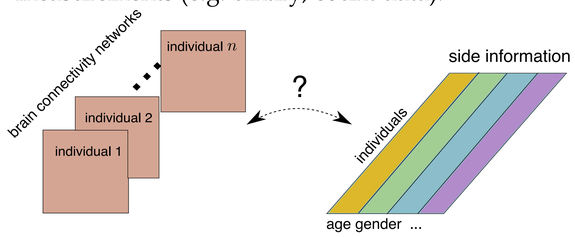

Figure 1. Brain connectivity networks (binary adjacency matrices) with individual side information.

Our Goal: Identify the structural variation in the data tensor affected by side information.

ALGORITHM

The alternating optimization algorithm updates one block of parameters at a time while keeping other fixed.

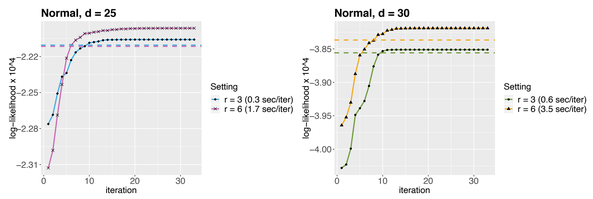

Figure 2. Likelihood trajectory for normal model. Dashed lines are likelihood with true parameters.

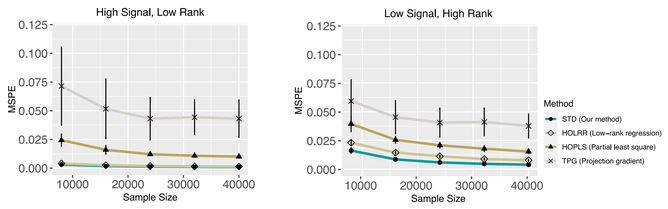

Figure 3. Mean squared prediction error (MSPE) comparison between different tensor methods. We consider rank $\boldsymbol{r} = (3,3,3)$ (low) vs (4,5,6) (high) and signal $\alpha = 3$ (low) vs 6 (high).

MODEL

Let $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote the tensor observation and $\boldsymbol{X}_i \in \mathbb{R}^{d_i \times p_i}$ denote the side information encoded as multiple feature matrices, for $i = 1, ..., K$. We propose a supervised tensor decomposition

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, ..., \boldsymbol{X}_K) = f(\mathcal{C} \times_1 \boldsymbol{X}_1\boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{X}_K\boldsymbol{M}_K),$$

where $f(\cdot)$ is the known link function depending on the data type, $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is the unknown core tensor, and $\boldsymbol{M}_i \in \mathbb{R}^{p_i \times r_i}$ are unknown factor matrices with orthonormal columns.

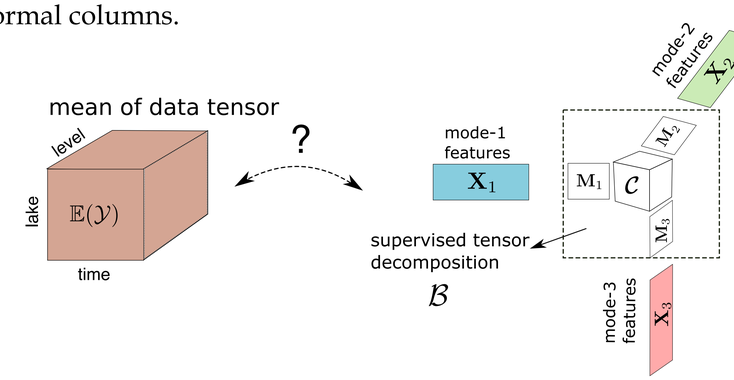

Figure 4. Supervised tensor decomposition for an order-3 tensor with multiple side information.

The features $\boldsymbol{X}_i$ affect the distribution of tensor entries in $\mathcal{Y}$ through ``supervised tensor factors" $\boldsymbol{X}_i\boldsymbol{M}_i$. We call $\boldsymbol{M}_i$ the ``dimension reduction matrix", and $\mathcal{C}$ collects the interaction effects between sufficient features.

We propose a likelihood-based estimator

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

where $\mathcal{L}_{\mathcal{Y}}(\cdot)$ is the quasi log-likelihood function

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_1, ..., \boldsymbol{M}_K) = \langle \mathcal{Y}, \Theta \rangle + \sum_{i_1,...,i_K} b\left(\Theta_{i_1,...,i_K}\right),$$

$$\text{with} \quad \Theta = \mathcal{C} \times_1 \boldsymbol{X}_1\boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{X}_K\boldsymbol{M}_K,$$

$b(\cdot)$ is a known function depending on the data type in $\mathcal{Y}$, and $\mathcal{P}$ is the parameter space

$$\mathcal{P} = \left\{(\mathcal{C}, \boldsymbol{M}_1, ..., \boldsymbol{M}_K) | \boldsymbol{M}_k^T\boldsymbol{M}_k = \mathbf{I}_{r_k}, \|\Theta(\mathcal{C}, \boldsymbol{M}_1, ..., \boldsymbol{M}_K)\|_{\max} \le \alpha\right\},$$

with the positive constant α.

STATISTICAL GUARANTEES

Theorem. Suppose the singular values of $\boldsymbol{X}_i, \boldsymbol{M}_i$ are bounded by constants and $\|\boldsymbol{X}_i\|_F \asymp \Omega(\sqrt{d_i})$. Under mild technical conditions, there exist constants $C_1, C_2 > 0$, such that, with high probability at least $1 - \exp(-C_1 \sum_k p_k)$,

$$\sin^2\Theta(\boldsymbol{M}_i, \hat{\boldsymbol{M}}_i) \le \frac{C_2 \prod_k r_k}{\max_k r_k \sigma_{\min}^2(\boldsymbol{M}_i)} \frac{\sum_k p_k}{\prod_k d_k}, \quad \text{for all } i = 1, ..., K,$$

and

$$\left\|\mathcal{B} - \hat{\mathcal{B}}\right\|_F^2 \le \frac{C_2 \prod_k r_k}{\max_k r_k} \frac{\sum_k p_k}{\prod_k d_k},$$

where $\mathcal{B} = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$, and C_1, C_2 are constants independent of $\{p_k\}$ and $\{d_k\}$.

APPLICATION

Our method identifies the global connectivity pattern as well as local regions associated with age and gender in the Human Connectome Project (HCP) data.

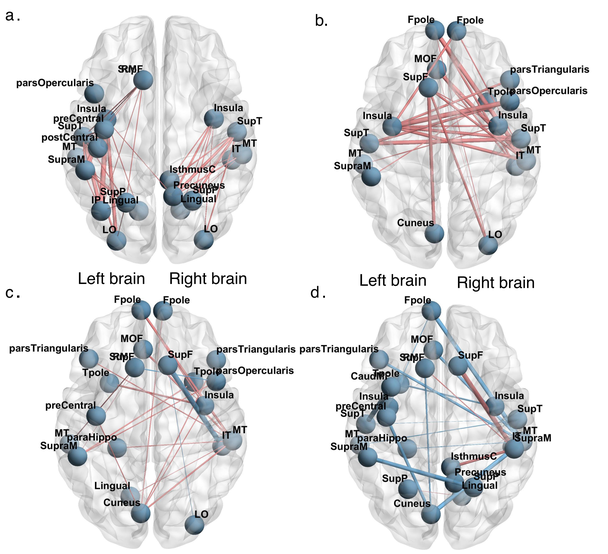

Figure 5. Top edges with large effects. (a) Global effect; (b) Female effect; (c) Age 22-25; (d) Age 31+. Red (blue) edges represent positive (negative) effects. Edge-widths are proportional to the magnitude of effect sizes.