t-BNE: Tensor-based Brain Network Embedding

Bokai Cao* Lifang He*† Xiaokai Wei* Mengqi Xing‡ Philip S. Yu*§
Heide Klumpp¶ Alex D. Leow‡¶

Abstract

Brain network embedding is the process of converting brain network data to discriminative representations of subjects, so that patients with brain disorders and normal controls can be easily separated. Computer-aided diagnosis based on such representations is potentially transformative for investigating disease mechanisms and for informing therapeutic interventions. However, existing methods either limit themselves to extracting graph-theoretical measures and subgraph patterns, or fail to incorporate brain network properties and domain knowledge in medical science. In this paper, we propose t-BNE, a novel Brain Network Embedding model based on constrained tensor factorization. t-BNE incorporates 1) symmetric property of brain networks, 2) side information guidance to obtain representations consistent with auxiliary measures, 3) orthogonal constraint to make the latent factors distinct with each other, and 4) classifier learning procedure to introduce supervision from labeled data. The Alternating Direction Method of Multipliers (ADMM) framework is utilized to solve the optimization objective. We evaluate t-BNE on three EEG brain network datasets. Experimental results illustrate the superior performance of the proposed model on graph classification tasks with significant improvement 20.51%, 6.38% and 12.85%, respectively. Furthermore, the derived factors are visualized which could be informative for investigating disease mechanisms under different emotion regulation tasks.

1 Introduction

Recent years have witnessed an increasing amount of brain network data in a variety of modalities, *e.g.*, functional magnetic resonance imaging (fMRI), diffusion tensor imaging (DTI) and electroencephalogram (EEG). These data are inherently represented as a set of vertices and edges, instead of feature vectors as traditional data, where vertices correspond to regions of interest (ROIs) in the brain and edges represent the connectivity strength or correlation between brain regions. Both structural and functional brain networks have been increasingly studied in recent years [3, 12], with potential applications to the early detection of brain diseases [30]. For example, functional brain networks provide a graph-theoretical viewpoint to investigate the collective pattern of functional activity across all brain regions, and have been shown to be abnormal in neuropsychiatric disorders [19].

The complex structures and the lack of vector representations within these brain network data raise research challenges for data mining. A straightforward solution that has been extensively explored is to first derive features from brain networks so that conventional machine learning algorithms could be applied. In general, two types of features are usually extracted: (1) graph-theoretical measures [32, 19] and (2) subgraph patterns [22, 9]. However, the expressiveness of the derived features is limited to the predefined formulations. To explore a larger space of potentially informative features to represent brain networks, it is critical to learn latent representations from the brain network data through factorization techniques.

In this paper, we propose t-BNE, a brain network embedding method based on constrained tensor factorization. The contributions of this work are threefold:

- The brain network embedding problem is modeled as partially symmetric tensor factorization which is suitable for inherently undirected graphs, *e.g.*, EEG brain networks.
- The self-report data is incorporated as guidance in the tensor factorization procedure to learn latent factors that are consistent with the side information. Moreover, orthogonal constraint is introduced to obtain distinct factors.
- The representation learning and classifier training are blended into a unified optimization problem, which allows the classifier parameters to interact with the

*Department of Computer Science, University of Illinois, Chicago. {caobokai, xwei2, psyu}@uic.edu
†Corresponding author. Institute for Computer Vision, Shenzhen University, Shenzhen. lifanghescut@gmail.com
‡Department of Bioengineering, University of Illinois, Chicago. mxing3@uic.edu, alexfeuillet@gmail.com
§Institute for Data Science, Tsinghua University, Beijing.
¶Department of Psychiatry, University of Illinois, Chicago. hklumpp@psych.uic.edu

original brain network data via latent factors, and the supervision information to be introduced to the representation learning process, so that discriminative representations can be obtained. It is demonstrated to be equivalent to partially coupled matrix and tensor factorization.

We evaluate t-BNE on three EEG datasets. Experimental results illustrate the superior performance of the proposed model on graph classification with significant improvement 20.51%, 6.38% and 12.85%, respectively. Furthermore, the derived factors are visualized which could be informative for investigating disease mechanisms under different emotion regulation tasks.

2 Problem Formulation

Let $\mathcal{D} = \{G_1, \cdots, G_n\}$ denote a graph dataset of brain networks where n is the number of subjects. All graphs in the dataset share a given set of vertices V, which corresponds to a specific brain parcellation scheme. Suppose the brain is parcellated via an atlas into m regions. A brain network G_i can be represented by an adjacency matrix $\mathbf{A}_i \in \mathbb{R}^{m \times m}$.

Definition 1. (Graph) *A graph is represented as $G = (V, E)$, where $V = \{v_1, \cdots, v_m\}$ is the set of vertices, $E \subseteq V \times V$ is the set of edges.*

We assume that the first l subjects within $\mathcal{D}$ are labeled and $\mathbf{Y} \in \mathbb{R}^{l \times c}$ is the class label matrix where c is the number of class labels. Each subject belongs to only one class where $\mathbf{Y}(i,j) = 1$ if G_i belongs to the j-th class, otherwise $\mathbf{Y}(i,j) = 0$. For convenience, we also denote the labeled graph dataset by $\mathcal{D}_l = \{G_1, \cdots, G_l\}$, and the unlabeled graph dataset as $\mathcal{D}_u = \{G_{l+1}, \cdots, G_n\}$, $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$.

The research problem studied in this paper, *i.e.*, brain network embedding, can be described as to learn latent representations of all brain networks, say, $\mathbf{S} \in \mathbb{R}^{n \times k}$ where each brain network G_i is represented in a k-dimensional space with coordinates $\mathbf{S}(i,:)$. It is desirable to let the latent representations be discriminative so that brain networks with different labels can be easily separated. Formally, given $\{\mathbf{S}(i,:) \mid i \in \mathcal{D}_l\}$, the labels of $\{\mathbf{S}(i,:) \mid i \in \mathcal{D}_u\}$ can be correctly classified.

In order to train an effective graph classifier for diagnostic classification, how to learn an informative representation of each subject from the brain network data? Learning such representations is a non-trivial task due to the following problems:

(P1) Brain networks are undirected graphs in most cases, thus, their corresponding connectivity matrices are symmetric. How can we carefully deal with such graph property in the representation learning process?

(P2) In medical experiments, we usually only have a limited number of subjects. However, in addition to the brain network data, other cognitive measures might be documented. How can we leverage such side information sources to facilitate the representation learning process?

(P3) The ultimate goal is to train an effective graph classifier based on the learned representations. If we can incorporate the classifier training procedure into the representation learning process, the classifier parameters will indirectly interact with the original brain network data via latent factors, and the supervision information will be introduced to the representation learning process, so that discriminative representations can be obtained. How can we effectively fuse these two procedures together?

3 Brain Network Embedding

3.1 Tensor Modeling We resort to tensor techniques for modeling the brain network data in this paper. We first address the problem (P1) discussed in Section 2 by stacking brain networks of n subjects, *i.e.*, $\{\mathbf{A}_i\}_{i=1}^n$, as a partially symmetric tensor $\mathcal{X} \in \mathbb{R}^{m \times m \times n}$.

Definition 2. (Partially Symmetric Tensor) *A mth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ is a rank-one partially symmetric tensor if it is partially symmetric on modes $i_1, ..., i_j \in \{1, ..., m\}$, and can be written as the tensor product of m vectors, i.e.,*

$$\mathcal{X} = \mathbf{x}^{(1)} \circ \cdots \circ \mathbf{x}^{(m)} \tag{3.1}$$

where $\mathbf{x}^{(i_1)} = \cdots = \mathbf{x}^{(i_j)}$.

We assume that the third-order tensor $\mathcal{X}$ can be decomposed into k factors

$$\mathcal{X} = \mathcal{C} \times_1 \mathbf{B} \times_2 \mathbf{B} \times_3 \mathbf{S} \tag{3.2}$$

where $\mathbf{B} \in \mathbb{R}^{m \times k}$ is the factor matrix for vertices, $\mathbf{S} \in \mathbb{R}^{n \times k}$ is the factor matrix for subjects, and $\mathcal{C} \in \mathbb{R}^{k \times k \times k}$ is the identity tensor, *i.e.*, $\mathcal{C}(i_1, i_2, i_3) = \delta(i_1 = i_2 = i_3)$. Basically, Eq. (3.2) is a CANDECOMP/PARAFAC (CP) factorization [20] as shown in Figure 1.

For brain network analysis, auxiliary statistics are usually associated with subjects, *e.g.*, demographic information and cognitive measures [7, 36]. As noted in previous work [26], the similarity/distance between subjects in the latent space should be consistent with that in the space of auxiliary features. That is to say, if two subjects are similar in the auxiliary space, they should also be close to each other in the latent space (*i.e.*, a small distance between latent factors). Therefore, we address the problem (P2) discussed in

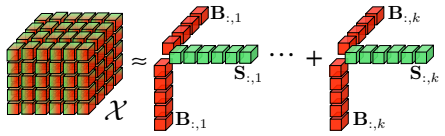


Figure 1: CP factorization. The third-order partially symmetric tensor $\mathcal{X}$ is approximated by k rank-one tensors. The f-th factor tensor is the tensor product of three vectors, *i.e.*, $\mathbf{B}_{:,f} \circ \mathbf{B}_{:,f} \circ \mathbf{S}_{:,f}$.

Section 2 by defining the objective as to minimize the distance between latent factors of each pair of similar subjects based upon the side information. It can be mathematically formulated as follows

$$\min_{\mathbf{S}} \sum_{i,j=1}^{n} \|\mathbf{S}(i,:) - \mathbf{S}(j,:)\|_F^2 \mathbf{Z}(i,j) \tag{3.3}$$

where $\mathbf{Z}$ is the kernel matrix whose entries $\mathbf{Z}(i,j)$ represent the similarity between G_i and G_j in the space of auxiliary features. A linear kernel is used in this study. In this way, the side information about subjects is effectively used as guidance to discover meaningful latent factors. Note that we can rewrite Eq. (3.3) as

$$\min_{\mathbf{S}} \ \mathrm{tr}(\mathbf{S}^{\mathrm{T}} \mathbf{L}_{\mathbf{Z}} \mathbf{S}) \tag{3.4}$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix, $\mathbf{L}_{\mathbf{Z}}$ is the Laplacian matrix induced from the similarity matrix $\mathbf{Z}$, *i.e.*, $\mathbf{L}_{\mathbf{Z}} = \mathbf{D}_{\mathbf{Z}} - \mathbf{Z}$, and $\mathbf{D}_{\mathbf{Z}}$ is the diagonal matrix whose entries are column sums of $\mathbf{Z}$, *i.e.*, $\mathbf{D}_{\mathbf{Z}}(i,i) = \sum_j \mathbf{Z}(i,j)$.

Moreover, it is desirable to discover distinct latent factors to obtain more concise and interpretable results, which can be achieved via the orthogonal constraint

$$\mathbf{S}^{\mathrm{T}} \mathbf{S} = \mathbf{I} \tag{3.5}$$

One of the targets is diagnostic classification based on the brain network data. We assume that there is a mapping matrix $\mathbf{W} \in \mathbb{R}^{k \times c}$ which assigns subjects with labels based on the subject factor matrix $\mathbf{S}$. The relation can be captured by the ridge regression problem [18]

$$\min_{\mathbf{W}} \ \|\mathbf{DSW} - \mathbf{Y}\|_F^2 + \gamma \|\mathbf{W}\|_F^2 \tag{3.6}$$

where $\mathbf{D} = [\mathbf{I}^{l \times l}, \ \mathbf{0}^{l \times (n-l)}] \in \mathbb{R}^{l \times n}$ and $\|\mathbf{W}\|_F^2$ controls the capacity of $\mathbf{W}$ with the parameter γ controlling its influence.

An intuitive idea is to use latent factors of subjects as features and then train a classifier on them in a serial two-step manner. However, the advantage is established in [6] of directly searching for classification-relevant structure in the original data, rather than solving the supervised and unsupervised problems independently. To address the problem (P3) discussed in Section 2, we propose to incorporate the classifier learning process (*i.e.*, $\mathbf{W}$) into the framework of learning latent feature representations of subjects (*i.e.*, $\mathbf{S}$). In this manner, the weight matrix $\mathbf{W}$ and the feature matrix $\mathbf{S}$ can interact with each other in the same learning framework. Note that it is equivalent to partially coupled matrix and tensor factorization [2], considering $\mathcal{X}$ and $\mathbf{Y}$ are coupled in part of the subject mode.

In summary, the proposed brain network embedding framework can be mathematically formulated as solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{B},\mathbf{S},\mathbf{W}} \ & \underbrace{\|\mathcal{X} - \mathcal{C} \times_1 \mathbf{B} \times_2 \mathbf{B} \times_3 \mathbf{S}\|_F^2}_{\text{factorization error}} \\ & + \alpha \underbrace{\mathrm{tr}(\mathbf{S}^{\mathrm{T}} \mathbf{L}_{\mathbf{Z}} \mathbf{S})}_{\text{guidance}} + \beta \underbrace{\|\mathbf{DSW} - \mathbf{Y}\|_F^2}_{\text{classification loss}} + \gamma \underbrace{\|\mathbf{W}\|_F^2}_{\text{reg.}} \\ \text{s.t.} \ & \underbrace{\mathbf{S}^{\mathrm{T}} \mathbf{S} = \mathbf{I}}_{\text{orthogonality}} \end{aligned} \tag{3.7}$$

where α, β, γ are all positive parameters which control contributions of side information guidance, classification loss and regularization, respectively.

3.2 Optimization Framework The model parameters that have to be estimated include $\mathbf{B} \in \mathbb{R}^{m \times k}$, $\mathbf{S} \in \mathbb{R}^{n \times k}$ and $\mathbf{W} \in \mathbb{R}^{k \times c}$. The optimization problem in Eq. (3.7) is not convex with respect to $\mathbf{B}$, $\mathbf{S}$ and $\mathbf{W}$ together. There is no closed-form solution for the problem. We now introduce an alternating scheme to solve the optimization problem. The key idea is to decouple the orthogonal constraints using an Alternating Direction Method of Multipliers (ADMM) scheme [5]. We optimize the objective with respect to one variable, while fixing others. The algorithm will keep updating the variables until convergence.

Update the vertex factor matrix. First, we optimize $\mathbf{B}$ while fixing $\mathbf{S}$ and $\mathbf{W}$. Note that $\mathcal{X}$ is a partially symmetric tensor and the objective function in Eq. (3.7) involves a fourth-order term w.r.t. $\mathbf{B}$ which is difficult to optimize directly. To obviate this problem, we use a variable substitution technique and minimize the following objective function

$$\begin{aligned} \min_{\mathbf{B},\mathbf{P}} \ & \|\mathcal{X} - \mathcal{C} \times_1 \mathbf{B} \times_2 \mathbf{P} \times_3 \mathbf{S}\|_F^2 \\ \text{s.t.} \ & \mathbf{P} = \mathbf{B} \end{aligned} \tag{3.8}$$

where $\mathbf{P}$ are auxiliary variables.

The augmented Lagrangian function for problem in

Eq. (3.8) is

$$\mathcal{L}(\mathbf{B},\mathbf{P}) = \|\mathcal{X} - \mathcal{C} \times_1 \mathbf{B} \times_2 \mathbf{P} \times_3 \mathbf{S}\|_F^2 + \operatorname{tr}(\mathbf{U}^{\mathrm{T}}(\mathbf{P}-\mathbf{B})) + \frac{\mu}{2}\|\mathbf{P}-\mathbf{B}\|_F^2 \tag{3.9}$$

where $\mathbf{U} \in \mathbb{R}^{m\times k}$ are Lagrange multipliers, μ is the penalty parameter which can be adjusted efficiently according to [25].

To compute $\mathbf{B}$, the optimization problem is formulated as

$$\min_{\mathbf{B}} \ \|\mathbf{B}\mathbf{E}^{\mathrm{T}} - \mathbf{X}_{(1)}\|_F^2 + \frac{\mu}{2}\|\mathbf{B} - \mathbf{P} - \frac{1}{\mu}\mathbf{U}\|_F^2 \tag{3.10}$$

where $\mathbf{E} = \mathbf{S} \odot \mathbf{P} \in \mathbb{R}^{(m*n)\times k}$ ($\odot$ is Khatri-Rao product) and $\mathbf{X}_{(1)} \in \mathbb{R}^{m\times(m*n)}$ is the mode-1 matricization of tensor $\mathcal{X}$.

By setting the derivative of Eq. (3.10) w.r.t. $\mathbf{B}$ to zero, we obtain the closed-form solution

$$\mathbf{B} = (2\mathbf{X}_{(1)}\mathbf{E} + \mu\mathbf{P} + \mathbf{U})(2\mathbf{E}^{\mathrm{T}}\mathbf{E} + \mu\mathbf{I})^{-1} \tag{3.11}$$

To efficiently compute $\mathbf{E}^{\mathrm{T}}\mathbf{E}$, we consider the following property of the Khatri-Rao product of two matrices ($*$ is Hadamard product) [20]

$$\mathbf{E}^{\mathrm{T}}\mathbf{E} = (\mathbf{S}\odot\mathbf{P})^{\mathrm{T}}(\mathbf{S}\odot\mathbf{P}) = \mathbf{S}^{\mathrm{T}}\mathbf{S} * \mathbf{P}^{\mathrm{T}}\mathbf{P} \tag{3.12}$$

Then the auxiliary matrix $\mathbf{P}$ can be optimized successively in a similar way

$$\mathbf{P} = (2\mathbf{X}_{(2)}\mathbf{F} + \mu\mathbf{B} - \mathbf{U})(2\mathbf{F}^{\mathrm{T}}\mathbf{F} + \mu\mathbf{I})^{-1} \tag{3.13}$$

where $\mathbf{F} = \mathbf{S}\odot\mathbf{B} \in \mathbb{R}^{(m*n)\times k}$ and $\mathbf{X}_{(2)} \in \mathbb{R}^{m\times(m*n)}$ is the mode-2 matricization of tensor $\mathcal{X}$.

Moreover, we optimize the Lagrange multipliers $\mathbf{U}$ using gradient ascent

$$\mathbf{U} \leftarrow \mathbf{U} + \mu(\mathbf{P} - \mathbf{B}) \tag{3.14}$$

Update the subject factor matrix. Next, we optimize $\mathbf{S}$ while fixing $\mathbf{B}$ and $\mathbf{W}$. We need to minimize the following objective function

$$\mathcal{L}(\mathbf{S}) = \|\mathbf{S}\mathbf{G}^{\mathrm{T}} - \mathbf{X}_{(3)}\|_F^2 + \alpha\operatorname{tr}(\mathbf{S}^{\mathrm{T}}\mathbf{L}_{\mathbf{Z}}\mathbf{S}) + \beta\|\mathbf{D}\mathbf{S}\mathbf{W} - \mathbf{Y}\|_F^2 \quad \text{s.t. } \mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I} \tag{3.15}$$

where $\mathbf{G} = \mathbf{P}\odot\mathbf{B} \in \mathbb{R}^{(m*m)\times k}$ and $\mathbf{X}_{(3)} \in \mathbb{R}^{n\times(m*m)}$ is the mode-3 matricization of tensor $\mathcal{X}$.

Such an optimization problem has been well studied and can be solved by many existing orthogonality preserving methods in the literature [1, 17, 33]. Here we employ the curvilinear search approach introduced in [33], for which we calculate the derivative of $\mathcal{L}(\mathbf{S})$ w.r.t. $\mathbf{S}$ as follows

$$\nabla_{\mathbf{S}}\mathcal{L}(\mathbf{S}) = \mathbf{S}\mathbf{G}^{\mathrm{T}}\mathbf{G} - \mathbf{X}_{(3)}\mathbf{G} + \alpha\mathbf{L}_{\mathbf{Z}}\mathbf{S} + \beta\mathbf{D}^{\mathrm{T}}(\mathbf{D}\mathbf{S}\mathbf{W} - \mathbf{Y})\mathbf{W}^{\mathrm{T}} \tag{3.16}$$

Algorithm 1 t-BNE

Input: $\mathcal{X}, \mathbf{Z}, \mathbf{Y}, \alpha, \beta, \gamma$
Output: $\mathbf{B}, \mathbf{S}, \mathbf{W}$

```
Set μ_max = 10^6, ρ = 1.15
Initialize B, S, W ~ N(0,1), U = 0, μ = 10^-6
repeat
    Update B and P by Eq. (3.11) and Eq. (3.13)
    Update U by Eq. (3.14)
    Update μ by μ ← min(ρμ, μ_max)
    Update S by Eq. (3.16) with the curvilinear search
    Update W by Eq. (3.18)
until convergence
```

Update the weight matrix. Last, we optimize $\mathbf{W}$ while fixing $\mathbf{B}$ and $\mathbf{S}$. We need to minimize the following objective function

$$\mathcal{L}(\mathbf{W}) = \|\mathbf{D}\mathbf{S}\mathbf{W} - \mathbf{Y}\|_F^2 + \gamma\|\mathbf{W}\|_F^2 \tag{3.17}$$

Note that Eq. (3.17) is a regularized least squares problem with the closed-form solution

$$\mathbf{W} = (\mathbf{S}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{S} + \gamma\mathbf{I})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{Y} \tag{3.18}$$

Based on the above analysis, we outline the optimization framework for brain network embedding in Algorithm 1. The stopping criterion is the difference between the explained variations of two consecutive estimations with the threshold value of 10^{-4}. The code has been made available at the author's homepage[1].

3.3 Time Complexity Each ADMM iteration consists of simple matrix operations. Therefore, rough estimates of its computational complexity can be easily derived [24].

The estimate for the update of $\mathbf{B}$ according to Eq. (3.11) is as follows: $O(m^2nk)$ for the computation of the term $2\mathbf{X}_{(1)}\mathbf{E} + \mu\mathbf{P} + \mathbf{U}$; $O((m+n)k^2)$ for the computation of the term $2\mathbf{E}^{\mathrm{T}}\mathbf{E} + \mu\mathbf{I}$ due to Eq. (3.12), and $O(k^3)$ for its Cholesky decomposition; $O(mk^2)$ for the computation of the system solution that gives the updated value of $\mathbf{B}$. An analogous estimate can be derived for the update of $\mathbf{P}$.

Considering $l < n$ and c is usually a small constant, the estimate for the update of $\mathbf{W}$ according to Eq. (3.18)

[1] https://www.cs.uic.edu/~bcao1/code/t-BNE.zip

is as follows: $O(n^2k + nk^2)$ for the computation of the term $\mathbf{S}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{S} + \gamma\mathbf{I}$, and $O(k^3)$ for its Cholesky decomposition; $O(nk)$ for the computation of the term $\mathbf{S}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{Y}$; $O(k^2)$ for the computation of the system solution that gives the updated value of $\mathbf{W}$.

Overall, the updates of model parameters, $\mathbf{B}$, $\mathbf{P}$ and $\mathbf{W}$, require $O(k^3+(m+n)k^2+(m^2n+n^2)k)$ arithmetic operations in total. Note that it excludes the update of $\mathbf{S}$ which depends on the orthogonality preserving method we use.

3.4 Discussion In this part, we introduce several potential extensions to t-BNE and related variations.

Guidance. The side information guidance introduced above essentially regularize the factor matrix row-wise ($\mathbf{S}(i,:)$ in Eq. (3.3)). Another approach to incorporating the side information is doing a coupled matrix and tensor factorization [2]. However, it would introduce additional model parameters, *i.e.*, a factor matrix for auxiliary features. On the other hand, the column-wise guidance information can be added on factors if we have prior knowledge about community information of subjects or brain regions [31]. Alternatively, such knowledge can be modeled as an augmented space [23].

Supervision. Rather than integrating the process of training a classifier with tensor factorization, we could borrow the idea from [9] that latent factors should satisfy the following constraints: (a) must-link: labeled subjects in the same class should be close to each other; (b) cannot-link: labeled subjects in different classes should be far away from each other; (c) separability: unlabeled subjects should be able to be separated from each other. Intuitively, these constraints tend to discover latent factors that can distinguish different class labels among labeled subjects and separate unlabeled subjects from their distribution. Similar to the side information guidance, this idea can be formulated as graph Laplacians induced from the relationships.

Multimodality. A tensor can be constructed from brain networks of each modality, *e.g.*, fMRI and DTI. Then, a joint tensor factorization can be conducted to capture the consensus information across multiple modalities [34].

4 Experiments

4.1 Data Collections Data were collected from 37 patients with primary diagnoses of social anxiety disorder (SAD) or generalized anxiety disorder (GAD), and 32 healthy participants. Each participant underwent an Emotion Regulation Task (ERT). During the ERT session, participants were instructed to look at pictures displayed on the screen. Emotionally neutral pictures (*e.g.*, landscape, everyday objects) and negative pictures (*e.g.*, car crash, natural disasters) would appear on the screen for seven seconds in random orders. One second after the picture on display, a corresponding auditory guide would instruct the participant to *look*: viewing the neutral pictures; to *maintain*: viewing the negative pictures as they normally would; or to *reappraise*: viewing the negative pictures while attempting to reduce their emotion response by re-interpreting the meaning of pictures. All EEG data were recorded using the Biosemi system equipped with an elastic cap with 34 scalp channels. A detailed description about data acquisition and preprocessing is available in [35].

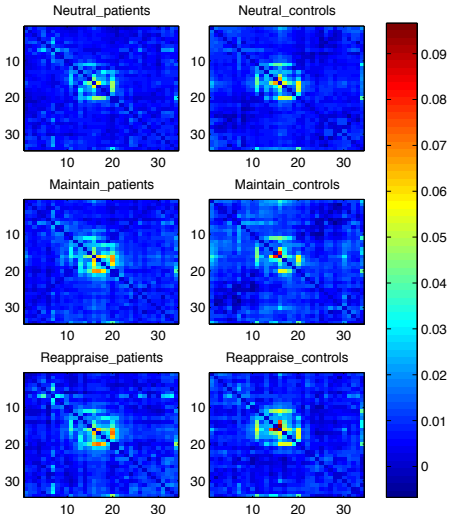


Figure 2: Average brain networks during neutral, maintain and reappraise in the anxiety group (left column) and the control group (right column).

We partitioned the data into three datasets based on each task, denoted as NEUTRAL, MAINTAIN and REAPPRAISE, respectively. Hence, in each dataset, there are $n = 69$ subjects with their corresponding task-specific EEG brain networks which contain $m = 34$ vertices. The target is to distinguish patients from healthy controls. The average brain networks are shown in Figure 2 where the x and y axes represent the vertex id, and the color of the cell represents the strength of the connection between vertex x and y. We can see that the connections between vertices from 11 to 20 are generally stronger than other vertex pairs.

In addition, common self-report scales including Beck Depression Inventory, State-Trait Anxiety Inventory and Leibowitz Social Anxiety Scale were obtained from all participants prior to the test. Each scale was

represented as a range of scores corresponding to the symptom severity rating of depression and anxiety. In general, participants with higher score suggest greater level of disease burden. We use these scales as side information to guide the tensor factorization procedure.

4.2 Compared Methods The compared methods are summarized as follows:

- **t-BNE**: The proposed tensor factorization model for brain network embedding.
- **CMTF**: Coupled matrix and tensor factorization where brain networks and side information are coupled in the subject mode [2].
- **Rubik**: Tensor factorization with orthogonality and sparsity constraints [31].
- **ALS**: Tensor factorization using alternating least squares without any constraint [11].
- **gMSV**: A discriminative subgraph selection approach using side information [9].
- **CC**: Local clustering coefficients, one of the most popular graph-theoretical measures that quantify the cliquishness of the vertices [28].

For a fair comparison, we used a ridge regression [18] as in t-BNE as the base classifier for all the compared methods. All factorization based methods use the same stopping threshold, *i.e.*, the difference between the explained variations of two consecutive estimations with the threshold value of 10^{-4}. Moreover, to assure that all the compared methods have the same data access, especially to the side information, we conducted feature selection using Laplacian Score [16] based on side information for CC. The number of selected features (including that in gMSV) are determined by a hyperparameter k which is equal to the number of latent factors in factorization models. In this manner, the number of (latent) features used for classification are the same in all the compared methods. In summary, k was tuned in the same range of $1, ..., 20$, the regularization parameter γ was tuned in the same range of $2^{-6}, ..., 2^{6}$ for all the compared methods, and other model-specific parameters were set as default, *e.g.*, $\alpha = \beta = 0.1$ in t-BNE and $\lambda_q = 0.1$ in Rubik. In the experiments, 10-fold cross validations were performed. The average accuracy with the best parameter configuration was reported.

4.3 Classification Performance Experimental results in Table 1 show classification performance of compared methods on three datasets. A significant improvement of 20.51%, 6.38% and 12.85% by t-BNE over the best baseline performance can be observed on Neutral, Maintain and Reappraise datasets, respectively. Although clustering coefficients have been widely used to identify Alzheimer's disease [32, 19], they appear to be less useful for distinguishing anxiety patients from normal controls. Note that they are filtered using Laplacian Score [16] based on side information. On the other hand, gMSV achieves better performance than CC on Neutral and Reappraise datasets by extracting connectivity patterns within brain networks that are consistent with the side information guidance.

Table 1: Classification performance (accuracy).

	Datasets		
Methods	Neutral	Maintain	Reappraise
t-BNE	0.7833	0.7548	0.7524
CMTF	0.5810	0.7095	0.6381
Rubik	0.6405	0.6833	0.6667
ALS	0.6119	0.6667	0.6524
gMSV	0.6500	0.6548	0.5952
CC	0.5357	0.6667	0.5357

In general, factorization based models demonstrate themselves with better accuracy. According to the low-rank assumption, a low-dimensional latent factor of each subject is obtained by factorizing the stacked brain network data of all subjects. ALS is a direct application of the alternating least squares technique to the tensor factorization problem without incorporating any domain knowledge. Rubik is a constrained tensor factorization method by regularizing the subject factors to be orthogonal and enforcing sparsity. CMTF incorporates side information by collectively factorizing the brain network tensor and side information matrix with shared subject factors. t-BNE achieves the best performance through the employment of factorizing a partially symmetric tensor, introducing the side information guidance and fusing the processes of tensor factorization and classifier learning.

4.4 Parameter Sensitivity In the experiments above, all the compared methods had the same degree of freedom, *i.e.*, 2, because of the hyperparameters k and γ, and other model-specific parameters were fixed. In this section, in order to evaluate how changes to the parameterization of t-BNE effect its performance on classification tasks, we study the influence of the hyperparameters k and γ, as well as α and β in the proposed t-BNE model.

In Figure 3, experimental results of t-BNE with different k show that a very small k would not be a wise choice in general, and the best performance can usually be achieved around $k = 10$. For the regularization parameter γ, we observe that t-BNE is insensitive to γ in a relatively large range, and a

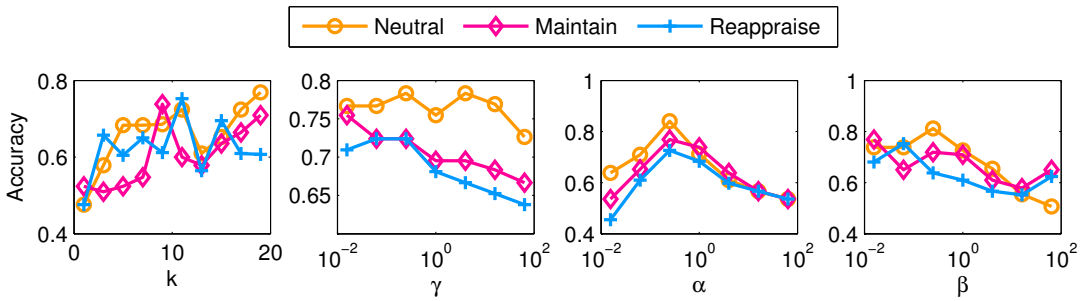
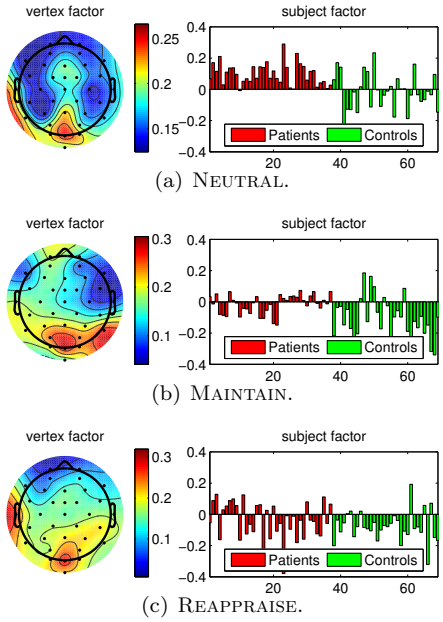


Figure 3: Sensitivity analysis of hyperparameters.



(a) NEUTRAL.

(b) MAINTAIN.

(c) REAPPRAISE.

Figure 4: The largest factors for disease diagnosis on different tasks.

small γ would be preferred (*e.g.*, $\gamma \leq 2^{-2}$). It makes sense because large γ will let the regularization term override the effect of other terms and thus dominate the objective. Considering the influence of α in Figure 3, the performance of t-BNE in Table 1 can be further improved when $\alpha = 2^{-2}$. Neither a small nor a large α would be preferred, because both of the brain network data and the side information guidance are important to learn discriminative representations of brain networks. It is critical to set a desirable α for the trade-off between the two different data sources, in order to acquire the complementary information from them. The sensitivity analysis of t-BNE w.r.t. β generally shows that a good choice of β can be found around 2^{-2}.

4.5 Factor Analysis A k-factor t-BNE model extracts $\mathbf{B}(:, i)$ and $\mathbf{S}(:, i)$, for $i = 1, ..., k$, where these factors indicate the signatures of sources in vertex and subject domain, respectively. We show the largest factors in terms of magnitude for NEUTRAL, MAINTAIN and REAPPRAISE in Figure 4. In the left panel, points indicate the spatial layout of electrodes (*i.e.*, vertices) on the scalp, and factor values of electrodes are demonstrated on a colormap using EEGLAB [14]. The right panel shows the factor strengths for each study participant (both patients and controls). A domain expert may examine the brain activity pattern in the left panel and the differences between participants in such a pattern in the right panel. For instance, visually we note that regardless of the condition (NEUTRAL, MAINTAIN and REAPPRAISE), the first factor always exhibits a pattern of occipital dominance, while the first factor of REAPPRAISE additionally includes a left temporal/parietal involvement, likely reflecting the integration between frontoparietal "cognitive" control networks and temporal limbic regions instrumental for emotion processing.

5 Related Work

To the best of our knowledge, this paper is the first work exploring constrained tensor factorization techniques in the task of brain network embedding for graph classification. In this section, we briefly discuss both of graph classification algorithms and tensor factorization models.

5.1 Graph Classification The recent development of brain network analysis has made characterization of brain disorders at a whole-brain connectivity level possible, thus providing a new direction for brain disease classification. Due to the complex structures and the lack of vector representations, graph data can not be directly used as the input for most data mining algorithms. A straightforward solution that has been extensively explored is to first derive features from brain networks. In general, two types of features are

usually extracted: (1) graph-theoretical measures and (2) subgraph patterns [8].

Wee et al. extract weighted local clustering coefficients of each ROI in relation to the remaining ROIs in brain connectivity networks to quantify the prevalence of clustered connectivity around the ROIs for disease diagnosis on Alzheimer's disease [32]. In addition to a local network property quantified by the weighted clustering coefficients, Jie et al. use a topology-based graph kernel, Weisfeiler-Lehman subtree kernel [29], to measure the topological similarity between paired fMRI brain networks [19]. However, graph kernel methods are not interpretable. Subgraph patterns are more suitable for brain networks, which can simultaneously model the network connectivity patterns around the vertices and capture the changes in local area [21]. Normalized brain networks can be modeled as weighted graphs where each edge is associated with a probability indicating the likelihood of whether this edge should exist or not, based on which Kong et al. propose a discriminative subgraph feature selection method based on dynamic programming to compute the probability distribution of the discrimination scores for each subgraph pattern within a set of weighted graphs [22]. In contrast to focusing on the graph view alone, Cao et al. introduce a subgraph mining algorithm using multiple vector-based side views as guidance to find an optimal set of subgraph features for graph classification [9].

5.2 Tensor Factorization There are increasing research efforts of incorporating constraints in tensor factorization. Carroll et al. describe a least squares fitting procedure with linear constraints for tensor data [10]. Narita et al. provide a framework to use relationships among data as auxiliary information in addition to the low-rank assumption to improve the quality of tensor decomposition [26]. Davidson et al. propose a constrained alternating least squares framework for network analysis of fMRI data [13]. To achieve the purpose of discovering vertices while preserving anatomical adjacency, known anatomical regions in the brain are used as masks and constraints are added to enforce that the discovered factors should closely match these masks. Wang et al. introduce knowledge guided tensor factorization for computational phenotyping [31]. However, some of the guidance and constraints are specifically designed for a domain, thereby making the methods might not work well in other areas, *e.g.*, brain network analysis.

Rather than embedding prior knowledge in the guidance or constraints, another approach to fusing heterogeneous information sources is coupled factorization where matrices and tensors sharing some common modes are jointly factorized [15]. Acar et al. propose a gradient-based optimization approach for joint analysis of matrices and higher-order tensors [2]. Scalable solutions for the coupled factorization problem are presented in [4, 27]. However, these frameworks are not directly applicable to partially symmetric tensor factorization, and they do not leverage all the crucial domain knowledge for brain network embedding.

6 Conclusion and Future Work

This paper presents t-BNE, a novel constrained tensor factorization model for brain network embedding. It handles partially symmetric tensors, incorporates side information guidance and orthogonal constraint to obtain informative and distinct latent factors, and fuses the classifier learning procedure to introduce supervision from labeled data. ADMM is used to solve the optimization objective. In the experiments on EEG datasets, we demonstrate the superior performance of t-BNE on graph classification tasks over baselines.

The brain network embedding problem can be further investigated in several directions for future work. For example, we would like to work with domain experts to incorporate a wider variety of guidance and supervision, and learn a joint representation from multimodal brain network data.

Acknowledgement This work is supported in part by NSF through grants IIS-1526499, CNS-1626432, NSFC 61672313 and NSFC 61503253. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009.

[2] E. Acar, T. G. Kolda, and D. M. Dunlavy, *All-at-once optimization for coupled matrix and tensor factorizations*, arXiv:1105.3422, (2011).

[3] O. Ajilore, L. Zhan, J. GadElkarim, A. Zhang, J. D. Feusner, S. Yang, P. M. Thompson, A. Kumar, and A. Leow, *Constructing the resting state structural connectome*, Frontiers in neuroinformatics, 7 (2013).

[4] A. Beutel, A. Kumar, E. E. Papalexakis, P. P. Talukdar, C. Faloutsos, and E. P. Xing, *Flexi-fact: Scalable flexible factorization of coupled tensors on hadoop*, in SDM, 2014.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends® in Machine Learning, 3 (2011), pp. 1–122.

[6] D. Bzdok, M. Eickenberg, O. Grisel, B. Thirion, and G. Varoquaux, *Semi-supervised factored logistic regression for high-dimensional neuroimaging data*, in NIPS, 2015, pp. 3330–3338.

[7] B. Cao, L. He, X. Kong, P. S. Yu, Z. Hao, and A. B. Ragin, *Tensor-based multi-view feature selection with applications to brain diseases*, in ICDM, IEEE, 2014, pp. 40–49.

[8] B. Cao, X. Kong, and P. S. Yu, *A review of heterogeneous data mining for brain disorder identification*, Brain Informatics, 2 (2015), pp. 253–264.

[9] B. Cao, X. Kong, J. Zhang, P. S. Yu, and A. B. Ragin, *Mining brain networks using multiple side views for neurological disorder identification*, in ICDM, IEEE, 2015, pp. 709–714.

[10] J. D. Carroll, S. Pruzansky, and J. B. Kruskal, *CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters*, Psychometrika, 45 (1980), pp. 3–24.

[11] P. Comon, X. Luciani, and A. L. De Almeida, *Tensor decompositions, alternating least squares and other tales*, Journal of Chemometrics, 23 (2009), pp. 393–405.

[12] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, *A whole brain fMRI atlas generated via spatially constrained spectral clustering*, Human brain mapping, 33 (2012), pp. 1914–1928.

[13] I. Davidson, S. Gilpin, O. Carmichael, and P. Walker, *Network discovery via constrained tensor analysis of fMRI data*, in KDD, 2013, pp. 194–202.

[14] A. Delorme and S. Makeig, *EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis*, Journal of neuroscience methods, 134 (2004), pp. 9–21.

[15] B. Ermiş, E. Acar, and A. T. Cemgil, *Link prediction in heterogeneous data via generalized coupled tensor factorization*, Data Mining and Knowledge Discovery, 29 (2015), pp. 203–236.

[16] X. He, D. Cai, and P. Niyogi, *Laplacian score for feature selection*, in NIPS, 2005, pp. 507–514.

[17] U. Helmke and J. B. Moore, *Optimization and dynamical systems*, Springer Science & Business Media, 2012.

[18] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.

[19] B. Jie, D. Zhang, W. Gao, Q. Wang, C. Wee, and D. Shen, *Integration of network topological and connectivity properties for neuroimaging classification*, Biomedical Engineering, 61 (2014), p. 576.

[20] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM review, 51 (2009), pp. 455–500.

[21] X. Kong and P. S. Yu, *Brain network analysis: a data mining perspective*, SIGKDD Explorations Newsletter, 15 (2014), pp. 30–38.

[22] X. Kong, P. S. Yu, X. Wang, and A. B. Ragin, *Discriminative feature selection for uncertain graph classification*, in SDM, 2013.

[23] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui, *GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation*, in KDD, ACM, 2014, pp. 831–840.

[24] A. P. Liavas and N. D. Sidiropoulos, *Parallel algorithms for constrained tensor factorization via the alternating direction method of multipliers*, arXiv:1409.2383, (2014).

[25] Z. Lin, R. Liu, and Z. Su, *Linearized alternating direction method with adaptive penalty for low-rank representation*, in NIPS, 2011, pp. 612–620.

[26] A. Narita, K. Hayashi, R. Tomioka, and H. Kashima, *Tensor factorization using auxiliary information*, Data Mining and Knowledge Discovery, 25 (2012), pp. 298–324.

[27] E. E. Papalexakis, T. M. Mitchell, N. D. Sidiropoulos, C. Faloutsos, P. P. Talukdar, and B. Murphy, *Turbo-smt: Accelerating coupled sparse matrix-tensor factorizations by 200x*, in SDM, 2014.

[28] M. Rubinov and O. Sporns, *Complex network measures of brain connectivity: uses and interpretations*, Neuroimage, 52 (2010), pp. 1059–1069.

[29] N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, *Weisfeiler-lehman graph kernels*, The Journal of Machine Learning Research, 12 (2011), pp. 2539–2561.

[30] X. Wang, P. Foryt, R. Ochs, J.-H. Chung, Y. Wu, T. Parrish, and A. B. Ragin, *Abnormalities in resting-state functional connectivity in early human immunodeficiency virus infection*, Brain connectivity, 1 (2011), pp. 207–217.

[31] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun, *Rubik: Knowledge guided tensor factorization and completion for health data analytics*, in KDD, ACM, 2015, pp. 1265–1274.

[32] C.-Y. Wee, P.-T. Yap, D. Zhang, K. Denny, J. N. Browndyke, G. G. Potter, K. A. Welsh-Bohmer, L. Wang, and D. Shen, *Identification of mci individuals using structural and functional connectivity networks*, Neuroimage, 59 (2012), pp. 2045–2056.

[33] Z. Wen and W. Yin, *A feasible method for optimization with orthogonality constraints*, Mathematical Programming, 142 (2013), pp. 397–434.

[34] H. Xiao, Y. Li, J. Gao, F. Wang, L. Ge, W. Fan, L. Vu, and D. Turaga, *Believe it today or tomorrow detecting untrustworthy information from dynamic multi-source data*, in SDM, 2015.

[35] M. Xing, R. Tadayonnejad, A. MacNamara, O. Ajilore, K. L. Phan, H. Klumpp, and A. Leow, *EEG based functional connectivity reflects cognitive load during emotion regulation*, in ISBI, IEEE, 2016.

[36] J. Zhang, B. Cao, S. Xie, C.-T. Lu, P. S. Yu, and A. B. Ragin, *Identifying connectivity patterns for brain diseases via multi-side-view guided deep architectures*, in SDM, 2016.