Label-Informed Non-negative Matrix Factorization with Manifold Regularization for Discriminative Subnetwork Detection

Takanori Watanabe[1](✉), Birkan Tunc[1], Drew Parker[1],
Junghoon Kim[2], and Ragini Verma[1]

[1] Section of Biomedical Image Analysis, University of Pennsylvania,
Philadelphia, PA, USA
watanabe@uphs.upenn.edu
[2] The City College of New York, New York, NY, USA

Abstract. In this paper, we present a novel method for obtaining a low dimensional representation of a complex brain network that: (1) can be interpreted in a neurobiologically meaningful way, (2) emphasizes group differences by accounting for label information, and (3) captures the variation in disease subtypes/severity by respecting the intrinsic manifold structure underlying the data. Our method is a supervised variant of non-negative matrix factorization (NMF), and achieves dimensionality reduction by extracting an orthogonal set of subnetworks that are interpretable, reconstructive of the original data, and also discriminative at the group level. In addition, the method includes a manifold regularizer that encourages the low dimensional representations to be smooth with respect to the intrinsic geometry of the data, allowing subjects with similar disease-severity to share similar network representations. While the method is generalizable to other types of non-negative network data, in this work we have used structural connectomes (SCs) derived from diffusion data to identify the cortical/subcortical connections that have been disrupted in abnormal neurological state. Experiments on a traumatic brain injury (TBI) dataset demonstrate that our method can identify subnetworks that can reliably classify TBI from controls and also reveal insightful connectivity patterns that may be indicative of a biomarker.

1 Introduction

Substantial evidence suggests that many major psychiatric and neurological disorders are associated with aberrations in the network structure of the brain [5, 7]. With the availability of modern neuroimaging modalities such as diffusion tensor (DTI) and functional (fMRI) imaging, there is currently an exciting potential for researchers to identify connectivity-based biomarkers of disease states. Since brain networks are known to exhibit complex interactions, multivariate pattern analysis (MVPA) methods are particularly suitable here, as they aim to identify the site of the pathology by examining the data as a whole, accounting for the correlations among the network features.

S. Ourselin et al. (Eds.): MICCAI 2016, Part I, LNCS 9900, pp. 166–174, 2016.
DOI: 10.1007/978-3-319-46720-7_20

In this work, we are interested in applying MVPA methods on diffusion-based structural connectomes (SCs) to identify the patterns of structural dysconnectivity induced by a brain disorder. However, due to the high dimensionality of SCs, standard MVPA methods such as the support vector machine (SVM) become prone to overfitting and thus tend to generalize poorly to test data. Even when generalizability is achieved, SVM lacks clinical interpretability since it returns a dense, high dimensional weight vector. One way to address this is by adding an L1-regularizer to the SVM objective for feature selection [6], but this approach is known to perform poorly when the features are highly correlated. Thus dimensionality reduction becomes critical for improving classification performance and interpretability. Some well-established dimensionality reduction methods in neuroimaging include the principal and independent component analysis (PCA and ICA). However, these approaches do not preserve the non-negativity of the SCs, thus return global representations of brain network that are highly overlapping and lack interpretability since negative structural connection is biologically ill-defined.

Non-negative matrix factorization (NMF) [9] is a relatively recent method that addresses this problem by incorporating non-negativity as a constraint. This constraint leads to a more localized “parts-based” representation where the data is decomposed into purely additive combinations of non-negative basis components. For our work, the bases can be interpreted as data-driven subnetworks, and the corresponding coefficients provide a low-dimensional representation of the SC that can be used in a classifier.

However, despite its success, NMF possesses several limitations. First, NMF does not guarantee the basis components to be local and parts-based, i.e., the subnetworks may be global representations that are overlapping and redundant. Moreover, standard NMF and many of its variants are unsupervised, thus they ignore discriminative structures that may signify important group differences. Finally, NMF assumes that the data are sampled from a Euclidean space, thus does not account for the intrinsic manifold structure underlying the data. While this last issue was addressed in a recent work by Ghanbari et al. [7] under a graph-embedding framework, their method is also unsupervised and thus ignores label information. On the other hand, although supervised subnetwork detection frameworks have been introduced in some recent works [2, 8], these methods do not account for the manifold structure underlying the data.

To overcome these limitations, in this paper we introduce a novel supervised NMF framework for identifying an orthogonal set of subnetworks that is interpretable and emphasizes group differences in structural connectivity. The method also respects the intrinsic geometric structure in the data through manifold regularization [7, 10], which encourages subnetwork representations to be smooth with respect to the data manifold. To solve the proposed objective function, we introduce an optimization algorithm based on the alternating direction method (ADM), which has recently been demonstrated to solve NMF with superior performance over other state-of-the-art algorithms [12]. The proposed framework was evaluated on a TBI dataset, and the results demonstrate the interpretability and the discriminative capacity of the subnetworks.

2 Method

Projective NMF. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$ and $\boldsymbol{y} = [y_1, \cdots, y_n]^T$ denote a set of training samples consisting of SCs $\boldsymbol{x}_i \in \mathbb{R}^p_+$, $i = 1, \cdots, n$, and $y_i \in \{\pm 1\}$ indicates the label of subject i. An SC is a vector representation of the brain network obtained via tractography, where each vector elements represents the strength of structural connection between distinct pair of brain regions (see Sect. 3 for details). Given a target dimension $r \ll p$, NMF learns a decomposition of the form $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$ by minimizing the Frobenius norm error $\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2$, where $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r] \in \mathbb{R}_+^{p \times r}$ is the basis matrix and $\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n] \in \mathbb{R}_+^{r \times n}$ is the coefficient matrix. In the context of our work, the columns of $\boldsymbol{W}$ are connectivity bases that represent subnetworks.

Following [10], we assume that $\boldsymbol{H}$ is obtained from a linear projection of $\boldsymbol{X}$, i.e., $\boldsymbol{H} = \boldsymbol{P}\boldsymbol{X}$, where $\boldsymbol{P} \in \mathbb{R}_+^{r \times p}$ is a nonnegative projection matrix that embeds the data onto the intrinsic subspace. Under this assumption, the objective function for NMF becomes

$$\min_{\boldsymbol{W},\boldsymbol{P} \geq 0} \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{W}(\boldsymbol{P}\boldsymbol{X})\|_F^2. \tag{1}$$

A key advantage of this projective NMF is that once an optimal projection $\boldsymbol{P}^*$ is learned from solving (1), the trained model can be readily generalized to unseen data. That is, given a new test data $\boldsymbol{x}^*$, we can immediately obtain its low dimensional representation by $\boldsymbol{h}^* = \boldsymbol{P}^*\boldsymbol{x}^*$. This is extremely important for running cross-validation (CV).

Orthogonal NMF with Manifold Regularization and Label Information. Despite the merits of the projective NMF, it has three key deficiencies. Firstly, it is often reported that NMF does not necessarily return meaningful parts-based decompositions for some datasets. Secondly, although many real-world data are found to lie in a low dimensional manifold, NMF assumes that the data are sampled from a Euclidean space, neglecting the intrinsic geometric structure in the data. Thirdly, traditional NMF models are unsupervised and thus ignore the discriminative information from the different label groups.

In light of these limitations, we propose to include the following terms in our model:

1. *Orthogonality constraint*: $\boldsymbol{F}_1(\boldsymbol{W}) = I_{\boldsymbol{\Omega}}(\boldsymbol{W})$, where $\Omega := \{\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{p} \times \boldsymbol{r}} | \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W} = \boldsymbol{I_r}\}$ and $I_C(\cdot)$ is the indicator function of a set $C : I_C(\boldsymbol{W}) = 0$ if $\boldsymbol{W} \in C$ and $I_C(\boldsymbol{W}) = \infty$ elsewise.
2. *Manifold regularization*: $\boldsymbol{F}_2(\boldsymbol{P}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \|\boldsymbol{P}\boldsymbol{x}_i - \boldsymbol{P}\boldsymbol{x}_j\| S_{ij}$.
3. *Classification error*: $\boldsymbol{F}_3(\boldsymbol{P}, \boldsymbol{\beta}, b) = \|\boldsymbol{y} - (\boldsymbol{P}\boldsymbol{X})^T\boldsymbol{\beta} - b1_n\|_2^2$, where $\boldsymbol{\beta} \in \mathbb{R}^{\boldsymbol{r}}$ and $b \in \mathbb{R}$ defines a hyperplane in the intrinsic subspace, and $1_n \in \mathbb{R}^{\boldsymbol{n}}$ is a vector of all ones.

The $\boldsymbol{F}_1$ term constrains the basis matrix to reside within the set Ω, which is the set of orthogonal matrices known as the Stiefel manifold [11]. Since $\boldsymbol{W}$ is non-negative, orthogonality implies that the bases representing the subnetworks are non-overlapping, which enhances interpretability and eliminates redundancy.

The $\boldsymbol{F}_2$ term ensures smoothness of the low dimensional representation with respect to the manifold structure encoded in affinity matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$. Intuitively, this regularizer preserves the intrinsic geometric structure in the data by encouraging representations $\boldsymbol{Px}_i$ and $\boldsymbol{Px}_j$ to be close if $S_{i,j}$ is large, i.e., subjects i and j are similar under some notion. This regularizer can also be expressed in terms of the trace operator: $\boldsymbol{F}_2(\boldsymbol{P}) = \mathrm{Tr}\left((\boldsymbol{PX})\boldsymbol{L}(\boldsymbol{PX})^{\boldsymbol{T}}\right)$, where $\boldsymbol{L} \in \mathbb{R}^{n \times n}$ is the graph Laplacian defined by $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, and $\boldsymbol{D}$ is a diagonal matrix with $D_{i,i} = \sum_{j=1}^{n} S_{i,j} \forall i$. While the type of inter-subject relationship that can be encoded via the affinity matrix $\boldsymbol{S}$ is general, in this work, we will take advantage of the clinical scores that are used to evaluate patients, and create a "disease-severity graph" to capture the disease-induced variation in the SCs. Specifically, we will assign higher value to $S_{i,j}$ if subjects i and j share similar severity scores.

Finally, the classification error term $\boldsymbol{F}_3$ enhances the discriminatory power of NMF by encouraging the label groups in the low dimensional embedding $\boldsymbol{PX}$ to be separated by a hyperplane $\boldsymbol{\beta}$ (for clarity, the intercept term b is dropped from our presentation hereon after). Thus, our proposed NMF model seeks to identify subnetwork bases that are not only reconstructive of data but also discriminative of label groups (note that the squared error is used here to allow the ADM algorithm to admit a closed form solution).

Integrating the above constraint terms into the projective NMF Eq. (1) gives us our final objective function ($\lambda_1, \lambda_2 \geq 0$ below are regularization parameters):

$$\min_{\boldsymbol{W}, \boldsymbol{P} \geq 0, \boldsymbol{\beta}} \|\boldsymbol{X} - \boldsymbol{W}(\boldsymbol{PX})\|_F^2 + \lambda_1 \mathrm{Tr}\left((\boldsymbol{PX})\boldsymbol{L}(\boldsymbol{PX})^{\boldsymbol{T}}\right) + \lambda_2 \left\|\boldsymbol{y} - (\boldsymbol{PX})^{\boldsymbol{T}}\boldsymbol{\beta}\right\|_2^2 + I_\Omega(\boldsymbol{W}). \quad (2)$$

ADM Algorithm. We now introduce an optimization algorithm based on the ADM algorithm [12] for solving the proposed cost function. Before applying ADM, we first convert objective function (2) into the following equivalent constrained form by introducing auxiliary variables $\{\boldsymbol{H}, \tilde{\boldsymbol{H}}, \tilde{\boldsymbol{W}}_1, \tilde{\boldsymbol{W}}_2, \tilde{\boldsymbol{P}}\}$ (a technique called *variable splitting*):

$$\min_{\substack{W, P, H, \beta, \\ \tilde{\boldsymbol{P}}, \tilde{\boldsymbol{H}}, \tilde{\boldsymbol{W}}_1, \tilde{\boldsymbol{W}}_2}} \|\boldsymbol{X} - \boldsymbol{WH}\|_F^2 + \lambda_1 \mathrm{Tr}\left(\tilde{\boldsymbol{H}}\boldsymbol{L}\tilde{\boldsymbol{H}}^{\boldsymbol{T}}\right) + \lambda_2 \left\|\boldsymbol{y} - \boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{\beta}\right\|_2^2 + I_+\left(\tilde{\boldsymbol{W}}_1\right) + I_\Omega\left(\tilde{\boldsymbol{W}}_2\right) + I_+\left(\tilde{\boldsymbol{P}}\right)$$

$$\text{such that } \boldsymbol{H} = \boldsymbol{PX}, \boldsymbol{W} = \tilde{\boldsymbol{W}}_1, \boldsymbol{W} = \tilde{\boldsymbol{W}}_2, \boldsymbol{P} = \tilde{\boldsymbol{P}}, \boldsymbol{H} = \tilde{\boldsymbol{H}},$$

where $I_+(\cdot)$ denotes the indicator function of the non-negative orthant. Although the auxiliary variables introduced from variable splitting may appear redundant, this strategy is commonly used in ADM frameworks (see [12] for example), as it allows the ADM subproblems to be solved in closed form. In the context of our work, the augmented Lagrangian (AL) function for the above constrained problem is given by:

$$\begin{aligned}&\mathcal{L}_{\mathrm{AL}}\left(\boldsymbol{W},\boldsymbol{P},\boldsymbol{\beta},\tilde{\boldsymbol{P}},\boldsymbol{H},\tilde{\boldsymbol{H}},\tilde{\boldsymbol{W}}_1,\tilde{\boldsymbol{W}}_2,\boldsymbol{\Lambda}_{\tilde{\boldsymbol{W}}_1},\boldsymbol{\Lambda}_{\tilde{\boldsymbol{W}}_2},\boldsymbol{\Lambda}_{\tilde{\boldsymbol{P}}},\boldsymbol{\Lambda}_{\boldsymbol{H}},\boldsymbol{\Lambda}_{\tilde{H}}\right)=\|\boldsymbol{X}-\boldsymbol{WH}\|_F^2\\&\quad+\lambda_1\mathrm{Tr}\left(\tilde{\boldsymbol{H}}\boldsymbol{L}\tilde{\boldsymbol{H}}^{\boldsymbol{T}}\right)+\lambda_2\|\boldsymbol{y}-\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{\beta}\|_2^2+I_+\left(\tilde{\boldsymbol{W}}_1\right)+I_\Omega\left(\tilde{\boldsymbol{W}}_2\right)+I_+\left(\tilde{\boldsymbol{P}}\right)\\&\quad+\left\langle\boldsymbol{\Lambda}_{\tilde{\boldsymbol{W}}_1},\boldsymbol{W}-\tilde{\boldsymbol{W}}_1\right\rangle+\left\langle\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_2},\boldsymbol{W}-\tilde{\boldsymbol{W}}_2\right\rangle+\left\langle\boldsymbol{\mathit{\Lambda}}_{\boldsymbol{P}},\boldsymbol{P}-\tilde{\boldsymbol{P}}\right\rangle+\left\langle\boldsymbol{\mathit{\Lambda}}_{\boldsymbol{H}},\boldsymbol{H}-\boldsymbol{PX}\right\rangle+\left\langle\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{H}}},\boldsymbol{H}-\tilde{\boldsymbol{H}}\right\rangle\\&\quad+\frac{\rho}{2}\left\{\|\boldsymbol{W}-\tilde{\boldsymbol{W}}_1\|_F^2+\|\boldsymbol{W}-\tilde{\boldsymbol{W}}_2\|_F^2+\|\boldsymbol{P}-\tilde{\boldsymbol{P}}\|_F^2+\|\boldsymbol{H}-\boldsymbol{PX}\|_F^2+\|\boldsymbol{H}-\tilde{\boldsymbol{H}}\|_F^2\right\},\end{aligned}$$

where $\left\{\boldsymbol{W},\boldsymbol{P},\boldsymbol{\beta},\tilde{\boldsymbol{W}}_1,\tilde{\boldsymbol{W}}_2,\tilde{\boldsymbol{P}},\boldsymbol{H},\tilde{\boldsymbol{H}}\right\}$ and $\left\{\boldsymbol{\mathit{\Lambda}}_{\tilde{W}_1},\boldsymbol{\mathit{\Lambda}}_{\tilde{W}_2},\boldsymbol{\mathit{\Lambda}}_{\tilde{P}},\boldsymbol{\mathit{\Lambda}}_{\boldsymbol{H}},\boldsymbol{\mathit{\Lambda}}_{\tilde{H}}\right\}$ are primal and dual variables, $\rho>0$ is the AL penalty parameter, and $\cdot,\cdot$ denotes the trace inner product. The ADM algorhm is derived by alternately minimizing $\mathcal{L}_{\mathrm{AL}}$ with respect to each primal variable while holding others fixed, followed by a gradient ascent step on dual variables. The overall ADM algorithm can be summarized as follows:

```
Repeat until convergence after variable initialization:
```

Primal updates (1)	*Primal updates (2)*	*Dual updates*
$\boldsymbol{P}\leftarrow\arg\min_{\boldsymbol{P}}\mathcal{L}_{AL}$	$\tilde{\boldsymbol{P}}\leftarrow\arg\min_{\tilde{\boldsymbol{P}}}\mathcal{L}_{AL}$	$\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{P}}}\leftarrow\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{P}}}+\rho(\boldsymbol{P}-\tilde{\boldsymbol{P}})$
$\boldsymbol{W}\leftarrow\arg\min_{\boldsymbol{W}}\mathcal{L}_{AL}$	$\tilde{\boldsymbol{W}}_1\leftarrow\arg\min_{\tilde{\boldsymbol{W}}_1}\mathcal{L}_{AL}$	$\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_1}\leftarrow\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_1}+\rho(\boldsymbol{W}-\tilde{\boldsymbol{W}}_1)$
$\boldsymbol{H}\leftarrow\arg\min_{\boldsymbol{H}}\mathcal{L}_{AL}$	$\tilde{\boldsymbol{W}}_2\leftarrow\arg\min_{\tilde{\boldsymbol{W}}_2}\mathcal{L}_{AL}$	$\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_2}\leftarrow\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_2}+\rho(\boldsymbol{W}-\tilde{\boldsymbol{W}}_2)$
$\boldsymbol{\beta}\leftarrow\arg\min_{\boldsymbol{\beta}}\mathcal{L}_{AL}$	$\tilde{\boldsymbol{H}}\leftarrow\arg\min_{\tilde{\boldsymbol{H}}}\mathcal{L}_{AL}$	$\boldsymbol{\mathit{\Lambda}}_{\boldsymbol{H}}\leftarrow\boldsymbol{\mathit{\Lambda}}_{\boldsymbol{H}}+\rho(\boldsymbol{H}-\boldsymbol{PX})$
		$\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{H}}}\leftarrow\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{H}}}+\rho(\boldsymbol{H}-\tilde{\boldsymbol{H}})$

The primal updates above can all be carried out efficiently in closed form:

$\boldsymbol{P}\leftarrow\left(\boldsymbol{H}\boldsymbol{X}^{\boldsymbol{T}}+\tilde{\boldsymbol{P}}+\left[\boldsymbol{\mathit{\Lambda}}_{\boldsymbol{H}}\boldsymbol{X}^{\boldsymbol{T}}-\boldsymbol{\mathit{\Lambda}}_{\boldsymbol{P}}\right]/\rho\right)\left(\boldsymbol{X}\boldsymbol{X}^{T}+\boldsymbol{I}_p\right)^{-1}$	$\tilde{\boldsymbol{P}}\leftarrow\max\left(0,\boldsymbol{P}+\boldsymbol{\Lambda}_{\tilde{\boldsymbol{P}}}/\rho\right)$
$\boldsymbol{W}\leftarrow\left(\boldsymbol{X}\boldsymbol{H}^{\boldsymbol{T}}+\rho\left[\tilde{\boldsymbol{W}}_1+\tilde{\boldsymbol{W}}_2\right]-\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_1}-\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_2}\right)\left(\boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}}+2\rho\boldsymbol{I}_{\boldsymbol{r}}\right)^{-1}$	$\tilde{\boldsymbol{W}}_1\leftarrow\max\left(0,\boldsymbol{W}_1+\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_1}/\rho\right)$
$\boldsymbol{H}\leftarrow\left(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W}+2\rho\boldsymbol{I}_{\boldsymbol{r}}+\lambda_2\boldsymbol{\beta}\boldsymbol{\beta}^{T}\right)^{-1}\left(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}+\rho\boldsymbol{P}\boldsymbol{X}-\boldsymbol{\mathit{\Lambda}}_{\boldsymbol{H}}+\lambda_2\boldsymbol{\beta}\boldsymbol{y}^{T}\right)$	$\tilde{\boldsymbol{H}}\leftarrow\left(\rho\boldsymbol{H}+\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{H}}}\right)\left(\lambda_1\boldsymbol{L}+\rho\boldsymbol{I}_{\boldsymbol{n}}\right)^{-1}$
$\boldsymbol{\beta}\leftarrow\left(\boldsymbol{H}\boldsymbol{H}^{T}\right)^{-1}\boldsymbol{y}$	$\tilde{\boldsymbol{W}}_2\leftarrow\mathrm{Pro}j_{\Omega}\left(\boldsymbol{W}+\boldsymbol{\mathit{\Lambda}}_{\tilde{\boldsymbol{W}}_2}/\rho\right)$

Note $\mathrm{Pro}j_{\Omega}(\cdot)$ for the $\tilde{\boldsymbol{W}}_2$ update denotes the Euclidean projection of a matrix onto the Stiefel manifold. Letting $\boldsymbol{A}\in\mathbb{R}^{p\times r}$ $(r\leq p)$ denote a rank-r matrix, this is given by:

$$\mathrm{Pro}j_{\Omega}(\boldsymbol{A})=\underset{\boldsymbol{Q}\in\Omega}{\mathbf{arg\,min}}\,\|\boldsymbol{A}-\boldsymbol{Q}\|_{\boldsymbol{F}}^2=\boldsymbol{U}\begin{bmatrix}\boldsymbol{I}_{\boldsymbol{r}}\\0\end{bmatrix}\boldsymbol{V}^{\boldsymbol{H}}\qquad(3)$$

Here $\boldsymbol{U\Sigma V^{H}}$ represents the SVD of $\boldsymbol{A}$ and $0\in\mathbb{R}^{(\boldsymbol{p}-\boldsymbol{r})\times\boldsymbol{r}}$ is a matrix of all zeros; solution (3) is unique as long as $\boldsymbol{A}$ is full column rank (see Proposition 7 in [11]).

3 Experiments and Conclusions

Dataset. We apply our method to a TBI dataset consisting of 34 TBI patients and 32 age-matched controls. While the control subjects were scanned only once, the TBI patients were scanned and evaluated at three different time points: 3, 6, and 12 months post-injury. Of the 34 TBI patients, 18 had all 3 time points, 9 had 2 and 7 had only one timepoint. The functional outcome of patients was evaluated using the Glasgow Outcome Scale Extended (GOSE) and Disability Rating Scale (DRS), which are commonly used in TBI. GOSE ranges from 1 = dead to 8 = good recovery, whereas DRS ranges from 0 = normal to 29 = extremely vegetated. In total, the dataset comprises 111 total scans, with 32 labeled *control* and 79 labeled *TBI.* All scans are accompanied with 11 clinical scores that are intended to assess the cognitive functioning of the subject.

Creating the SCs. DTI data was acquired for each subject (Siemens 3T TrioTim, 8 channel head coil, single shot spin echo sequence, TR/TE = 6500/84 ms, b = 1000 s/mm^2, 30 gradient directions). 86 ROIs from the Desikan atlas were extracted to represent the nodes of the structural network. Probabilistic tractography [3] was performed from each of these regions with 100 streamline fibers sampled per voxel, resulting in an 86 × 86 matrix of weighted connectivity values, where each element represents the conditional probability of a pathway between regions, normalized by the active surface area of the seed ROI. Finally, the 86 × 86 connectivity matrix of each subject was vectorized to its p = 3655 lower triangular elements, resulting in $\boldsymbol{x} \in \mathbb{R}_+^{\boldsymbol{p}}$ representing the SC.

Implementation Details. We applied our method to SCs computed from the TBI dataset to compute the subnetwork bases and their corresponding NMF coefficients; here we let y = + 1 indicate TBI and y = - 1 indicate control. The disease-severity graph was created using the functional outcome indices of GOSE/DRS as follows. First, we constructed a symmetrized k-nearest-neighbor (k-NN) graph with k = 5, where the distance between scans i and j was measured as $d_{i,j} = (\mathrm{GOS}E_i\text{-}\mathrm{GOS}E_j)^2 + (\mathrm{DR}S_i - \mathrm{DR}S_j)^2$. Then a binary affinity graph was created by setting $S_{i,j}$ to 1 if and only if scans i and j were connected by the k-NN graph and did not represent the same subject (to avoid connecting same TBI patients who underwent multiple scans); controls were left un-connected.

We identified r = 5 subnetwork bases using this affinity graph, and the regularization parameters were set at $\lambda_1 = \lambda_2 = 0.25$, as the model became stable around this value (degradation in classification performance was observed when parameters were set at $\lambda_1 = \lambda_2 = 0$, i.e., a setup equivalent to traditional NMF). To initialize the ADM variables, we use the strategy introduced in [4] to deterministically initialize $\boldsymbol{W}$ and $\boldsymbol{H}$ and set all other variables to zero for replicability. The AL parameter value was set to ρ = 1000 based on empirical test runs, and the ADM algorithm was terminated when the relative change in the objective function value (Eq. 2) at successive iterations fell below 10^{-4} and the following primal residual condition was met:

$$\max\left(\frac{\|\boldsymbol{W}-\tilde{\boldsymbol{W}}_1\|_F}{\|\boldsymbol{W}\|_F},\frac{\|\boldsymbol{W}-\tilde{\boldsymbol{W}}_2\|_F}{\|\boldsymbol{W}\|_F},\frac{\|\boldsymbol{H}-\boldsymbol{PX}\|_F}{\|\boldsymbol{H}\|_F},\frac{\|\boldsymbol{H}-\tilde{\boldsymbol{H}}\|_F}{\|\boldsymbol{H}\|_F},\frac{\|\boldsymbol{P}-\tilde{\boldsymbol{P}}\|_F}{\|\boldsymbol{P}\|_F}\right)<10^{-4}.$$

To remove features that are likely non-biological, we applied feature selection using the aforementioned 11 clinical scores. Precisely, we first correlated individual SC features with each clinical score to obtain 11 separate p-value rankings (rank = 1 the smallest), and summed these rankings to obtain a rank-sum value for each feature. We then selected 1000 features having the smallest rank-sum that were then standardized via linear scaling to the range [0,1]. This feature selection and standardization procedures were conducted within the CV-folds to avoid biasing the classification performance.

We compared the performance for the following classifiers (implemented using Liblinear [6]). The first three methods are applied to the 1000 features selected using the above procedure: (1) L1-loss L2-regularized SVM (SVM), (2) L2-loss, L1 regularized SVM (SVM + L1), (3) L1-regularized Logistic regression (LogReg + L1), and (4) L1-loss L2-regularized SVM applied to the projected NMF coefficients with our method. A weighted loss function was used for all classifiers, where the weights assigned to each label class is inversely proportional to the class frequency. Since subjects have multiple timepoints, the classification accuracy was assessed using a Leave-One-Subject-Out CV (LOSO-CV) procedure, where all scans from a test subject are iteratively left out during training. Finally, the hyperparameter C, which is common to all classifiers, were tuned via an internal LOSO-CV over the range $C \in \{2^{-10}, 2^{-9}, \cdots, 2^{10}\}$.

3.1 Experimental Results and Conclusions

Classification Results. Table 1 reports the classification results from LOSO-CV for different methods, showing overall accuracy, specificity (type I error), sensitivity (type II error), and balanced score rate (BSR), which is the mean of specificity and sensitivity. The results show that the classification performance obtained using the proposed subnetwork features demonstrates a noticeable improvement over using the SC features in its original form, achieving accuracy of 82.0 % and a BSR of 81.8 %. The SVM achieves the next best performance, but the model is hard to interpret since all 1000 edge features contribute to the classifier. Finally, despite using a weighted loss function, we see the sparsity-promoting L1-regularized classifiers suffer from low sensitivity, which

Table 1. Classification results from “leave-one-subject-out” cross-validation.

Classifier	Accuracy	Sensitivity	Specificity	BSR
SVM	76.6 %	77.2 %	75.0 %	76.1 %
SVM + L1	69.4 %	73.4 %	59.4 %	66.4 %
LogReg + L1	67.6 %	70.9 %	59.4 %	65.1 %
Proposed NMF + SVM	**82.0 %**	**82.3 %**	**81.3 %**	**81.8 %**

is likely caused by data label imbalance, as well as the correlated structures among the features (a case where L1-regularizations tend to suffer).

Effect of Manifold Regularization. We next assessed whether the manifold regularizer with the disease-severity graph has successfully preserved the inter-patient relationship in terms of GOSE/DRS functional outcome indices. To do this, we computed Spearman's rank correlation between the subnetwork bases coefficients and GOSE/DRS indices from the 79 TBI scans. The results reported in Table 2 reveal that for all basis coefficients, consistently positive and negative correlations (statistically significant) are obtained for GOSE and DRS, respectively. This result indicates that subjects with similar level of disease-severity share similar representations in the embedding space, demonstrating the impact of manifold regularization.

Table 2. Spearman's correlation coefficients and corresponding p values between the $r = 5$ subnetwork basis coefficients and DRS/GOSE severity scores among TBI patients.

Basis label	Basis coefs' correlation with DRS	Basis coefs' correlation with GOSE
1	−0.538 (p = 3.24e-7)	0.596 (p = 6.75e-9)
2	−0.464 (p = 1.65e-5)	0.584 (p = 1.63e-8)
3	−0.387 (p = 4.19e-4)	0.408 (p = 1.88e-4)
4	−0.516 (p = 1.12e-6)	0.607 (p = 3.00e-9)
5	−0.517 (p = 1.08e-6)	0.605 (p = 3.54e-9)

Subnetwork Visualization. Given the high predictive capacity of subnetwork coefficients, we next examine their corresponding subnetwork bases $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_5]$ to assess the pathological impact TBI may have induced on structural connectivity. For visualization and interpretation, we retrained the proposed NMF model using the entire dataset, and learned an SVM hyperplane $\boldsymbol{\beta} \in \mathbb{R}^5$ in the corresponding embedding space. The resulting subnetworks are rendered in 3-D brain space in Fig. 1 (figures generated using Python module Nilearn [1]); the color of the edges represent the sign of the hyperplane coefficients in $\boldsymbol{\beta}$, with red indicating contribution towards TBI (positive) and blue indicating contribution towards control (negative). From the figure, we can see

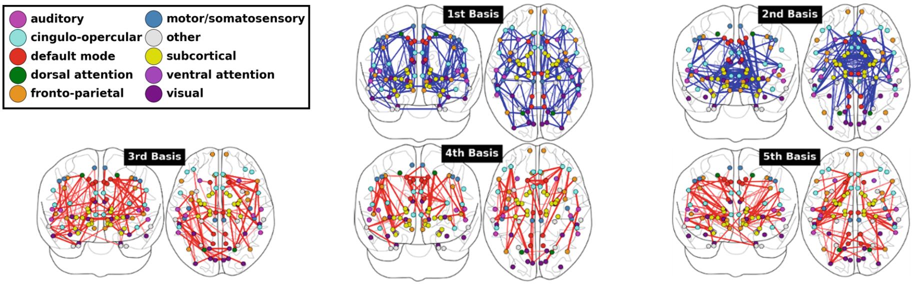

Fig. 1. The subnetwork bases obtained with $r = 5$. The edge color represents the sign of the corresponding hyperplane coefficient $\boldsymbol{\beta} \in \mathbb{R}^r$ (blue = negative/control, red = positive/TBI).

that the network structure of the first basis exhibits strong bilateral symmetry with notable inter-hemispheric connections between the cerebellar, precuneus, and cingulate regions. Moreover, the second subnetwork basis resembles dense inter-hemispheric connections among the subcortical regions, with the sign indicating that these edges tend to be the weaker among TBI patients. On the other hand, subnetwork bases 3–5 represents connection towards TBI. Overall, the subnetworks exhibit a diffuse connectivity pattern that spans across the cortex, suggesting that damages from TBI results in a widespread disturbance in brain network. Interestingly, the connectivity patterns in the first two bases exhibit rich connectivity pattern within the subcortical and medial posterior regions, which are frequently reported to be vulnerable in TBI.

Conclusions. We have presented a supervised NMF framework for extracting a disjoint set of subnetworks that are interpretable and highlight group differences in structural connectivity. The method is also capable of preserving the manifold structure in the data encoded by an affinity graph, thereby respecting the intrinsic geometry of the data. Experiment on a TBI dataset shows that the subnetworks identified from our method can not only be used to reliably discriminate TBI from controls, but also exhibit tight correlation with TBI-outcome indices, indicating that subjects with similar level of TBI-severity share similar subnetwork representations due to manifold regularization.

References

1. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., et al.: Machine learning for neuroimaging with scikit-learn. Front. Neuroinformatics **8**(14) (2014)
2. Allahyar, A., Ridder, J.: FERAL: network-based classifier with application to breast cancer outcome prediction. Bioinformatics **31**(12), i311–i319 (2015)
3. Behrens, T., et al.: Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. Nat. Neurosci. **6**(7), 750–757 (2003)
4. Boutsidis, C., Gallopoulos, E.: SVD based initialization: a head start for nonnegative matrix factorization. Pattern Recognit. **41**, 1350–1362 (2008)
5. Cheplygina, V., Tax, D.M., Loog, M., Feragen, A.: Network-guided group feature selection for classification of autism spectrum disorder. In: Wu, G., Zhang, D., Zhou, L. (eds.) MLMI 2014. LNCS, vol. 8679, pp. 190–197. Springer, Heidelberg (2014)
6. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
7. Ghanbari, Y., Smith, A.R., Schultz, R.T., Verma, R.: Identifying group discriminative and age regressive sub-networks from DTI-based connectivity via a unified framework of non-negative matrix factorization and graph embedding. Med. Image Anal. **18**(8) (2014)
8. Kasenburg, N., et al.: Supervised hub-detection for brain connectivity. In: Proceedings of the SPIE, vol. 9784, Medical Imaging 2016: Image Processing, p. 978409 (2016)
9. Lee, D.D., Seung, H.S.: Learning the parts of objects by NMF. Nature **401**, 788–791 (1999)
10. Liu, X., et al., H.: Projective nonnegative graph embedding. IEEE Trans. Image Proc. (2010)
11. Manton, J.H.: Optimization algorithms exploiting unitary constraints. IEEE Trans. Signal Process. **50**(3), 635–650 (2002)
12. Xu, Y., Yin, W., Wen, Z., Zhang, Y.: An alternating direction algorithm for matrix completion with nonnegative factors. Front. Math. China **7**(2), 365–384 (2012)