Unified and Contrasting Graphical Lasso for Brain Network Discovery

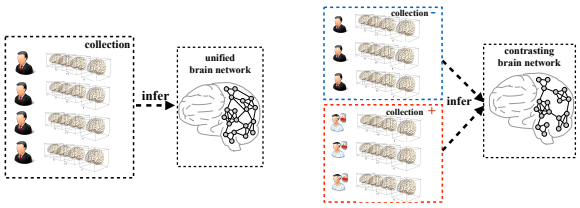
Xinyue Liu* Xiangnan Kong* Ann B. Ragin†

Abstract

The analysis of brain imaging data has attracted much attention recently. A popular analysis is to discover a network representation of brain from the neuroimaging data, where each node denotes a brain region and each edge represents a functional association or structural connection between two brain regions. Motivated by the multi-subject and multi-collection settings in neuroimaging studies, in this paper, we consider brain network discovery under two novel settings: 1) *unified setting:* Given a collection of subjects, discover a single network that is good for all subjects. 2) *contrasting setting*: Given two collections of subjects, discover a *single* network that best discriminates two collections. We show that the existing formulation of graphical Lasso (GLasso) cannot address above problems properly. Two novel models, UGLasso (Unified Graphical Lasso) and CGLasso(Contrasting Graphical Lasso), are proposed to address these two problems respectively. We evaluate our methods on synthetic data and two real-world functional magnetic resonance imaging (fMRI) datasets. Empirical results demonstrate the effectiveness of the proposed methods.

1 Introduction

Recent years have witnessed an increasing amount of data in the form of graph representations, which involve complex structures, *e.g.*, brain networks and social networks. For instance, a brain network is composed of brain regions as the nodes and functional/structural connectivities between the brain regions as the links. The network representation of human brain as shown in right hand side of Fig. 1(a) is useful in many ways. For example, one can learn subgraph patterns in the brain networks to build classification models for disease diagnosis [11]. However, in many real-world studies, the connectivities between different brain regions are not available and should be derived from the neuroimaging data, *e.g.*, fMRI data. In this paper, we study the problem of brain network discovery, which aims at inferring the functional connectivities among a set of predefined non-overlapping brain regions. Previous studies usually focus on inferring a network for a single subject or treating a collection of subjects as a single subject by concatenating the data of multiple subjects [10, 14]. As the increasing availability of neuroimaging data in recent years, we usually have one or more collections of subjects in brain datasets. The problem of discovering a network across collection of subjects is interesting and important. In this paper, we explore two novel settings on brain network discovery. The first one is to find a representative brain network on a single collection, where the discovered network is good for all subjects in the collection, while it is similar to each subject's best network. We call this setting *unified network discovery*, which is illustrated in Fig. 1(a). The second setting we study is to discover a discriminative network between two collections of subjects, where the inferred network retains the differential connectivities between two collections. We call the second setting *contrasting network discovery*, which is illustrated in Fig. 1(b). Such scenario is very common in neuroimaging analysis, where subjects can be grouped using different attributes, such as genders, ages, neurological diseases *etc.*



(a) Unified Brain Network Discovery (b) Contrasting Brain Network Discovery

Figure 1: An illustration of two tasks of brain network discovery.

Problems Studied: Fig. 2 presents the dicovered individual brain network for five healthy subjects in ADNI dataset using standard GLasso. Even all five subjects come from the healthy collection, one can observe that the networks are widely different from each other. This example illustrates the difficulty of discovering a representative network for a collection of subjects. Hence,

*Department of Computer Science, Worcester Polytechnic Institute

†Feinberg School of Medicine, Northwestern University

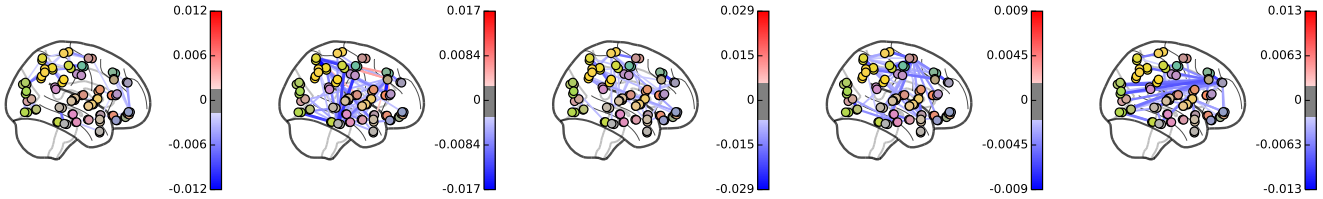


Figure 2: The individual networks derived by GLasso from fMRI scans of five healthy subjects in ADNI dataset.

inferring a network without considering the differences among subjects may lead to unsatisfactory results. Unified network discovery aims at finding a single representative network that is good for all subjects in a collection, which helps the neurology professionals to derive the common connectivity patterns for certain group of individuals. In this paper, we propose a novel algorithm called UGLasso (Unified Graphical Lasso) to address the unified network discovery problem.

In the contrasting setting where we have two collections of subjects, an usual question one may ask is what are the differences between the two collections. A common attempt to answer the question is to infer a network for each collection respectively and to compare them. However, this approach can be hindered in certain circumstances. For instance, in some neuroimaging datasets, the number of nodes in the network can be as large as 40 thousands. Thus, inferring two separated large-scale networks can be expensive. Besides, due to the underlying unreliability and the existence of noises in the signals, it would be difficult and time consuming for one to extract the differences between two inferred networks. Hence, solving the proposed contrasting network discovery problem is a much more efficient way to obatin the discriminative connectivity patterns between two collections. In this paper, we propose another novel algorithm called CGLasso (Contrasting Graphical Lasso) to address the contrasting network discovery problem.

The contributions of this paper are as follows.

- We formulate the novel problems of unified network discovery and contrasting network discovery.
- We show how to use a modified gradient projected to solve the two proposed problems while preserve the solution to be positive-definite.
- We demonstrate the effectiveness of our proposed methods on synthetic datasets with ground truth and on two real-world neuroimaging datasets.

2 Problem Formulation

2.1 Preliminary Assume we are given n observations $\mathbf{X} \in \mathbb{R}^{n\times m}$ from a m-variate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where n denotes the number of samples, m denotes the number of variables and $\mathbf{\Sigma}$ denotes the covariance matrix of the distribution. The problem of estimating the inverse of covariance matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ from $\mathbf{X}$ is known as the inverse covariance estimation [2, 7]. If the (i, j)-th entry of $\mathbf{\Theta}_{ij}$ is zero, then variables i and j are conditionally independent, given the other variables. The inverse covariance estimation can be cast as the problem of minimizing ℓ_1-regularized *negative log likelihood* as

$$\underset{\mathbf{\Theta}\succ 0}{\text{minimize}} \; -\log\det \mathbf{\Theta} + \text{tr}(\mathbf{S\Theta}) + \lambda||\mathbf{\Theta}||_1 \qquad (2.1)$$

where $\mathbf{S} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ is the empirical covariance matrix, $||\mathbf{\Theta}||_1$ is the ℓ_1-norm regularization that encourages sparse solutions, and λ is a positive parameter denotes the strength of regularization. In the case where $\mathbf{S} \succ 0$, the maximum likelihood estimate (MLE) of $\mathbf{\Sigma}^{-1}$ can be recovered by setting $\lambda = 0$. However, in many high dimensional datasets, the number of samples n can be smaller than the number of variables m, and $\mathbf{S}$ can be singular. In such cases, additional regularizations, such as ℓ_1-norm, are usually used to estimate $\mathbf{\Theta}$. It is obvious that solving Eq. (2.1) leads to the ℓ_1-regularized maximum likelihood estimation (MLE) of $\mathbf{\Sigma}^{-1}$.

Stacking Approach to Multi-subject Study: In brain imaging studies, researchers usually collect data from multiple subjects. The data of the i-th subject can be represented by $\mathbf{X}_i \in \mathbb{R}^{n_i\times m}$, where n_i is the number of samples of subject i. Conventional approaches [10, 14] on multi-subject studies usually stack the data matrices of different subjects into $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N) \in \mathbb{R}^{(\sum_{i=1}^N n_i)\times m}$. Then $\frac{1}{\sum_{i=1}^N n_i}\mathbf{X}^T\mathbf{X}$ can be used as $\mathbf{S}$ in Eq. (2.1) to obtain a unified network $\mathbf{\Theta}$ of all subjects. However, this approach does not consider the differences between subjects.

2.2 Unified Graphical Lasso In the unified setting, we are given a collection of data matrices $\{\mathbf{X}_1, \ldots, \mathbf{X}_p\}$ with the same sets of m variables, where $\mathbf{X}_i \in \mathbb{R}^{n_i\times m}$. And thus, we can compute a corresponding collection of empirical covariance matrices $\{\mathbf{S}_1, \ldots, \mathbf{S}_p\}$, where $\mathbf{S}_i = \frac{1}{n_i}\mathbf{X}_i^T\mathbf{X}_i$. The goal is to derive a single estimated inverse covariance matrix $\hat{\mathbf{\Theta}}$ that obeys the following two criteria: *i) Has high likelihood for all subjects. ii) The single estimated inverse covariance matrix is*

similar to each subject's individual estimated inverse covariance matrix. The first criterion requires the solution to be quantitatively good on every subject. The second criterion put enforcement on the solution to be quantitatively similar to the estimated inverse covariance matrix of each subject in the collection.

The *negative log likelihood* of a estimated inverse covariance matrix $\mathbf{\Theta}$ for subject i is defined as $-\log\det\mathbf{\Theta} + \mathrm{tr}(\mathbf{S}_i\mathbf{\Theta})$. And the overall likelihood can be expressed by the average the likelihood of $\mathbf{\Theta}$ for each subject:

$$\begin{aligned} L(\mathbf{\Theta}, \mathbf{S}_1, \ldots, \mathbf{S}_p) &= \frac{1}{p}\sum_{i=1}^{p}\big(-\log\det\mathbf{\Theta} + \mathrm{tr}(\mathbf{S}_i\mathbf{\Theta})\big) \\ &= -\log\det\mathbf{\Theta} + \mathrm{tr}(\bar{\mathbf{S}}\mathbf{\Theta}) \end{aligned} \tag{2.2}$$

where $\bar{\mathbf{S}} = \frac{1}{p}\sum_{i=1}^{p}\mathbf{S}_i$. Further we define the similarity between two inverse covariance matrices $\mathbf{\Theta}_i$ and $\mathbf{\Theta}_j$ as the square of the Frobenius norm of $\mathbf{\Theta}_i - \mathbf{\Theta}_j$. So the criterion 2) can be written as the minimization of $\sum_{i=1}^{p}||\mathbf{\Theta}-\hat{\mathbf{\Theta}}_i||_F^2$, where $\hat{\mathbf{\Theta}}_i = \arg\min_{\mathbf{\Theta}\succ 0} -\log\det\mathbf{\Theta} + \mathrm{tr}(\mathbf{S}_i\mathbf{\Theta}) + \lambda||\mathbf{\Theta}||_1$ is the estimated individual inverse covariance matrix for subject i.

Maximizing criterion 1) is equivalent to minimizing Eq. (2.2). With the objective 1) and 2), we adopt a standard approach of combining them into a objective function with a weighting parameter α and an ℓ_1-norm regularization as follows, which is solved by UGLasso:

$$\underset{\mathbf{\Theta}\succ 0}{\text{minimize}}\ L(\mathbf{\Theta}, \mathbf{S}_1, \ldots, \mathbf{S}_p) + \frac{\alpha}{p}\sum_{i=1}^{p}||\mathbf{\Theta} - \hat{\mathbf{\Theta}}_i||_F^2 + \lambda||\mathbf{\Theta}||_1 \tag{2.3}$$

2.3 Contrasting Graphical Lasso In the contrasting setting, we are given two collections of data matrices $\{\mathbf{X}_1^{(A)}, \ldots, \mathbf{X}_p^{(A)}\}$ and $\{\mathbf{X}_1^{(B)}, \ldots, \mathbf{X}_q^{(B)}\}$ with the same set of variables. We can compute the corresponding empirical covariance matrices $\{\mathbf{S}_1^{(A)}, \ldots, \mathbf{S}_p^{(A)}\}$ and $\{\mathbf{S}_1^{(B)}, \ldots, \mathbf{S}_q^{(B)}\}$ accordingly. The goal is to derive a single discriminative inverse covariance matrix between two collections in that the likelihood is small for the first collection but large for the second. The estimated contrasting inverse covariance matrix is the one whose likelihoods on the each subject best distinguishes the two collections. Our objective function is defined as follows, which is solved by CGLasso:

$$\begin{aligned} &\underset{\mathbf{\Theta}\succ 0}{\text{minimize}}\ \frac{1}{p}\sum_{i=1}^{p}\mathrm{tr}(\mathbf{S}_i^{(A)}\mathbf{\Theta}) - \frac{1}{q}\sum_{j=1}^{q}\mathrm{tr}(\mathbf{S}_j^{(B)}\mathbf{\Theta}) + \lambda||\mathbf{\Theta}||_1 \\ &\qquad = \mathrm{tr}(\hat{\mathbf{S}}\mathbf{\Theta}) + \lambda||\mathbf{\Theta}||_1 \end{aligned} \tag{2.4}$$

where $\hat{\mathbf{S}} = \frac{1}{p}\sum_{i=1}^{p}\mathbf{S}_i^{(A)} - \frac{1}{q}\sum_{j=1}^{q}\mathbf{S}_j^{(B)}$. The log determinant terms for two collections are canceled under the contrasting setting. Although Eq. (2.4) resembles linear programming problem, the existence of ℓ_1-norm regularization and positive-definite constraint prohibits the standard approach and makes it challenging to solve.

3 Algorithms

In this section we present the UGLasso and CGLasso algorithms in detail. Both objectives proposed in Sec. 2 can be transformed to the following form:

$$\underset{\mathbf{\Theta}\succ 0}{\text{minimize}}\ g(\mathbf{\Theta}) + \lambda||\mathbf{\Theta}||_1 \tag{3.5}$$

where $g(\mathbf{\Theta})$ is a differentiable smooth function and $\lambda||\mathbf{\Theta}||_1$ is the non-differentiable regularization function. Specifically, we have $g(\mathbf{\Theta}) = L(\mathbf{\Theta}, \mathbf{S}_1, \ldots, \mathbf{S}_p) + \alpha\sum_{i=1}^{p}||\mathbf{\Theta} - \hat{\mathbf{\Theta}}_i||_F^2$ for unified setting and $g(\mathbf{\Theta}) = \mathrm{tr}(\hat{\mathbf{S}}X)$ for contrasting setting.

Following the classic strategy in linear programming for addressing the ℓ_1-norm minimization problems, we transform the elements of the norm into positive and negative parts and reform Eq. (3.5) as follows:

$$\begin{aligned} \underset{\mathbf{\Theta}^+,\mathbf{\Theta}^-}{\text{minimize}}\quad & g(\mathbf{\Theta}^+ - \mathbf{\Theta}^-) + \lambda\ \mathrm{vec}(\mathbf{\Theta}^+)^\top\mathbf{1} + \lambda\ \mathrm{vec}(\mathbf{\Theta}^-)^\top\mathbf{1} \\ \text{subject to}\quad & \mathbf{\Theta}^+ - \mathbf{\Theta}^- \succ 0 \\ & \mathbf{\Theta}^+ \geq 0, \mathbf{\Theta}^- \geq 0 \end{aligned} \tag{3.6}$$

where $\mathbf{\Theta}_{ij}^+ = \max(\mathbf{\Theta}_{ij}, 0)$, $\mathbf{\Theta}_{ij} = \max(-\mathbf{\Theta}_{ij}^-, 0)$, $\mathbf{1}$ is the column-vector of all ones which has the same length as $\mathrm{vec}(\mathbf{\Theta})$, so $\mathrm{vec}(\mathbf{\Theta})^T\mathbf{1} = \sum_{ij}\mathbf{\Theta}_{ij}$. Thus, it is obvious that $\mathrm{vec}(\mathbf{\Theta}^+)^\top\mathbf{1} + \mathrm{vec}(\mathbf{\Theta}^-)^\top\mathbf{1} = ||\mathbf{\Theta}||_1$, and $\mathbf{\Theta} = \mathbf{\Theta}^+ - \mathbf{\Theta}^-$, then Eq. (3.5) and Eq. (3.6) shares the same minimizer. We also use $\mathbf{\Theta}^+ \geq 0$ to denote that $\mathbf{\Theta}_{ij}^+ \geq 0$, for all $1 \leq i, j \leq m$. Same thing for $\mathbf{\Theta}^- \geq 0$. We further use $\tilde{\mathbf{\Theta}} = (\mathbf{\Theta}^+, \mathbf{\Theta}^-)$ to simplify Eq. (3.6)

$$\begin{aligned} \underset{\tilde{\mathbf{\Theta}}}{\text{minimize}}\quad & \underbrace{\tilde{g}(\tilde{\mathbf{\Theta}}) + \lambda\ \mathrm{vec}(\tilde{\mathbf{\Theta}})^\top\mathbf{1}}_{f(\tilde{\mathbf{\Theta}})} \\ \text{subject to}\quad & \mathbf{\Theta}^+ - \mathbf{\Theta}^- \succ 0 \\ & \tilde{\mathbf{\Theta}} \geq 0 \end{aligned} \tag{3.7}$$

where $\tilde{g}(\tilde{\mathbf{\Theta}})$ is the corresponding equivalent dual function of $g(\mathbf{\Theta})$. In contrasting setting where $g(\mathbf{\Theta}) = \mathrm{tr}(\hat{\mathbf{S}}\mathbf{\Theta})$, the corresponding dual can be represented as $\tilde{g}(\tilde{\mathbf{\Theta}}) = \mathrm{tr}(\overset{*}{\mathbf{S}}\tilde{\mathbf{\Theta}})$, where $\overset{*}{\mathbf{S}} = (\hat{\mathbf{S}}, -\hat{\mathbf{S}})$. Similar construction can be performed on the unified setting too.

We note that Eq. (3.7) is a smooth optimization problem with non-negativity and positive-definite constraints. If we treat the positive-definite constraint

Algorithm 1 Algorithm for Solving Eq. (3.7)

Require: $\mathbf{S}, \lambda, iter_{max}$

Initialize $\mathbf{\Theta}_0 \leftarrow \mathbf{I}^{m\times m}$, $iter \leftarrow 0$
Project the initial estimation $\mathbf{\Theta}_0 \leftarrow \mathcal{P}_{\mathcal{C}}(\mathbf{\Theta}_0)$
$f_t \leftarrow f(\mathbf{\Theta}^{(0)})$, $g_t \leftarrow \nabla f(\mathbf{\Theta}^{(0)})$
repeat
    Initialize s using Eq. (3.12)
    Find the largest s satisfies Eq. (3.11) and positive definite constraint by performing the non-monotonic Armijo backtracking line search.
    Compute the new projection $\mathbf{\Theta}_{t+1} \leftarrow \mathcal{P}_{\mathcal{C}}(\mathbf{\Theta}_t - s\mathbf{g}_t)$ using Eq. (3.10).
    Compute the new objective function $f_{t+1} \leftarrow f(\mathbf{\Theta}_{t+1})$
    Compute the new gradient $g_{t+1} \leftarrow \nabla f(\mathbf{\Theta}_{t+1})$
until $iter = iter_{max}$ or convergence
Return $\mathbf{\Theta}_{t+1}$

as inactive, the remaining constraint is a convex non-negative cone, then one can apply projected gradient method to solve Eq. (3.7). In this work, we consider a variant of the projected gradient method that updates the solution in each iteration as

$$\tilde{\mathbf{\Theta}}^{(i+1)} \leftarrow \mathcal{P}_{\mathcal{C}}(\tilde{\mathbf{\Theta}}^{(i)} - s\nabla f(\tilde{\mathbf{\Theta}}^{(i)})) \tag{3.8}$$

where s is the step size to be selected by backtracking line search strategy and $\mathcal{P}_{\mathcal{C}}$ is a function defined by

$$\mathcal{P}_{\mathcal{C}}(\mathbf{\Theta}) \stackrel{\triangle}{=} \arg\min_{\mathbf{y}\in\mathcal{C}} ||\mathbf{\Theta} - \mathbf{y}||_2 \tag{3.9}$$

is the Euclidean projection of $\mathbf{\Theta}$ onto convex set $\mathcal{C}$. Here, we have $\mathcal{C} = \{\mathbf{y} : \mathbf{y} > 0\}$ is a non-negative cone. With this convex set, the solution to Eq. (3.9) is trivial, we simply project every dimension of $\mathbf{\Theta}$ to the non-negative part as

$$\mathbf{y}_{ij} \leftarrow \max(\mathbf{\Theta}_{ij}, 0). \tag{3.10}$$

Algorithm 2 UGLasso

Require: $\mathbf{S}_1, \ldots, \mathbf{S}_p, \lambda, \alpha$

Compute $\bar{\mathbf{S}} \leftarrow \frac{1}{p}\sum_{i=1}^{p}\mathbf{S}_i$.
Infer $\hat{\mathbf{\Theta}}_i \leftarrow \text{GLasso}(\mathbf{S}_i, \lambda)$ for $i \in [1, p]$.
Let $g(\mathbf{\Theta}) \leftarrow -\log\det\mathbf{\Theta} + \text{tr}(\bar{\mathbf{S}}\mathbf{\Theta}) + \frac{\alpha}{p}\sum_{i=1}^{p}||\mathbf{\Theta} - \hat{\mathbf{\Theta}}_i||_F^2$
Solve Eq. (3.5) to get $\hat{\mathbf{\Theta}}$ using Algorithm 1.
Return $\hat{\mathbf{\Theta}}$

For the selection of step size s in Eq. (3.9), we employ non-monotonic Armijo backtracking line search

Algorithm 3 CGLasso

Require: $\mathbf{S}_1^{(A)}, \ldots, \mathbf{S}_p^{(A)}, \mathbf{S}_1^{(B)}, \ldots, \mathbf{S}_q^{(B)}, \lambda$

Compute $\hat{\mathbf{S}} \leftarrow \frac{1}{p}\sum_{i=1}^{p}\mathbf{S}_i^{(A)} - \frac{1}{q}\sum_{i=1}^{q}\mathbf{S}_i^{(B)}$.
Let $g(\mathbf{\Theta}) \leftarrow \text{tr}(\hat{\mathbf{S}}\mathbf{\Theta})$
Solve Eq. (3.5) to get $\hat{\mathbf{\Theta}}$ using Algorithm 1.
Return $\hat{\mathbf{\Theta}}$

[8], which accepts the largest s that satisfies

$$f(\tilde{\mathbf{\Theta}}^{(i+1)}) \leq \max_{k=i-j:i}\Big(f(\tilde{\mathbf{\Theta}}^{(k)})\Big) + \eta\nabla f(\tilde{\mathbf{\Theta}}^{(i)}))^{\top}(\tilde{\mathbf{\Theta}}^{(i+1)} - \tilde{\mathbf{\Theta}}^{(i)}) \tag{3.11}$$

where $\eta \in (0, 1)$ is the sufficient decrease parameter (usually small) and j is the reference memory parameter typically set as 10. Armijo backtracking line search does not always decrease the objective function, but it can ensure the global convergence of the projected gradient method as well as enhance the convergence rate [6]. We also use Barzilai-Borwein initialization proposed in [3] to setup the step size in the k-th iteration:

$$s_0^{(k)} = \frac{(\mathbf{w}^{(k)})^{\top}(\tilde{\mathbf{\Theta}}^{(k)} - \tilde{\mathbf{\Theta}}^{(k-1)})}{(\mathbf{w}^{(k)})^{\top}\mathbf{w}^{(k)}} \tag{3.12}$$

where $\mathbf{w}^{(k)} = \nabla f(\tilde{\mathbf{\Theta}}^{(\mathbf{k})}) - \nabla f(\tilde{\mathbf{\Theta}}^{(\mathbf{k-1})})$.

Now we recall the positive-definite constraint. We first note that it is obvious the projection in Eq. (3.9) with $\mathcal{C} = \{\mathbf{y} : \mathbf{y} > 0\}$ does not affect the positive-definiteness of a matrix, *i.e.*, if $\mathbf{\Theta}$ is a positive-definite matrix, then $\mathcal{P}_{\mathcal{C}}(\mathbf{\Theta})$ is also positive-definite. Thus, to guarantee that our algorithm always find a positive-definite solution, we need prove the following theorem.

THEOREM 3.1. *For any symmetric matrix $\mathbf{\Theta} \succ 0$ and symmetric $\mathbf{D}$, there exists an $\bar{\alpha} > 0$ such that for all $\alpha < \bar{\alpha}$ the matrix $\mathbf{\Theta} - \alpha\mathbf{D} \succ 0$.*

Proof. We first let $\hat{\mathbf{D}} = -\mathbf{D}$, where $\hat{\mathbf{D}}$ is also symmetric. Let $\sigma_{min}(\mathbf{\Theta})$ to be the smallest eigenvalue of $\mathbf{\Theta}$. When $\alpha < \frac{\sigma_{min}(\mathbf{\Theta})}{||\hat{\mathbf{D}}||_2}$, we have $||\alpha\hat{\mathbf{D}}||_2 < \sigma_{min}(\mathbf{\Theta})$. From Lemma 3.1 we can conclude that $\mathbf{\Theta} + \alpha\mathbf{D} = \mathbf{\Theta} - \alpha\hat{\mathbf{D}} \succ 0$

LEMMA 3.1. *For any symmetric matrix $\mathbf{\Theta} \succ 0$ and symmetric $\mathbf{D}$, if $||\mathbf{D}||_2 < \sigma_{min}(\mathbf{\Theta})$ then $\mathbf{\Theta} + \mathbf{D} \succ 0$.*

Proof. Since $\mathbf{\Theta} \succ 0 \iff \mathbf{z}^{\top}\mathbf{\Theta}\mathbf{z} > 0, \forall\mathbf{z} \neq \mathbf{0}$, with the spectral theorem, we have $\mathbf{z}^{\top}\mathbf{\Theta} \geq \sigma_{min}(\mathbf{\Theta})||\mathbf{z}||_2^2$. By using Cauchy-Schwartz's inequality, we have $|\mathbf{z}^{\top}\mathbf{D}\mathbf{z}| \leq ||\mathbf{D}||_2||\mathbf{z}||_2^2, \forall\mathbf{z}$. Hence, it implies $\mathbf{z}^{\top}\mathbf{D}\mathbf{z} \geq -|\mathbf{z}^{\top}\mathbf{D}\mathbf{z}| \geq -||\mathbf{D}||_2||\mathbf{z}||_2^2$. Therefore,

$$\begin{aligned}\mathbf{z}^{\top}(\mathbf{\Theta} + \mathbf{D})\mathbf{z} &= \mathbf{z}^{\top}\mathbf{\Theta}\mathbf{z} + \mathbf{z}^{\top}\mathbf{D}\mathbf{z} \\ &\geq \sigma_{min}(\mathbf{\Theta})||\mathbf{z}||_2^2 - ||\mathbf{D}||_2||\mathbf{z}||_2^2 \\ &= \big(\sigma_{min}(\mathbf{\Theta}) - ||\mathbf{D}||_2\big)||\mathbf{z}||_2^2 > 0, \forall\mathbf{z} \neq \mathbf{0}.\end{aligned}$$

since $||\mathbf{D}||_2 < \sigma_{min}(\mathbf{\Theta})$. So $\mathbf{\Theta} + \mathbf{D} \succ 0$ holds.

Accordingly, given $\hat{\mathbf{\Theta}}^{(i)}$ in iteration i that is positive-definite, the line search process can always find a step size $s > 0$ that make $\hat{\mathbf{\Theta}}^{(i+1)} = \mathcal{P}_{\mathcal{C}}(\tilde{\mathbf{\Theta}}^{(i)} - s\nabla f(\tilde{\mathbf{\Theta}}^{(i)}))$ to be positive-definite as well. The modified projected gradient algorithm for solving Eq. 3.7 is summarized in Algorithm 1. The UGLasso and CGLasso algorithms are summarized in Algorithm 2 and Algorithm 3 respectively.

4 Simulated Study on Synthetic Data

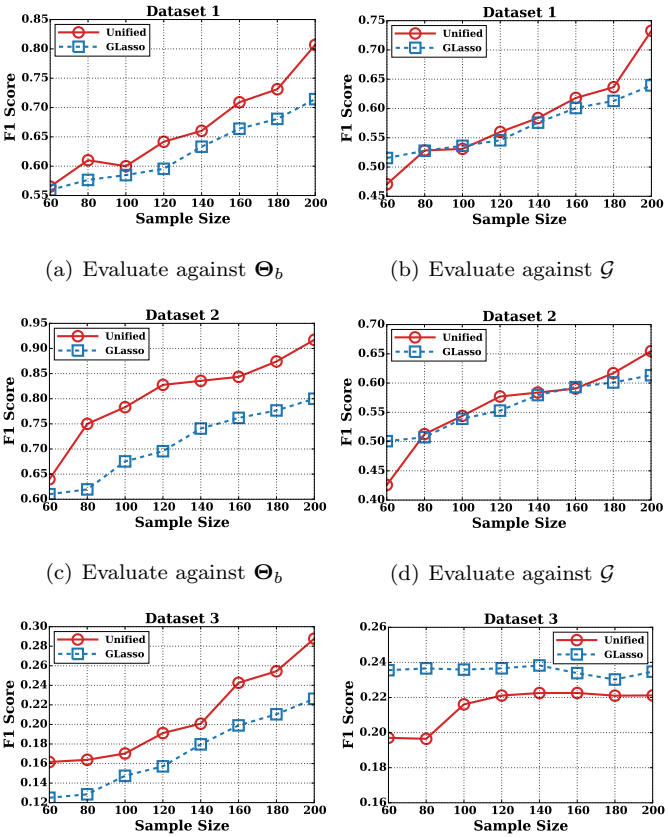
Due to the lack of ground truth in real neuroimaging datasets, synthetic data is considered to be an important tool to evaluate the effectiveness of brain network discovery methods. We first evaluate our methods on synthetic data where ground truth (network structure) is available.

4.1 Evaluation on UGLasso

• *Dataset:* The first set of synthetic data is generated to comparing the effectiveness of the proposed unified graphical Lasso method with GLasso. We adopt the approach in [16] with some modifications to generate the synthetic precision matrices. To simulate subjects in single collection, we generate p separate sparse precision matrices of size $m \times m$ with similar structure. Specifically, the first step is to randomly generate a basal positive definite matrix $\mathbf{\Theta}_b$ of size $m \times m$, where we control the density of $\mathbf{\Theta}_b$ to be $\rho_b \in (0, 1)$. Then we generate p different positive definite noise matrices $\{\mathbf{N}_1, \ldots, \mathbf{N}_p\}$, each of size $m \times m$ and density ρ_n. At last we add each noise matrix to the basal matrix respectively to get the collection of ground truth matrices $\mathcal{G} = \{\mathbf{G}_1, \ldots, \mathbf{G}_p\}$, where $\mathbf{G}_i = \mathbf{\Theta}_b + \mathbf{N}_i$. By doing so, we retain the positive definiteness of each ground truth matrices as well as control the similarity among them. With the collection of ground truth matrices, we can draw p separate sample matrices of size $n \times m$ from the Gaussian distribution for each subject to simulate the fMRI signals, where n denotes the number of samples (or the number of time steps in fMRI). Without losing generality, we simply use the same n for all subjects in the collection. To numerically evaluate the compared methods, we prepare three synthetic datasets with following parameters:

- **Dataset 1 (Weak Noises)** : $m = 50, p = 50, \rho_b = 0.01, \rho_n = 0.005, n = 60, 80, \ldots, 200$.
- **Dataset 2 (Moderate Noises)**: $m = 50, p = 100, \rho_b = 0.01, \rho_n = 0.01, n = 60, 80, \ldots, 200$.
- **Dataset 3 (Strong Noises)**: $m = 50, p = 100, \rho_b = 0.01, \rho_n = 0.05, n = 60, 80, \ldots, 200$.

• *Experimental Protocols:* We follow the approach as described above to obtain the collection of ground truths $\mathcal{G}$. For each choice of sample size n, we randomly draw a collection of sample matrices $\mathcal{X}^{(m \times n)} = \{\mathbf{X}_1^{(m \times n)}, \ldots, \mathbf{X}_p^{(m \times n)}\}$ from Gaussian distribution based on $\mathcal{G}$. Then the empirical covariance matrix $\mathbf{S}$ for the collection can be computed using the stack approach described in Section 2. GLasso uses $\mathbf{S}$ to estimate the precision matrix for the collection, and UGLasso uses both $\mathbf{S}$ and $\mathcal{X}^{(m \times n)}$. To be fair, we set the parameters for both methods to make the estimated matrices to have similar number of nonzero entities. We repeat this process 5 times for each choice of n.



(a) Evaluate against $\mathbf{\Theta}_b$ (b) Evaluate against $\mathcal{G}$

(c) Evaluate against $\mathbf{\Theta}_b$ (d) Evaluate against $\mathcal{G}$

(e) Evaluate against $\mathbf{\Theta}_b$ (f) Evaluate against $\mathcal{G}$

Figure 3: Comparison between UGLasso and GLasso on three synthetic datasets in terms of F1 score on connectivity inference.

• *Evaluation Metrics:* We follow [16] to define the F1 score of connectivity inference as $F1 = \frac{2n_d^2}{n_a n_d + n_g n_d}$, where n_d is the number of true edges detected by the algorithm, n_g is the number of true edges and n_a is the total number of edges detected. Larger F1 score is better. In our experiments we have a collection of ground truth matrices $\mathcal{G}$ and a single unified basal

ground truth matrix $\mathbf{\Theta}_b$. To evaluate the performance of finding representative networks, we report two F1 scores in each experiment. The first one is obtained by evaluating the inferred network against $\mathbf{\Theta}_b$, where the noises are excluded. This evaluation aims at assessing the ability of recovering the *real* representative structure from noisy signals. The second one is obtained by evaluating the inferred network against each network in $\mathcal{G}$, where noises are not excluded in the evaluation.

• *Results Analysis:* The results on synthetic dataset 1-3 are demonstrated in Fig. 3, where we compare the proposed UGLasso with GLasso in terms of F1 score. The left column of Fig. 3 shows the evaluation against the basal truth matrix $\mathbf{\Theta}_b$. The right column of Fig. 3 shows the evaluation against noisy ground truth matrices $\mathcal{G}$. We have following observations.

- UGLasso recovers the basal truth network better than GLasso (Fig. 3(a), 3(c), 3(e)) consistently.
- When we include the noises in evaluation, UGLasso achieves competitive performance compared to GLasso, and it usually outperforms GLasso when the number of samples $n > 100$ (Fig. 3(b), 3(d)).
- Due to the existence of strong noises in dataset 3, UGLasso is outperformed by GLasso (Fig. 3(f)), where the ground truths contain much more noises than authentic signals. It is likely that GLasso over-fits the noises to achieve higher F1 score in this case.

4.2 Evaluation on CGLasso • *Dataset:* The second set of synthetic data is generated to comparing CGLasso with GLasso. As in the contrasting setting where we have two collections of subjects, we need generate a ground truth matrix for each collection. Since we do not enforce the inferred network to be similar to any individual network, we simply use a precision matrix to represent the ground truth network of a collection. To make the network easy to visualize, we divide the $m \times m$ matrix into several $l \times l$ square blocks. Then we randomly select some blocks (symmetrically) to fill in values, and leave the rest all 0s. At last we add some random noises to the matrix. we generate three synthetic datasets with parameters as follows:

- **Dataset 4**: $m = 50, l = 10, n = 200$. The generated ground truth for collection (A) and (B) are shown in Fig. 4(a) and Fig. 4(b) respectively.
- **Dataset 5**: $m = 50, l = 5, n = 200$. The generated ground truth for collection (A) and (B) are shown in Fig. 4(f) and Fig. 4(g) respectively.
- **Dataset 6**: Same parameters as Dataset 5, different random seed is used. The generated ground truth for collection (A) and (B) are shown in Fig. 4(k) and Fig. 4(l) respectively.

• *Experimental Protocols:* We compare the inferred discriminative network derived by CGLasso with the two inferred networks for collection (A) and collection (B) derived by standard GLasso. We choose the same value of λ for both methods in all experiments.

• *Results Analysis:* Since there is no standard protocol to evaluate contrasting inference, we demostrate the ground truths and inferred networks for synthetic datasets 4-6 in Fig. 4. The ground truth precision matrices for collection (A) and (B) are shown in the first column and the second column respectively; the difference of the ground truths between two collections is shown in the third column; the discriminative network inferred by the proposed contrasting GLasso is shown in the fourth column; at last, the fifth column shows the network structure of $\hat{\mathbf{\Theta}}^{(A)} - \hat{\mathbf{\Theta}}^{(B)}$, where $\hat{\mathbf{\Theta}}^{(A)}$ and $\hat{\mathbf{\Theta}}^{(B)}$ denote the precision matrix inferred for collection (A) and collection (B) by GLasso respectively. One can observe that CGLasso captures the differences between two collections fairly good with less noises compared to the ones derived by GLasso ($\hat{\mathbf{\Theta}}^{(A)} - \hat{\mathbf{\Theta}}^{(B)}$). Besides, one can also observe that GLasso has much more false positive in than CGLasso does. These results demonstrate that CGLasso is a more suitable tool for discriminative network discovery.

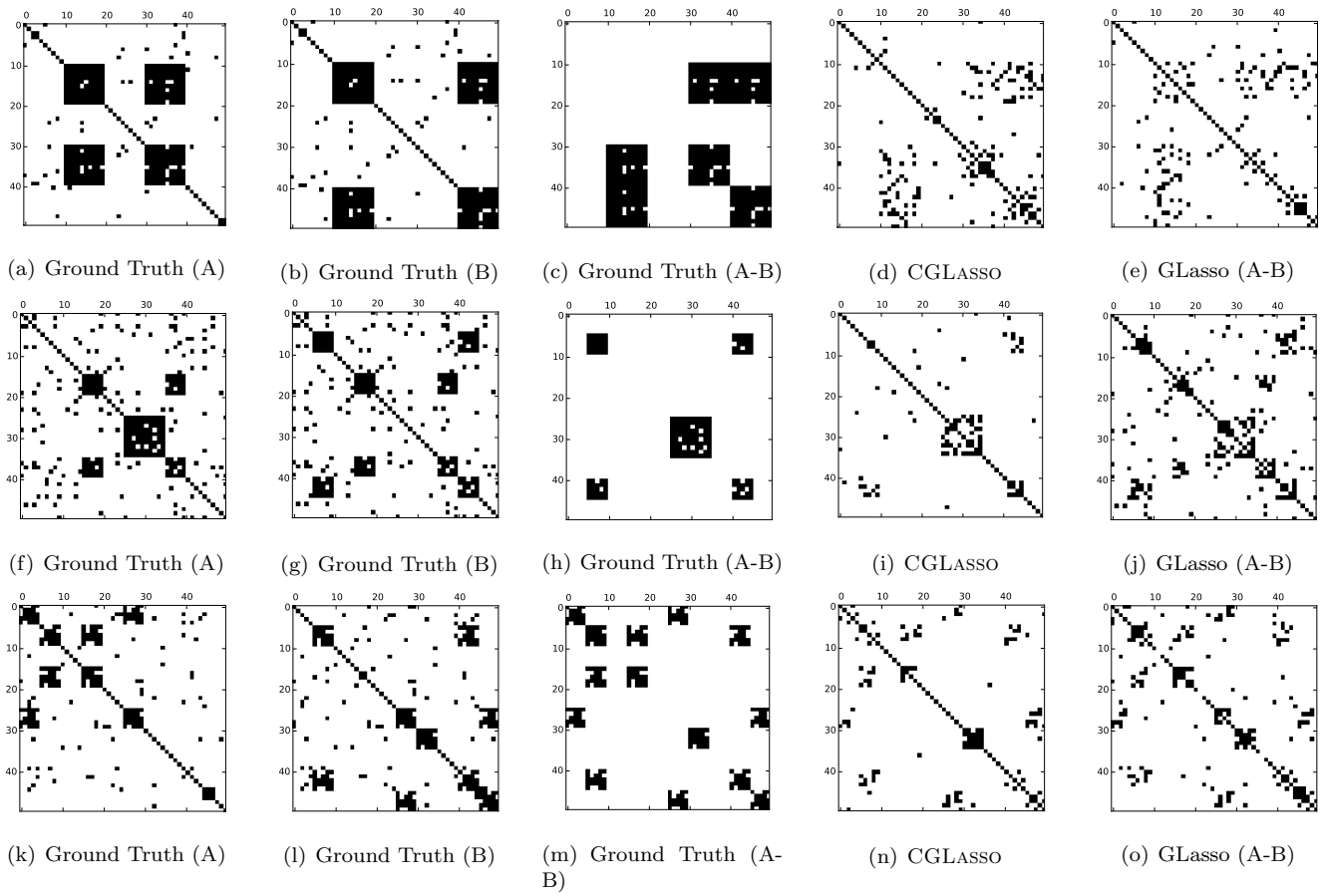
5 Real Data

5.1 Data Collection • *Alzheimer's Disease (ADNI)*: The first dataset is collected from the Alzheimer's Disease Neuroimaging Initiative[1]. The dataset consists of records of patients with Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI). We downloaded all records of resting-state fMRI images and treated the normal brains as *healthy* subjects, and AD+MCI as the *morbid* subjects. We applied Automated Anatomical Labeling (AAL[2]) to extract a sequence of responds from each of the 116 anatomical volumes of interest (AVOI), where each AVOI represents a different brain region. We keep 90 cerebral regions, excluding 26 cerebellar regions. We follow the same preprocess steps in [11] to obtain the cleaned time series data.

• *Human Immunodeficiency Virus Infection (HIV)*: The second dataset is collected from the Chicago Early HIV Infection Study in Northwestern University [15]. The

[1] http://adni.loni.ucla.edu/

[2] http://neuro.imm.dtu.dk/wiki/Automated_Anatomical_Labeling



(a) Ground Truth (A) (b) Ground Truth (B) (c) Ground Truth (A-B) (d) CGLasso (e) GLasso (A-B)

(f) Ground Truth (A) (g) Ground Truth (B) (h) Ground Truth (A-B) (i) CGLasso (j) GLasso (A-B)

(k) Ground Truth (A) (l) Ground Truth (B) (m) Ground Truth (A-B) (n) CGLasso (o) GLasso (A-B)

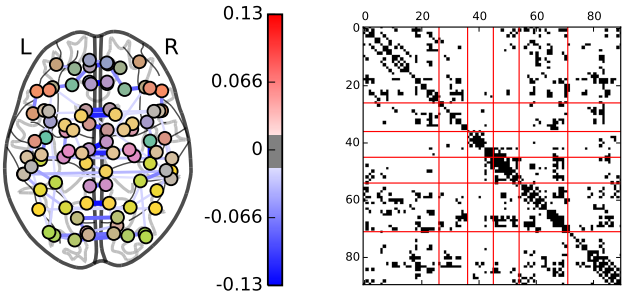
Figure 4: Comparison between Contrasting Graphical Lasso and GLasso on three synthetic datasets.

dataset contains fMRI brain images of patients with early HIV infection (morbid) as well as normal controls (healthy). The same preprocessing steps as in ADNI dataset were used to extract the cleaned time series.

5.2 Results and Discussion • *Unified Setting*: The results of UGLasso on HIV data are shown in Fig. 5 and Fig. 7. The results for unified graphical Lasso of on HIV data are shown in Fig. 8 and Fig. 9. In all figures, we present the visualization of brain connectivity on the left and the corresponding precision matrix on the right. In each precision matrix, the diagonal blocks are referring to prefrontal lobe, other parts of frontal lobe, corpus striatum, occipital lobe, parietal lobe and temporal lobe respectively. All results are derived using $\lambda = 2.0$ and $\alpha = 0.5$.

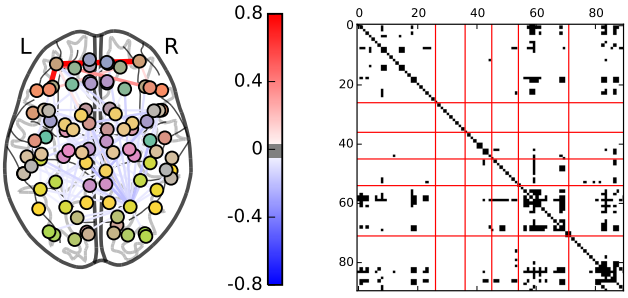
By comparing Fig. 5 and Fig. 7, we observe that the overall interconnection between different lobes are weaker in AD patients than the ones in healthy people. These degeneration may explain the AD symptoms such as difficulty thinking and understanding, confusion in the evening. Because understanding and sense of timing usually require the collaboration of several regions in brain, with degenerated connectivity between lobes, AD patients may not function well as normal people. By comparing Fig. 8 and Fig. 9, we observe that HIV patients have increased connectivity inside occipital lobe compared to the healthy people, which is consistent with previous studies [4]. This may be explained by that people infected by HIV usually vision problems and the major functions of occipital lobe are receives visual information and interprets color, shape and distance [1]. • *Contrasting Setting*: Here we attempt to see if the CGLasso finds any reasonable discriminative patterns between the healthy collection and the morbid collection in real fMRI data. The inferred networks for ADNI dataset and HIV datasets are illustrated in Fig. 6 and Fig. 10. For ADNI dataset, as can be seen from Fig. 6, the major differences between AD patients and normal people are located in parietal lobe and temporal lobe, which is consistent with previous studies [9, 16]. Strong decrease of connectivity in these lobes have been detected for AD patients before, which explains the symptoms such as memory loss, mental confusion *etc.*

Besides, we also observe a strong connection between "Frontal_Sup_Orb_L" and "Frontal_Sup_Orb_R"



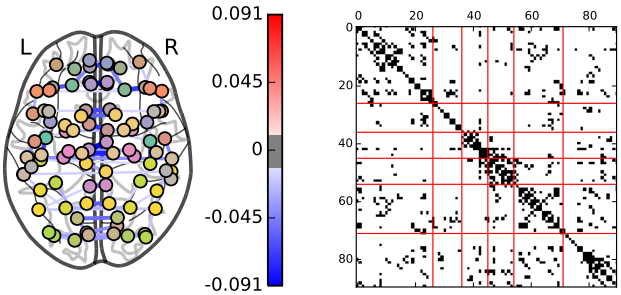
(a) ADNI dataset (b) ADNI dataset

Figure 5: Results of UGLasso for the healthy collection in ADNI data. The diagonal blocks in Fig. (b) refers to prefrontal lobe, other parts of frontal lobe, corpus striatum, occipital lobe, parietal lobe and temporal lobe respectively (Same for Figure 6-10).



(a) ADNI dataset (b) ADNI dataset

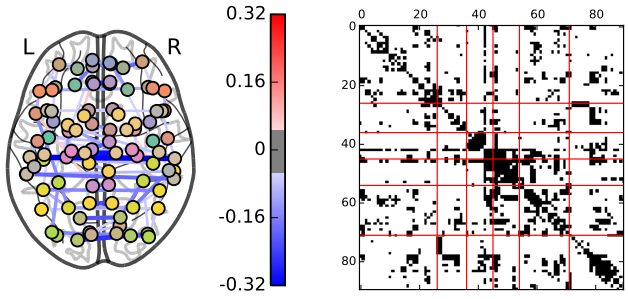
Figure 6: Results of CGLasso for ADNI data.



(a) ADNI dataset (b) ADNI dataset

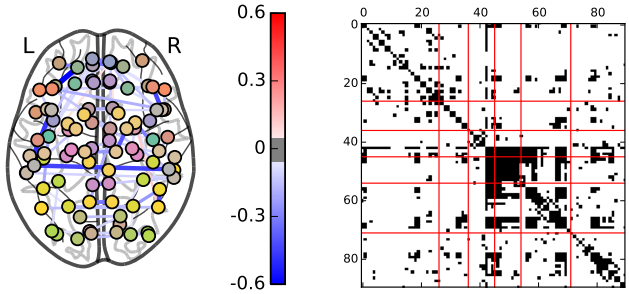
Figure 7: Results of UGLasso for the morbid collection in ADNI data.

in Figure 6(a) (the red line across left hemisphere and right hemisphere on the top), indicating that AD patients exhibit a significant different patterns toward the activity between these two regions in frontal lobe. Previous studies [14] sometimes exclude the frontal lobe in analysis since it is thought to be unrelated to AD. However, recent works show there exists increased connectivity in the frontal lobe of AD patients [16]. CGLasso also reveals such pattern in the frontal lobe.



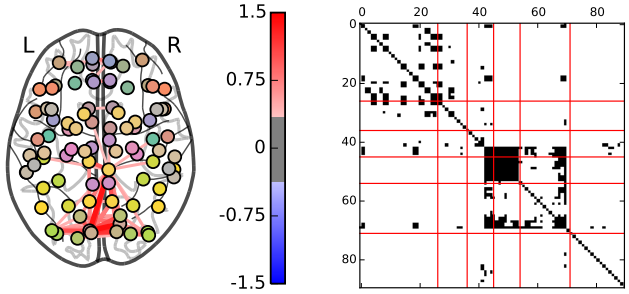
(a) HIV dataset (b) HIV dataset

Figure 8: Results of UGLasso for the healthy collection in HIV data.



(a) HIV dataset (b) HIV dataset

Figure 9: Results of UGLasso for the morbid collection in HIV Data.



(a) HIV dataset (b) HIV dataset

Figure 10: Results of CGLassofor HIV data.

As to the HIV dataset, from Fig. 10 we can observe that the major differences between HIV patients and healthy people are located in occipital lobe and part of parietal lobe, which is also supported by previous studies [4]. Several connections within occipital lobe are proved to be discriminative subgraph patterns that are considered to be associated with HIV. No connection is detected in temporal lobe for HIV dataset under contrasting setting, this may because HIV patients do not exhibit the mental symptoms as AD patients do.

6 Related Works

To the best of our knowledge, this paper is the first work exploring the brain network discovery under unified

and contrasting settings. Our work is related to brain network discovery and contrasting learning.

6.1 Brain Network Discovery Most works in this line focus on finding a network representation using sparse Gaussian graphical model (sGGM). Banerjee et al. [2] first formulated the problem of sparse maximum likelihood estimation, where they assumed that the multi-variate variables follow a certain multi-variant Gaussian distribution. Friedman et al. [7] reform the dual problem of Eq. (2.1) as a Lasso-type problem and apply the model on graphs, their method is widely referred to as GLasso. Sun et al. [14] and Huang et al. [10] utilize sGGM to infer networks for three collections of subjects related to Alzheimer's disease, where they treat each collection as a single subject. Davidson et al. [5] propose a supervised tensor-based framework to infer both brain regions and brain connectivity from fMRI data, where strong domain knowledges are required in training. Yang et al. [16] formulate a variant of GLasso called fused multiple graphical Lasso (FMGL) to derive p networks for p similar collections of subjects. FMGL is closely related to our contrasting setting, but with several major differences: (i) FMGL usually infer p separate networks for p collections where $p > 2$ while contrasting network discovery infers a single network between two collections; (ii) FMGL assumes that one can order the p collections properly where neighbored collections share similar network structure, while contrasting network discovery does not. (iii) For $p = 2$, FMGL inferres two similar networks.

6.2 Contrast Learning Contrast learning aims at finding discriminative patterns between classes of data. Ramamohanarao et al. [13] study the problem of extracting subgraphs that is frequent in one database but infrequent in another database. Kuo et al. [12] propose to finding a contrasting cut in two collections of graphs, where the cut has a low cost for one collection but has a high cost for the other collection. The problem proposed in [12] is similar to our contrasting network discovery, but they aims at learning the imaging segmentation while we aims at learning the connectivities.

7 Conclusion

Neuroimaging analysis usually involves one or more collections of subjects, *e.g.* healthy collection v.s. morbid collection. In this paper we explore the problems of unified network discovery on a collection of subjects and contrasting network discovery on two collections of subjects. Two novel algorithms, UGLasso and CGLasso, are proposed to solve them respectively. Extensive experiments conducted on synthetic datasets and real-world datasets demonstrate the outstanding performance and usefulness of the proposed methods.

References

[1] Brain map and functions. `https://www.rah.sa.gov.au/birs/bi_brain.php`.

[2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

[3] J. Barzilai and J. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[4] B. Cao, X. Kong, J. Zhang, P. Yu, and A. Ragin. Mining brain networks using multiple side views for neurological disorder identification. In *ICDM*, 2015.

[5] I. Davidson, S. Gilpin, O. Carmichael, and P. Walker. Network discovery via constrained tensor analysis of fmri data. In *KDD*, 2013.

[6] M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *Selected Topics in Signal Processing, IEEE Journal of*, 1(4):586–597, 2007.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[8] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

[9] B. Horwitz, C. Grady, NL. Schlageter, R. Duara, and SI. Rapoport. Intercorrelations of regional cerebral glucose metabolic rates in alzheimer's disease. *Brain research*, 407(2):294–306, 1987.

[10] S. Huang, J. Li, L. Sun, J. Liu, T. Wu, K. Chen, A. Fleisher, E. Reiman, and J. Ye. Learning brain connectivity of alzheimer's disease from neuroimaging data. In *NIPS*, pages 808–816, 2009.

[11] X. Kong, A. Ragin, X. Wang, and P. Yu. Discriminative feature selection for uncertain graph classification. In *SDM*, 2013.

[12] C.T. Kuo, X. Wang, P. Walker, O. Carmichael, J. Ye, and I. Davidson. Unified and contrasting cuts in multiple graphs: Application to medical imaging segmentation. In *KDD*, 2015.

[13] K. Ramamohanarao, J. Bailey, and H. Fan. Efficient mining of contrast patterns and their applications to classification. In *ICISIP*, pages 39–47. IEEE, 2005.

[14] L. Sun, R. Patel, J. Liu, K. Chen, T. Wu, J. Li, E. Reiman, and J. Ye. Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation. In *KDD*, 2009.

[15] X. Wang, P. Foryt, R. Ochs, J. Chung, Y. Wu, T. Parrish, and A. Ragin. Abnormalities in resting-state functional connectivity in early human immunodeficiency virus infection. *Brain Connectivity*, 1(3):207, 2011.

[16] S. Yang, Z. Lu, X. Shen, P. Wonka, and J. Ye. Fused multiple graphical lasso. *SIAM Journal on Optimization*, 25(2):916–943, 2015.