Discriminative Sparse Inverse Covariance Matrix: Application in Brain Functional Network Classification

Luping Zhou, Lei Wang, and Philip Ogunbona
School of Computer Science and Software Engineering, University of Wollongong, Australia
lupingz, leiw, philipo@uow.edu.au

Abstract

Recent studies show that mental disorders change the functional organization of the brain, which could be investigated via various imaging techniques. Analyzing such changes is becoming critical as it could provide new biomarkers for diagnosing and monitoring the progression of the diseases. Functional connectivity analysis studies the covary activity of neuronal populations in different brain regions. The sparse inverse covariance estimation (SICE), also known as graphical LASSO, is one of the most important tools for functional connectivity analysis, which estimates the interregional partial correlations of the brain. Although being increasingly used for predicting mental disorders, SICE is basically a generative method that may not necessarily perform well on classifying neuroimaging data. In this paper, we propose a learning framework to effectively improve the discriminative power of SICEs by taking advantage of the samples in the opposite class. We formulate our objective as convex optimization problems for both one-class and two-class classifications. By analyzing these optimization problems, we not only solve them efficiently in their dual form, but also gain insights into this new learning framework. The proposed framework is applied to analyzing the brain metabolic covariant networks built upon FDG-PET images for the prediction of the Alzheimer's disease, and shows significant improvement of classification performance for both one-class and two-class scenarios. Moreover, as SICE is a general method for learning undirected Gaussian graphical models, this paper has broader meanings beyond the scope of brain research.

1. Introduction

Modern neuroscience has greatly benefited from numerous imaging techniques that enable non-invasive investigations into the anatomy and functions of the brain. In recent years, we have witnessed an increasing interest and pace of development in imaging-based brain connectivity analysis ([2, 3]), which aims at studying how anatomically segregated brain regions are functionally connected for cognitive or other tasks. In such analysis, brain images are partitioned into "nodes" of interest and the connections between nodes are estimated from imaging-based features. Brain connectivity analysis has been extensively employed in analyzing mental disorders, such as the Alzheimer's disease [6, 12, 7] and Schizophrenia [4], among others. When affected by these diseases, some interregional brain connections may be interrupted and the brain may be reorganized. Mining such changes from neuroimaging data can provide crucial biomarkers for the diagnosis of the diseases, and enrich our knowledge about the disease mechanisms.

One important type of brain connectivity, known as functional connectivity, studies the covary activity of neuronal populations in different brain regions. A large body of research work models the functional connectivity by correlation-based statistics [12, 7, 6], and has been reported in [10] to be relatively more sensitive for detecting network connections than lag-based (such as Granger causality or dynamic causal models) or higher-order statistic-based methods. Early methods of correlation analysis use sample-based convariance matrix to estimate the pair-wise regional correlation, which cannot factor out the effects of other regions and hence has been gradually replaced by partial correlations. Partial correlations are usually estimated via the inverse covariance matrix (ICOV), as they correspond to the off-diagonal entries of the ICOV [8]. However, the accuracy of the maximum-likelihood estimation (MLE) of ICOV is notoriously sensitive to the number of available samples. To deal with this problem, in [7] it is proposed to regularize MLE with network sparsity by imposing a constraint to the l-1 norm of the ICOV. This method is termed sparse inverse covariance estimation (SICE), or called graphical LASSO [5, 13] in pattern recognition. By controlling the number of zero entries, SICE allows a reliable estimation of inverse covariance matrix even when the number of samples is small. This favorable property makes SICE the method of choice to learn the structure of undirected Gaussian graphical models. Incidentally, such models are required in the analysis of brain functional connectivity.

Although SICE is often used to investigate the predominant functional connectivity patterns in the studied populations, there is an evident increase of applications that use functional connectivity as an endophenotype for the prediction / classification of mental disorders [12, 7]. Two kinds of approaches have been employed for this purpose: i) extracting graph-based features (such as the local clustering coefficients) from SICE so that these features could be used by classifiers such as Support Vector Machines (SVMs) [12]; and ii) directly using SICE for classification [7]. The success of both approaches depends on whether the SICE contains sufficient discriminative information. However, in its original formulation SICE is a generative method that learns the ICOV to model a single class rather than discriminating between two classes. Therefore, some subtle but critical network differences may be inappropriately ignored, leading to inferior classification performance.

In this paper, we propose a learning framework to achieve SICE with improved discriminative power and apply it for brain network classification tasks. To the best of our knowledge, this is the first time discriminative learning has been incorporated into SICE. In a broader sense, our work has significance beyond the scope of brain research, because SICE is a widely used model in pattern recognition. This paper has the following contributions. First, we propose a learning framework to improve the discriminative power of SICE by taking advantage of the availability of samples in the opposite class. We formulate our discriminative learning objectives as convex optimization problems for both one-class (with negative samples available) and two-class classifications. Second, we carefully study our optimization problems and demonstrate how to efficiently solve them in their dual form. By converting the primal problems to dual problems, we can take full advantage of existing efficient SICE solvers for our discriminative learning tasks. Moreover, formulating in dual form also gives us more insights into our proposed learning framework. Third, we apply our proposed framework to analyzing brain metabolic covariant networks constructed from FDG (fluorodeoxyglucose)-PET (Positron Emission Tomography) images for the prediction of Alzheimer's disease. Our learned SICEs show significant improvement of discriminative power with better classification performances in both one-class and two-class classification tasks, indicating the effectiveness of our method.

2. Background

2.1. From image to brain connectivity

Decades of neuroimaging research shows that many mental disorders are associated with subtle abnormalities distributed over the brain rather than a damage of an individual brain region, implying the alteration of interactions between neuronal systems. The interactions of brain regions (referred as brain connectivity or brain network) could be studied by brain images at macroscopical level, such as functional magnetic resonance imaging (fMRI) and PET, with theoretical graphical models. The nodes of the network model correspond to clustered imaging voxels that are determined by predefined brain parcellations (atlas) or statistical methods such as the independent component analysis, etc. Each node is characterized by features extracted from the corresponding voxels, for example the averaged radiotracer uptake for PET images. There could be three types of connections between two nodes, namely the anatomical, functional or effective connectivity. In this paper, we focus on the functional connectivity, which is basically a statistical concept, measuring the covary patterns of brain regions. In functional connectivity analysis, the connection between two nodes is commonly determined by the correlation or partial correlation of their imaging-based features. The latter is more advantageous because when the partial correlation becomes zero, the corresponding two regions are conditionally independent given all other regions considered.

2.2. Sparse inverse covariance estimation

Partial correlation corresponds to the off-diagonal entries of ICOV. A reliable estimation of partial correlation calls for a reliable estimation of ICOV. The latter however often requires a sufficiently large number of training samples. Here we briefly review a representative work termed as sparse inverse covariance estimation (SICE), or graphical LASSO, which can circumvent this problem by introducing l-1 norm regularization into the estimation of ICOV. SICE is proposed as a general pattern recognition method [5], which is introduced to analyze brain network in [7].

Let $\mathbf{X}$ denote a set of n samples composed of imaging-based features from p brain regions, and $\mathbf{\Theta}$ the inverse covariance matrix to be estimated from $\mathbf{X}$. Assuming the samples following normal distribution, SICE solves

$$\min_{\mathbf{\Theta} \succeq 0} \; -\log|\mathbf{\Theta}| + \mathrm{tr}(\mathbf{S\Theta}) + \lambda\|\mathbf{\Theta}\|_1, \tag{1}$$

where $\mathbf{S}$ is the sample-based covariance matrix estimated from $\mathbf{X}$. The symbols $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ denote the determinant and the trace of a matrix. The symbol $\|.\|_1$ denotes the sum of the absolute values of all entries in a matrix, and λ is a user-defined parameter. Minimizing $-\log|\mathbf{\Theta}| + \mathrm{tr}(\mathbf{S\Theta})$ maximizes the log-likelihood of $\mathbf{\Theta}$, which is further regularized by the sparseness requirement of $\mathbf{\Theta}$ through minimizing $\|\mathbf{\Theta}\|_1$. The constraint $\mathbf{\Theta} \succeq 0$ ensures $\mathbf{\Theta}$ to be positive semi-definite.

The sparseness regularization used in SICE correlates with the fact that one brain region predominately interacts with only a small number of other regions. SICE can better recover the zero entries in $\mathbf{\Theta}$ than the maximum-likelihood, and becomes a key tool for structure estimation of undirected Gaussian graphical models. Eqn.(1) is a convex optimization problem. Various methods have been developed to efficiently solve it, such as [13, 5, 1], just name a few.

3. Proposed methods

As mentioned above, although the inverse covariance matrix has been increasingly used as an endophenotype for predicting or classifying the patient group with mental disorders, SICE is basically a generative method, focusing on modeling instead of discriminating the data. Therefore it may not perform satisfactorily in classification tasks, especially for subtle but critical group differences, which, however, is often encountered in neuroimaging data. In this paper, we propose a learning framework to improve the discriminative power of SICEs for better classification. Our framework covers the scenarios of both one-class (with some negative samples) and two-class classifications, explained in the following sections, respectively.

Some symbols and notations that are often used in this paper are defined as follows.

Let $f_{\mathrm{SICE}}(\mathbf{\Theta}_a) = -\log|\mathbf{\Theta}_a| + \mathrm{tr}(\mathbf{S}_a\mathbf{\Theta}_a) + c_1\|\mathbf{\Theta}_a\|_1$ denote the objective function to learn the sparse inverse covariance matrix $\mathbf{\Theta}_a$ for class a. $\mathbf{S}_a$ is the sample-based covariance matrix defined as $\mathbf{S}_a = n_a^{-1}\sum_{i=1}^{n_a}(\mathbf{x}_a^i - \boldsymbol{\mu}_a)(\mathbf{x}_a^i - \boldsymbol{\mu}_a)^\top$, where $\mathbf{x}_a^i$ denotes the ith sample in class a and $\boldsymbol{\mu}_a$ is the sample-based class mean. Let $(\mathbf{x}_a^i - \boldsymbol{\mu}_b)^\top\mathbf{\Theta}_b(\mathbf{x}_a^i - \boldsymbol{\mu}_b)$ denote the Mahalanobis distance of $\mathbf{x}_a^i$ from class b. This distance can be written in a compact form as

$$\mathrm{tr}(\mathbf{T}_{a,b}^i\mathbf{\Theta}_b) = (\mathbf{x}_a^i - \boldsymbol{\mu}_b)^\top\mathbf{\Theta}_b(\mathbf{x}_a^i - \boldsymbol{\mu}_b), \tag{2}$$

where $\mathbf{T}_{a,b}^i \triangleq (\mathbf{x}_a^i - \boldsymbol{\mu}_b)(\mathbf{x}_a^i - \boldsymbol{\mu}_b)^\top$.

3.1. Discriminative learning of Θ_a (or Θ_b)

We consider the scenario of one-class classification first. In this scenario, we are only interested in estimating the SICE model for one class, and classify a new subject by estimating its distance (or likelihood) to this class. When some samples in the opposite class are also available, we could improve the discriminative power of the one-class SICE by taking advantage of these negative samples. One-class classification may be preferred than two-class classification when the following concerns arise: i) for one of the two class, there are insufficient samples for estimating a reliable SICE; or ii) the patient group might be too divergent to reasonably follow a Gaussian distribution. One potential application of one-class classification is to detect abnormal subjects from the model purely built upon a large number of healthy subjects.

Without loss of generality, let's assume class a is the "normal" class and class b is the "abnormal" class. We are only interested in estimating $\mathbf{\Theta}_a$ for classification, but taking advantage of the samples in class b to improve the discriminative power of $\mathbf{\Theta}_a$. Our discriminative learning employs the following criteria: i) the samples in class b should be away from the distribution $P(\mathbf{x}|\boldsymbol{\mu}_a, \mathbf{\Theta}_a)$; ii) the samples in class a should be close to $P(\mathbf{x}|\boldsymbol{\mu}_a, \mathbf{\Theta}_a)$; and iii) the estimation of $\mathbf{\Theta}_a$ should respect the distribution of class a. The following objective function is therefore minimized:

$$\begin{aligned} &\min_{\rho,\boldsymbol{\xi},\boldsymbol{\eta},\mathbf{\Theta}_a\succeq 0} f_{\mathrm{SICE}}(\mathbf{\Theta}_a) + c_2\left[c_3\boldsymbol{\xi}^\top\mathbf{1} + c_4\boldsymbol{\eta}^\top\mathbf{1} - \rho\right] \qquad (3)\\ s.t. \quad & \mathrm{tr}(\mathbf{T}_{b,a}^i\mathbf{\Theta}_a) \geq \rho - \xi_i, \quad i = 1, \cdots, n_b\\ & \mathrm{tr}(\mathbf{T}_{a,a}^i\mathbf{\Theta}_a) \leq \rho + \eta_i, \quad i = 1, \cdots, n_a\\ & \xi_i \geq 0, \quad i = 1, \cdots, n_b\\ & \eta_i \geq 0, \quad i = 1, \cdots, n_a. \end{aligned}$$

As shown, we require the Mahalanobis distance of any $\mathbf{x}_b^i$ to class a, i.e., $\mathrm{tr}(\mathbf{T}_{b,a}^i\mathbf{\Theta}_a)$, to be larger than a margin ρ, and the Mahalanobis distance of any $\mathbf{x}_a^i$ to class a, i.e., $\mathrm{tr}(\mathbf{T}_{a,a}^i\mathbf{\Theta}_a)$, to be smaller than the margin ρ. To deal with difficult separation, we employ a soft-margin approach that allows misclassification with the slack variables $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_{n_b}]^\top$ and $\boldsymbol{\eta} = [\eta_1, \cdots, \eta_{n_a}]^\top$. To improve the classification performance, we minimize the misclassification as well as maximizing the separation margin. Meanwhile, we also maximize the log-likelihood of $\mathbf{\Theta}_a$ by minimizing $f_{\mathrm{SICE}}(\mathbf{\Theta}_a)$. The user-defined parameters c_2, c_3 and c_4 balance the corresponding terms in the objective function, which are suggested to be set proportionally to $\frac{1}{\sqrt{n_a\times n_b}}$, $\frac{1}{n_b}$ and $\frac{1}{n_a}$, respectively.

It is not difficult to see that Eqn.(3) is a convex optimization ($f_{\mathrm{SICE}}(\mathbf{\Theta}_a)$ is convex to $\mathbf{\Theta}_a$; $\mathrm{tr}(\mathbf{T}_{b,a}^i\mathbf{\Theta}_a)$ and $\mathrm{tr}(\mathbf{T}_{a,a}^i\mathbf{\Theta}_a)$ in the constraints are linear functions of $\mathbf{\Theta}_a$), which guarantees a global optimality. As is known, a convex optimization could be solved either in its primal or dual form. Here we choose to solve the dual form due to the following reason. In the literature, efficient algorithms have been proposed to minimize $f_{\mathrm{SICE}}(\mathbf{\Theta})$, the objective function of SICE. We find that **when formulating Eqn.(3) into its dual form, we only need to iteratively solve SICE with modified empirical covariance matrices**. This not only allows us to take full advantage of those efficient algorithms, but also gives us more insights into our optimization problem. The dual form of Eqn.(3) is derived as follows.

The Lagrangian $L(\rho, \boldsymbol{\xi}, \boldsymbol{\eta}, \mathbf{\Theta}_a; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ is

$$\begin{aligned} & f_{\mathrm{SICE}}(\mathbf{\Theta}_a) + c_2\left[c_3\boldsymbol{\xi}^\top\mathbf{1} + c_4\boldsymbol{\eta}^\top\mathbf{1} - \rho\right] \qquad (4)\\ + & \sum_{i=1}^{n_b}\alpha_i(\rho - \xi_i - \mathrm{tr}(\mathbf{T}_{b,a}^i\mathbf{\Theta}_a)) - \sum_{i=1}^{n_b}\gamma_i\xi_i\\ + & \sum_{i=1}^{n_a}\beta_i(-\rho - \eta_i + \mathrm{tr}(\mathbf{T}_{a,a}^i\mathbf{\Theta}_a)) - \sum_{i=1}^{n_a}\lambda_i\eta_i, \end{aligned}$$

where the multipliers $\alpha_i, \beta_i, \gamma_i, \lambda_i \geq 0, \forall i$. By manipulation, the Lagrangian L can be rearranged as

$$\begin{aligned}
& -\log|\boldsymbol{\Theta}_a| + \operatorname{tr}\left[\bar{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta})\boldsymbol{\Theta}_a\right] + c_1\|\boldsymbol{\Theta}_a\|_1 \qquad (5)\\
+ & \sum_{i=1}^{n_b}(c_2c_3 - \alpha_i - \gamma_i)\xi_i + \sum_{i=1}^{n_a}(c_2c_4 - \beta_i - \lambda_i)\eta_i\\
+ & \left(\sum_{i=1}^{n_b}\alpha_i - \sum_{i=1}^{n_a}\beta_i - c_2\right)\rho,
\end{aligned}$$

where

$$\bar{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta}) \triangleq \mathbf{S}_a + \sum_{i=1}^{n_a}\beta_i\mathbf{T}_{a,a}^i - \sum_{i=1}^{n_b}\alpha_i\mathbf{T}_{b,a}^i. \qquad (6)$$

Computing the derivatives of the Lagrangian with respect to the primal variables ξ_i, η_i and ρ and letting them vanish gives

$$\begin{aligned}
\frac{\partial L}{\partial \xi_i} &= c_2c_3 - \alpha_i - \gamma_i = 0, \qquad (7)\\
\frac{\partial L}{\partial \eta_i} &= c_2c_4 - \beta_i - \lambda_i = 0,\\
\frac{\partial L}{\partial \rho} &= \sum_{i=1}^{n_b}\alpha_i - \sum_{i=1}^{n_a}\beta_i - c_2 = 0.
\end{aligned}$$

Substituting Eqn.(7) into Eqn.(5) and with some manipulations, the Lagrange dual function $g(\boldsymbol{\alpha},\boldsymbol{\beta}) \triangleq \inf_{\rho,\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{\Theta}_a\succeq 0} L(\rho,\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{\Theta}_a;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda})$ can be derived as

$$g(\boldsymbol{\alpha},\boldsymbol{\beta}) = \inf_{\boldsymbol{\Theta}_a\succeq 0}\left(-\log|\boldsymbol{\Theta}_a| + \operatorname{tr}[\bar{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta})\boldsymbol{\Theta}_a] + \lambda_1\|\boldsymbol{\Theta}_a\|_1\right), \qquad (8)$$

such that $\boldsymbol{\alpha}^\top\mathbf{1} - \boldsymbol{\beta}^\top\mathbf{1} = c_2$, $0 \leq \alpha_i \leq c_2c_3\ \forall i$, and $0 \leq \beta_i \leq c_2c_4\ \forall i$.

As the primal problem in Eqn.(3) is convex, it can be equivalently solved by maximizing its dual in Eqn.(8) as

$$\begin{aligned}
& \max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\inf_{\boldsymbol{\Theta}_a\succeq 0}\left(-\log|\boldsymbol{\Theta}_a| + \operatorname{tr}\left[\bar{\mathbf{S}}\boldsymbol{\Theta}_a\right] + \lambda_1\|\boldsymbol{\Theta}_a\|_1\right) \qquad (9)\\
s.t. \quad & \boldsymbol{\alpha}^\top\mathbf{1} - \boldsymbol{\beta}^\top\mathbf{1} = c_2\\
& 0 \leq \alpha_i \leq c_2c_3, \quad i = 1,\cdots,n_b.\\
& 0 \leq \beta_i \leq c_2c_4, \quad i = 1,\cdots,n_a.
\end{aligned}$$

It is not difficult to see that the inner minimization problem is just a SICE problem, with the mere difference that $\mathbf{S}_a$ is now replaced with $\bar{\mathbf{S}}$. For any given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the inner minimization can be efficiently solved by off-the-shelf packages for graphical LASSO.

Solving the dual problem has additional benefit that it makes the discriminative learning of SICE easier to understand. As shown, it can be regarded as optimizing the $\boldsymbol{\alpha}, \boldsymbol{\beta}$ values to maximize the minimum of the inner SICE problem. The value of α_i or β_i indicates the importance of each $\mathbf{x}_b^i$ or $\mathbf{x}_a^i$ in this optimization. Specifically, the KKT conditions of Eqn.(9) are

$$\begin{aligned}
\alpha_i\left(\rho - \xi_i - \operatorname{tr}(\mathbf{T}_{b,a}^i\boldsymbol{\Theta}_a)\right) &= 0, \qquad (10)\\
\beta_i\left(-\rho - \eta_i + \operatorname{tr}(\mathbf{T}_{a,a}^i\boldsymbol{\Theta}_a)\right) &= 0,\\
\gamma_i\xi_i = 0, \quad i &= 1,\cdots,n_b,\\
\lambda_i\eta_i = 0, \quad i &= 1,\cdots,n_a.
\end{aligned}$$

It can be seen that the samples $\mathbf{x}_b^i$ with $\alpha_i > 0$ and the samples $\mathbf{x}_a^i$ with $\beta_i > 0$ are "support samples". Their Mahalanobis distances from class b are exactly at the boundary, with the value of $\rho - \xi_i$ or $\rho + \eta_i$.

We can also compute the optimal value of the primal variable ρ by looking for the samples whose optimal α_i $(i = 1,\cdots,n_b)$ does not reside at the boundaries (i.e., $0 < \alpha_i < c_2c_3$). According to Eqn.(7) and Eqn.(10), for such an α_i, we can infer that: i) $\rho - \xi_i - \operatorname{tr}(\mathbf{T}_{b,a}^i\boldsymbol{\Theta}_a) = 0$, ii) $\gamma_i \neq 0$ and $\xi_i = 0$. Therefore, ρ can be worked out as $\operatorname{tr}(\mathbf{T}_{b,a}^i\boldsymbol{\Theta}_a)$ for each i. The similar result can be obtained for β_i $(i = 1,\cdots,n_b)$. In practice, to obtain a reliable estimate, the average of all the $(n_a + n_b)$ values of ρ is calculated for use.

3.2. Joint learning of $\boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_b$

The scenario of two-class classification is more traditional in SICE-based applications. In this scenario, we learn SICEs for both classes and assign a new subject to the class with higher log likelihood. In traditional methods, the SICEs $\boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_b$ are learned separately for class a and class b, respectively. As mentioned before, this may inadvertently ignore subtle but critical network structures that distinguish the two classes. Therefore, we jointly learn $\boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_b$ to overcome this drawback. Similarly to the scenario of one-class classification, we require i) for each subject in class a, its Mahalanobis distance to class a should be smaller than its Mahalanobis distance to class b; ii) for each subject in class b, its Mahalanobis distance to class a should be larger than its Mahalanobis distance to class b; and iii) the distribution of both class a and class b should be respected. Specifically, we optimize the following objective function:

$$\min_{\rho,\boldsymbol{\xi},\boldsymbol{\eta};\boldsymbol{\Theta}_a,\boldsymbol{\Theta}_b\succeq 0} f_{\text{SICE}}(\boldsymbol{\Theta}_a) + f_{\text{SICE}}(\boldsymbol{\Theta}_b) + c_2\left[c_3\boldsymbol{\xi}^\top\mathbf{1} + c_4\boldsymbol{\eta}^\top\mathbf{1} - \rho\right] \qquad (11)$$

$$\begin{aligned}
& \operatorname{tr}(\mathbf{T}_{b,a}^i\boldsymbol{\Theta}_a) - \operatorname{tr}(\mathbf{T}_{b,b}^i\boldsymbol{\Theta}_b) \geq \rho - \xi_i, \quad i = 1,\cdots,n_b,\\
& \operatorname{tr}(\mathbf{T}_{a,a}^i\boldsymbol{\Theta}_a) - \operatorname{tr}(\mathbf{T}_{a,b}^i\boldsymbol{\Theta}_b) \leq \rho + \eta_i, \quad i = 1,\cdots,n_a,\\
& \xi_i \geq 0, \quad i = 1,\cdots,n_b,\\
& \eta_i \geq 0, \quad i = 1,\cdots,n_a.
\end{aligned}$$

As before, the variable ρ is the margin and ξ_i and η_i are the slack variables. Minimizing f_{SICE} regularizes the solutions so that they also reasonably represent the data. Again, this is a convex optimization problem. Using the techniques described in the one-class scenario, we can obtain the dual problem of this optimization problem as

$$\begin{aligned}
&\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \inf_{\boldsymbol{\Theta}_a,\boldsymbol{\Theta}_b \succeq 0} \Big[\big(-\log|\boldsymbol{\Theta}_a| + \mathrm{tr}[\bar{\mathbf{S}}_a\boldsymbol{\Theta}_a] + \lambda_1\|\boldsymbol{\Theta}_a\|_1\big) \qquad (12)\\
&\qquad + \big(-\log|\boldsymbol{\Theta}_b| + \mathrm{tr}[\bar{\mathbf{S}}_b\boldsymbol{\Theta}_b] + \lambda_1\|\boldsymbol{\Theta}_b\|_1\big) \Big]\\
&s.t. \qquad \boldsymbol{\alpha}^\top\mathbf{1} - \boldsymbol{\beta}^\top\mathbf{1} = c_2\\
&\qquad 0 \le \alpha_i \le c_2c_3, \;\; i = 1,\cdots,n_b.\\
&\qquad 0 \le \beta_i \le c_2c_4, \;\; i = 1,\cdots,n_a.
\end{aligned}$$

where

$$\begin{aligned}
\bar{\mathbf{S}}_a &\triangleq \mathbf{S}_a + \sum_{i=1}^{n_a}\beta_i\mathbf{T}_{a,a}^i - \sum_{i=1}^{n_b}\alpha_i\mathbf{T}_{b,a}^i, \qquad (13)\\
\bar{\mathbf{S}}_b &\triangleq \mathbf{S}_b - \sum_{i=1}^{n_a}\beta_i\mathbf{T}_{a,b}^i + \sum_{i=1}^{n_b}\alpha_i\mathbf{T}_{b,b}^i.
\end{aligned}$$

Note that in the inner minimization, the components for $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are totally separable. This leads to two individual SICE optimization problems, which can be readily solved as previously noted.

There are two possible ways to extend our method to multi-class: 1) for each class, treat all the samples in the other classes as negative samples and learn the SICE using our one-class formulation in Eqn.(3); and 2) Our two-class formulation in Eqn.(11) naturally extends to multi-class by including the SICE terms for all classes and adding corresponding linear constraints similarly to those in Eqn.(11).

3.3. Implementation Issues

Both the one-class and two-class discriminative learning methods can be formulated as

$$\begin{aligned}
&\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} J(\boldsymbol{\Theta};\boldsymbol{\alpha},\boldsymbol{\beta}) \qquad (14)\\
&s.t. \qquad \boldsymbol{\alpha}^\top\mathbf{1} - \boldsymbol{\beta}^\top\mathbf{1} = c_2\\
&\qquad 0 \le \alpha_i \le c_2c_3, \;\; i = 1,\cdots,n_b.\\
&\qquad 0 \le \beta_i \le c_2c_4, \;\; i = 1,\cdots,n_a,
\end{aligned}$$

where $J(\boldsymbol{\Theta};\boldsymbol{\alpha},\boldsymbol{\beta})) = \min_{\boldsymbol{\Theta}_a} f_{\mathrm{SICE}}(\boldsymbol{\Theta}_a;\bar{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta}))$ for the one-class case, and $J(\boldsymbol{\Theta}) = \min_{\boldsymbol{\Theta}_a} f_{\mathrm{SICE}}(\boldsymbol{\Theta}_a;\bar{\mathbf{S}}_a(\boldsymbol{\alpha},\boldsymbol{\beta})) + \min_{\boldsymbol{\Theta}_b} f_{\mathrm{SICE}}(\boldsymbol{\Theta}_b;\bar{\mathbf{S}}_b(\boldsymbol{\alpha},\boldsymbol{\beta}))$ for the two-class case. We use matlab "fmincon"-sqp (sequential quadratic programming) as our solver to find the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For each function evaluation of $J(\boldsymbol{\Theta};\boldsymbol{\alpha},\boldsymbol{\beta}))$ within "fmincon", we solve either one SICE or two SICE optimization problems with the modified $\mathbf{S}$. Multiple efficient algorithms have been proposed to solve SICE (or Graphical LASSO). In this paper, we use alternating directions method of multipliers (ADMM) [1] for the solution. ADMM has recently received a lot of attentions due to its fast convergence of optimization. We use the ADMM based graphical-LASSO solver in [1], which usually converges within 20 iterations and in each iteration we only need to compute the analytical solutions of two sub-optimization problems (please refer to [1] for more details). Typically, it takes a desktop computer with 3.0GHz CPU and 8.0G RAM only 0.04s to estimate a 40×40 SICE from 50 samples by ADMM. The whole discriminative learning process usually takes less than 2 minutes for the one-class case, and 6 minutes for the two-class case. Note that, increasing the network dimension leads to solving a larger SICE (which can be well-handled by the ADMM solver), but does not introduce additional constraints into Eqn.(14). In this way, we can efficiently solve our discriminative learning problems.

4. Experiment

In our experiment, the proposed discriminative SICE methods (both the one-class and the two-class formulations) are tested by classifying brain metabolic networks constructed from FDG-PET images for the prediction of Alzheimer's disease (AD). AD is the most prevalent dementia, characterized by cognitive and intellectual deficits.

4.1. Data Preparation

We download 163 FDG-PET images from the open-accessible database of Alzheimer's Disease Neuroimaging Initiative (ADNI)[1], and form them as two data sets for our experiment: i) PET-AD data set including 51 AD patients and 53 normal controls (NC); and ii) PET-MCI data set including 60 mild cognitive impairment (MCI) patients and 52 NCs. MCI is an intermediate stage of brain cognitive decline between normal ageing and AD. In addition to the PET images, we also download their accompanying T1-weighted MR images for spatial normalization described below. We conduct experiments to predict AD and MCI from NC.

Before the brain functional networks could be constructed, the PET images have to be spatially normalized for atlas-based brain parcellation. The normalization includes: i) an affine registration (using FSL package) of each PET image to its accompanying T1-weighted MR image, and ii) a deformable registration (using HAMMER package [2]) of each T1-weighted MR image to a given template MR image, for which predefined brain ROI parcellation is available. After normalization, each PET image is brought to the common space of the template image, and thus can be parcellated into ROIs according the ROI atlas in that space.

After parcellation, each ROI is characterized by the averaged uptake of radiotracer in that area. We select 40 brain ROIs whose radiotracer uptakes have the highest correlations with the class labels to build the graphical model of

[1] http://www.adni-info.org/

[2] http://www.med.unc.edu/bric/ideagroup/tools/projects-1/brain/pages-1/hammer

the functional brain network. Each node corresponds to a brain ROI, and each edge corresponds to the partial correlation of the two nodes, ie., an entry in the learned sparse inverse covariance matrix. Note that, slightly different ROIs are used for the PET-AD and PET-MCI data sets.

For each data set, we randomly partition it into 30 training-test groups, with about 60% for training and 40% for test. The classification performance is measured by both the averaged AUC (area under curve) of the ROC curves and the averaged test accuracies. For clarity, in the following we call the original SICE method without discriminative learning "**orig-SICE**", our one-class discriminative learning method "**1-disc-SICE**", and our two-class discriminative learning method "**2-disc-SICE**".

4.2. Discrimination by One-Class Formulation

As noted previously that, one-class classification can be used to detect abnormal subjects, which are the AD or MCI patients in our application. Specifically, we learn the SICE of the normal controls in the training data, and compute the likelihood of a test subject belonging to that distribution. If the value of the likelihood is lower than a threshold, the test subject is declared as an AD or MCI patient. Because classification accuracy depends on the threshold setting, we compute the ROC curve for each of the training-test group, and compare the AUCs between our one-class-SICE and the orig-SICE method. To build the ROC curve, we use the likelihood of test subjects to the NC class as scores. The lower the score, the more likely abnormal the subject. Fig. 1 shows the averaged ROC curves of the 30 training-test groups for both PET-AD (shown in red) and PET-MCI (shown in blue) datasets. It can be seen that the averaged ROC curve of our 1-disc-SICE (solid lines) always resides beyond that of the orig-SICE (dashed lines), indicating a clear advantage of the discriminatively learned SICE for classification: regardless of the threshold, 1-disc-SICE consistently has lower false positive ratio (misclassifying AD as NC) than the orig-SICE when the true positive ratio (correctly classifying NC) is controlled. Consequently, in Table 1, we observe significantly larger AUC values averaged over all the training-test groups for both classification tasks, as indicated by the small p-values in paired-t-tests.

Table 1. ***AUC for ROC curves Averaged over 30 Training-Test Groups in One-class Case***

	orig-SICE	1-disc-SICE	p-value
NC vs AD	0.8243	**0.8818**	0
NC vs MCI	0.7176	**0.7683**	0

To fully demonstrate the necessity of one-class classification, we also test the situation when there are much less training samples for one class than the other. For that purpose, the number of AD or MCI subjects for training is reudced to 30%, while the number of NC subjects for training and all the test subjects remain the same. We compare the one-class classification with the more traditional two-class classification (in Section 4.3) in this situation. The averaged ROC curves are given in Fig. 2 for PET-AD dataset and Fig. 3 for PET-MCI dataset, whose corresponding AUCs are given in Table 2. It can be seen that the orig-SICE performs better in one-class setting (solid blue line) than in two-class setting (dashed blue line). While our discriminative learning (1-disc-SICE or 2-disc-SICE, red lines) can significantly improve the AUC of the orig-SICE (blue lines) in both settings, the overall best classification performance is achieved by our 1-disc-SICE (red solid line). This demonstrates when our 1-disc-SICE will be preferred than two-class classification.

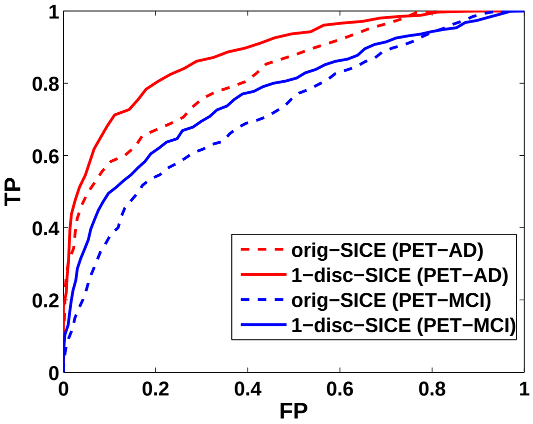


Figure 1. *One class classification: averaged ROC curves using likelihood to NC class as scores. The proposed 1-dist-SICE (solid lines) outperforms the orig-SICE (dashed lines) on both PET-AD (in red color) and PET-MCI (in blue color) datasets, respectively.*

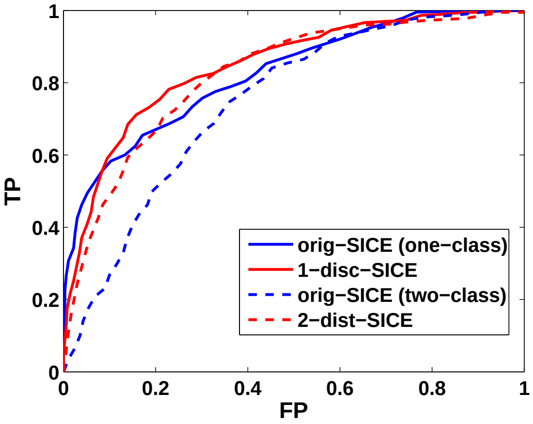


Figure 2. *Comparison of averaged ROC curves between one-class (solid lines) and two-class (dashed lines) classifications on the PET-AD dataset when only* 30% *of the training AD subjects are used (test subjects remained the same).*

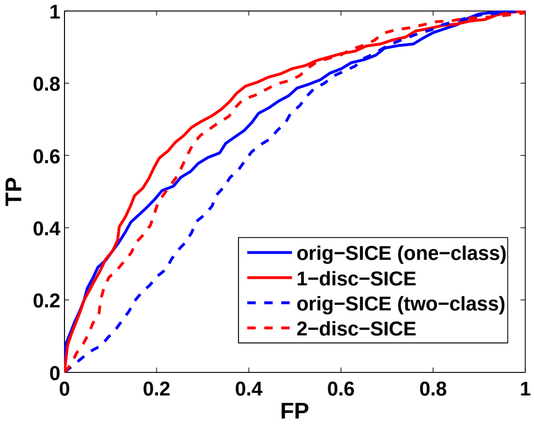


Figure 3. *Comparison of averaged ROC curves between one-class (solid lines) and two-class (dashed lines) classifications on the PET-MCI dataset when only* 30% *of the training MCI subjects are used (test subjects remained the same).*

Table 2. ***Comparison of Averaged AUC when only*** 30% ***AD or MCI training subjects are used ("disc" is our proposed method).***

AUC	One Class		Two Class	
	orig	disc	orig	disc
PET-AD-DATA	0.8243	**0.8460**	0.7458	0.8247
PET-MCI-DATA	0.7081	**0.7445**	0.6299	0.7219

4.3. Discrimination by Two-Class Formulation

We follow the common practice to employ SICE for two-class classification: training SICE for each of the two classes in comparison, and assigning the new subject to the class with higher log likelihood. The classification performance is measured in both AUCs (Fig. 4 and Table 3) and the test accuracies (Table 4), which are further compared between our proposed and the original SICE methods by paired-t-tests. For AUC, the ROC curves are computed with respect to the scores of the log likelihood difference between the two classes. Similar to the one-class case, our 2-disc-SICE produces ROC curves beyond those of the orig-SICE for both the PET-AD (shown in red) and the PET-MCI (shown in blue) data sets. The proposed two-class discriminative learning brings significant improvements to the orig-SICE in both AUCs and test accuracies as evidenced by the small p-values (< 0.05). This improvement remains salient even when the training number of AD or MCI patients is reduced as shown in Fig. 2 and Fig. 3 (comparing the red and the blue dashed lines).

4.4. Functional Connectivity

In addition to testing discrimination, we also explore the alteration of brain network structures between the patients and the healthy population. Our learned SICEs from different training-test groups are normalized to have a unit trace and binarized with a threshold of 0.005. They are then added up within the NC or the AD/MCI classes and visualized in Fig. 5. Each (i, j)-th entry represents an edge (connection) between node i and node j. The color code indicates the occurrence frequency of an edge in different training-test groups. Note that, due to the slightly different ROIs and the different (random) training-test partitions used in our PET-AD and PET-MCI data sets, the learned connectivity of NC (Fig. 5 (a) and (c)) is not identical. Nevertheless, similar patterns can be seen for the two data sets. When comparing Fig. 5 (a) with Fig. 5 (b), significant decrease of connectivity could be found in the temporal lobe (indicated by the brown box) and the subcortical region (indicated by the purple box) for AD. Such phenomenon is also observed for MCI, whose connectivity loss is less than that of AD. This trend has been extensively reported in AD-related research works using different imaging modalities [12, 7, 6], including FDG-PET [9, 7]. Interestingly, when looking into the temporal lobe, we find that many inter-hemispherical

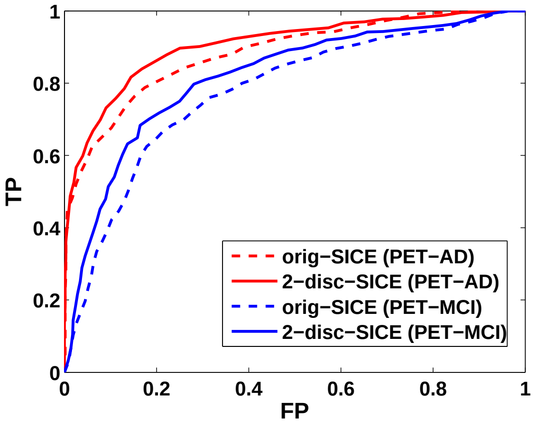


Figure 4. *Two class classification: averaged ROC curves using log likelihood difference as scores. The proposed 2-dist-SICE (solid lines) outperforms the orig-SICE (dashed lines) on both PET-AD (in red color) and PET-MCI (in blue color) datasets, respectively.*

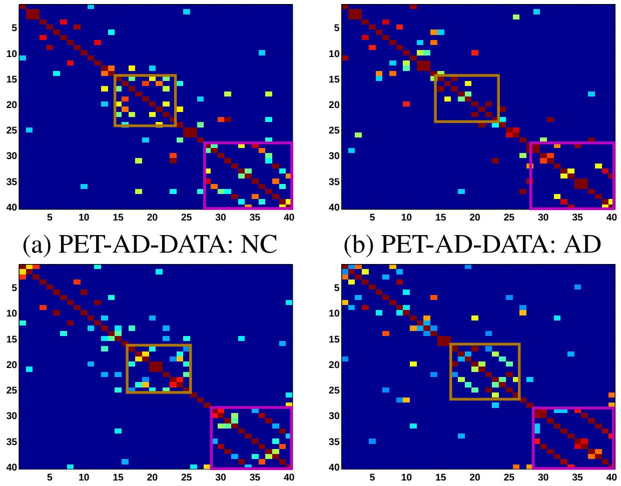
Table 3. ***AUC for ROC curves Averaged over 30 Training-Test Groups in Two-class Case***

	orig-SICE	2-disc-SICE	p-value
NC vs AD	0.8833	**0.9060**	2.3e-3
NC vs MCI	0.7755	**0.8164**	1e-8

Table 4. ***Test Accuracy*** (%) ***Averaged over 30 Training-Test Groups in Two-class Case***

	orig-SICE	2-disc-SICE	p-value
NC vs AD	78.60	**83.18**	3.8e-4
NC vs MCI	70.58	**74.87**	3.4e-5

connections between the same regions in the left and right hemispheres are lost in AD, but not in MCI. This coincides with the findings in [7].

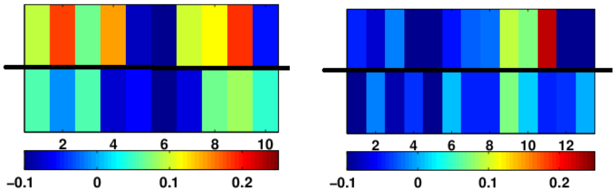


(a) PET-AD-DATA: NC (b) PET-AD-DATA: AD

(c) PET-MCI-DATA: NC (d) PET-MCI-DATA: MCI

Figure 5. *Comparison of functional connectivity between AD/MCI and NC. The brown box indicates the temporal lobe and the purple box indicates the subcortical region).*

The orig-SICE method produces similar SICE patterns as our methods. In order to conduct a detailed comparison, we compute the LCC (local clustering coefficient) for each node. LCC measures the level of local neighborhood clustering in the brain network. The changes of LCC in AD from NC are visualized in Fig. 6. The results of our method (the bottom row) are more reasonable than those of the orig-SICE (the top row). For example, our LCC changes in the temporal lobe (Fig. 6 (a)) are mostly negative and much smaller than those of orig-SICE, showing a loss of local efficiency in AD temporal lobe, aligning well with the literature. In addition, in Fig. 6 (b), node 3 and node 10 are left and right hippocampus respectively, whose loss of local clustering has been extensively reported [7, 6]. Our method again produces much smaller LCCs for these two nodes, with more loss for the left hippocampus than the right. This correlates with the findings that the left hippocampus is on average more severely affected by AD when AD has reached a moderate stage [11].



(a) Temporal Cortex (b) Subcortical Region

Figure 6. *Averaged LCC for nodes in (a) temporal cortex and (b) subcortical region for the PET-AD-data (Top: orig-SICE, Bottom: 2-disc-SICE).*

5. Conclusion

SICE is a key technique for constructing undirected graphical models of brain imaging connectomics. In this paper, we propose a learning-based framework to improve the discriminative power of SICE. Our discriminative learning problems are formulated as convex optimizations that can be solved effectively and efficiently. Compared with the existing SICE, our methods demonstrate superior classification performance and probably more reasonable discriminative patterns for AD classification. Moreover, our framework contributes to the general discriminative learning of SICE, which has broader meanings beyond the scope of brain research.

References

[1] 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(1):1–122, 2010.

[2] S. Bressler and V. Menon. Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn Sci*, 14(6):227–290, 2010.

[3] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*, 10(3):186–198, 2009.

[4] A. Fornito, A. Zalesky, C. Pantelis, and E. Bullmore. Schizophrenia, neuroimaging and connectomics. *Neuroimage*, 62(4):2296–2314, 2012.

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

[6] Y. He, Z. Chen, and A. Evans. Small-world anatomical networks in the human brain revealed by cortical thickness from mri. *Cerebral Cortex*, 17(10):2407–2419, 2007.

[7] 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*, 2009.

[8] D. Koller and N. Friedman. *Probabilistic Graphical Models Principles and Techniques*. The MIT Press, 2009.

[9] G. Sanabria-Diaz, E. Martnez-Montes, and L. Melie-Garcia. Glucose metabolism during resting state reveals abnormal brain networks organization in the alzheimer's disease and mild cognitive impairment. *PlosOne*, 8(7):e68860, 2013.

[10] S. Smith, K. Miller, G. Salimi-Khorshidi, M. Webster, C. Beckmann, T. Nichols, J. Ramsey, and M. Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

[11] P. Thompson, K. Hayashi, G. de Zubicaray, and et al. Mapping 20 hippocampal and ventricular change in alzheimer disease. *Neuroimage*, 22:1754–1766, 2004.

[12] C. Wee, P. Yap, D. Zhang, L. Wang, and D. Shen. Constrained sparse functional connectivity networks for mci classification. In *MICCAI*, pages 212–219, 2012.

[13] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.