University of Wollongong

Research Online

Faculty of Engineering and Information
Sciences - Papers: Part A

Faculty of Engineering and Information
Sciences

1-1-2016

Subject-adaptive Integration of Multiple SICE Brain Networks with Different Sparsity

Jianjia Zhang

University of Wollongong, jz163@uowmail.edu.au

Luping Zhou

University of Wollongong, lupingz@uow.edu.au

Lei Wang

University of Wollongong, leiw@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/eispapers>

Part of the [Engineering Commons](#), and the [Science and Technology Studies Commons](#)

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Subject-adaptive Integration of Multiple SICE Brain Networks with Different Sparsity

Abstract

As a principled method for partial correlation estimation, sparse inverse covariance estimation (SICE) has been employed to model brain connectivity networks, which holds great promise for brain disease diagnosis. For each subject, the SICE method naturally leads to a set of connectivity networks with various sparsity. However, existing methods usually select a single network from them for classification and the discriminative power of this set of networks has not been fully exploited. This paper argues that the connectivity networks at different sparsity levels present complementary connectivity patterns and therefore they should be jointly considered to achieve high classification performance. In this paper, we propose a subject-adaptive method to integrate multiple SICE networks as a unified representation for classification. The integration weight is learned adaptively for each subject in order to endow the method with the flexibility in dealing with subject variations. Furthermore, to respect the manifold geometry of SICE networks, Stein kernel is employed to embed the manifold structure into a kernel-induced feature space, which allows a linear integration of SICE networks to be designed. The optimization of the integration weight and the classification of the integrated networks are performed via a sparse representation framework. Through our method, we provide a unified and effective network representation that is transparent to the sparsity level of SICE networks, and can be readily utilized for further medical analysis. Experimental study on ADHD and ADNI data sets demonstrates that the proposed integration method achieves notable improvement of classification performance in comparison with methods using a single sparsity level of SICE networks and other commonly used integration methods, such as Multiple Kernel Learning.

Disciplines

Engineering | Science and Technology Studies

Publication Details

Zhang, J., Zhou, L. & Wang, L. (2017). [Subject-adaptive Integration of Multiple SICE Brain Networks with Different Sparsity](#). Pattern Recognition, 63 642-652.

Subject-adaptive Integration of Multiple SICE Brain Networks with Different Sparsity

Jianjia Zhang, Luping Zhou, Lei Wang

School of Computing and Information Technology, University of Wollongong, Wollongong, 2522, Australia.

Abstract

As a principled method for partial correlation estimation, sparse inverse covariance estimation (SICE) has been employed to model brain connectivity networks, which holds great promise for brain disease diagnosis. For each subject, the SICE method naturally leads to a set of connectivity networks with various sparsity. However, existing methods usually select a single network from them for classification and the discriminative power of this set of networks has not been fully exploited. This paper argues that the connectivity networks at different sparsity levels present complementary connectivity patterns and therefore they should be jointly considered to achieve high classification performance.

In this paper, we propose a subject-adaptive method to integrate multiple SICE networks as a unified representation for classification. The integration weight is learned adaptively for each subject in order to endow the method with the flexibility in dealing with subject variations. Furthermore, to respect the manifold geometry of SICE networks, Stein kernel is employed to embed the manifold structure into a kernel-induced feature space, which allows a linear integration of SICE networks to be designed. The optimization of the integration weight and the classification of the integrated networks are performed via a sparse representation framework. Through our method, we provide a unified and effective network representation that is transparent to the sparsity level of SICE networks, and can be readily utilized for further medical analysis. Experimental study on ADHD and ADNI data sets demonstrates that the proposed integration method achieves notable improvement of classification performance in comparison with methods using a single sparsity level of SICE networks and other commonly used integration methods, such as Multiple Kernel Learning.

Keywords:

Brain network integration, Sparse representation classification, Subject-adaptive learning, Brain disease diagnosis

1. Introduction

Functional Magnetic Resonance Imaging (fMRI) has emerged as a powerful tool to investigate the functional connectivities between different regions of human brain. A large amount of the investigations center around constructing and analyzing functional brain networks based on resting-state fMRI (rs-fMRI). Rs-fMRI focuses on the low frequency ($<0.1\text{Hz}$) oscillations of blood-oxygen-level-dependent signal which presents the task-free neuronal activation patterns of brain regions [1, 2, 3, 4]. Rs-fMRI based brain network analysis holds great promise for brain disease diagnosis [5]. For example, Attention Deficit Hyperactivity Disorder (ADHD) affects at least 5-10% of school-age children and, unfortunately, there does not exist well-known biological diagnosis. Recent studies have found increased connectivities between the right pulvinar and bilateral occipital regions and reduced connectivities between bilateral pulvinar and right prefrontal regions for ADHD patients in comparison with healthy controls [6, 7, 8]. Another example is Alzheimer's disease (AD). Compared with the healthy, AD patients have been found decreased functional connectivity between hippocampus and other brain regions [9].

With these disease-related connectivity differences, diagnosis of brain diseases could be achieved by modeling and classifying brain connectivity networks.

In recent years, several methods have been proposed to model brain connectivity networks based on the co-varying patterns of rs-fMRI time series across brain regions. The network modeling can be roughly split into two stages: i) identifying network nodes, and ii) inferring the functional connectivity between nodes. The network nodes are often defined as anatomically separated brain regions of interest (ROIs) or alternatively as latent components in some data-driven methods, e.g. independent component analysis [10] and clustering-based methods [11, 12]. Once network nodes are identified, the functional connectivity strength between a pair of nodes is conventionally measured as the correlation of the averaged time series across the brain regions associated with the two nodes [13]. In this way, the brain network is represented by a correlation matrix. Recently, it has been argued that compared with correlation, partial correlation could be a better choice since it regresses out the effects from all other nodes [14, 15]. This often results in a more accurate estimate of network structure in comparison with those correlation-based methods. Sparse inverse covariance estimation (SICE) is a principled method for partial correlation estimation, and it often produces a stable estimation with the

Email addresses: jz163@uowmail.edu.au (Jianjia Zhang), lupingz@uow.edu.au (Luping Zhou), lei@uow.edu.au (Lei Wang)

help of the sparsity regularization [16]. The resulting brain network modeled by SICE method is an inverse covariance matrix, and each of its off-diagonal entries corresponds to the partial correlation between two nodes.

Previous studies have shown the great potential of SICE networks for analyzing brain connectivity patterns [17, 18, 19, 20] and diagnosing brain diseases [21, 22, 23, 24]. For each subject, the SICE method naturally produces a set of connectivity networks by specifying different sparsity regularization parameters. Monotonic increase of this parameter would lead to gradually sparser networks. Sparse connectivity networks present the stronger connectivities while dense ones also include weaker connectivities. In this case, at which sparsity level the SICE network should be used for classification becomes a critical issue, because different connectivity patterns could possess different discriminative power. A common practice is to select one sparsity level from this set of networks [20, 22, 23] and ignore other sparsity levels. However, there are at least two drawbacks with this approach: 1) It does not fully exploit the information contained in these ignored sparsity levels; 2) The selection of the most appropriate sparsity level is usually carried out by multi-fold cross-validation, which often has high computational complexity and becomes unreliable when the sample size is small.

We argue that SICE networks with different sparsity levels could provide complementary information that is of great value for classifying the SICE networks. They should be jointly considered in order to improve the discriminative power. Also, this will circumvent the rigid selection of a single sparsity level. To this end, a straightforward method might be to concatenate the features extracted from these SICE networks. However, this method suffers from two limitations: 1) not all the extracted features are sufficiently discriminative; and 2) this method treats each network separately and indiscriminately, failing to integrate them in an inherent and adaptive manner. Another appealing method may be multiple kernel learning (MKL). However, although MKL combines multiple kernels from different channels, it ignores the inter-channel information, as will be explained in Section 3.5.1.

In this paper, we propose a learning based framework that integrates a set of SICE networks with the aim of attaining more discriminative power. Our framework has at least four contributions.

1. It makes use of the whole spectrum of SICE matrices at different sparsity levels, which not only improves the classification performance, but also circumvents the need of presetting the employed sparsity level.
2. The proposed framework provides subject-adaptive integration of SICE networks. It is noticed that some sparsity levels that are useful for one subject could become less useful for another due to the variation of subject-specific characteristics, e.g., disease phase, age and gender. In this case, we allow the integration to be subject-adaptive, and achieve this through a sparse representation framework.
3. Our integration of SICE networks respects the specific geometric property of SICE matrix. As known, SICE matrices

are symmetric positive definite (SPD) and form a Riemannian manifold [25, 26], which makes a linear combination of them in the input (Euclidean) space improper. To address this issue, we propose to embed the Riemannian manifold into a reproducing kernel Hilbert space (RKHS), where linear operations become sufficiently good to handle SICE matrices. This embedding is achieved through SPD-matrix based kernels, such as the Stein kernel [27]. Following that, a linear combination of multiple SICE networks is optimized in this kernel induced feature space through the sparse representation framework mentioned above.

4. By the integration, our learning framework provides a unique, enhanced, and new network representation for each subject. Although the integration takes place in a kernel induced feature space, it is feasible to project the integration result back into the original network space for visualization and further medical analysis. This could help to understand the underlying pathophysiology of brain diseases.

The rest of the paper is organized as follows: Section 2 reviews the SICE algorithm, the properties and measurements of SICE networks and the sparse representation technique. Section 3 details the proposed subject-adaptive integration method and discusses several issues regarding the proposed method. Section 4 presents the experimental results on ADHD and ADNI rs-fMRI data sets. And Section 5 concludes this paper.

2. Related Work

2.1. Constructing Brain Networks Using SICE

Let $\{x_1, x_2, \dots, x_M\}$ be a time series of length L , where x_i is a d -dimensional vector, corresponding to an observation of d brain nodes. Following the literature of SICE [16, 21], x_i is assumed to follow a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. Each off-diagonal entry of Σ^{-1} indicates the partial correlation between two nodes by eliminating the effect of all other nodes. Σ_{ij}^{-1} will be zero if nodes i and j are independent of each other when conditioned on other nodes. In this sense, Σ_{ij}^{-1} can be interpreted as the existence and strength of the connectivity between nodes i and j . The estimate of Σ^{-1} , denoted as S , can be obtained by maximizing the penalized log-likelihood over positive definite matrix S ($S > 0$) [16, 21]:

$$S^* = \arg \max_{S > 0} \log(\det(S)) - \text{tr}(\tilde{\Sigma}S) - \lambda \|S\|_1 \quad (1)$$

where $\tilde{\Sigma}$ is the sample-based covariance matrix; $\det(\cdot)$, $\text{tr}(\cdot)$ and $\|\cdot\|_1$ denote the determinant, the trace and the sum of the absolute values of the entries of a matrix, respectively. Here $\|S\|_1$ imposes sparsity on S to achieve more reliable estimation by considering the fact that a brain region often has limited direct connections with other brain regions in neurological activities. The tradeoff between the degree of sparsity and the log-likelihood estimation of S is controlled by the regularization parameter λ . A larger λ makes S^* sparser. The maximization problem in Eq. (1) can be efficiently solved by the off-the-shelf

Figure 1: An example of 10×10 -dimensional SICE networks with different λ values. The colorbar indicates the strength of connectivities. A smaller λ results in a denser connectivity network while a larger λ leads to a sparser network.

packages, such as SLEP [28]. In the following, we may use SICE matrix to refer the SICE network estimated by Eq. (1).

2.2. Properties and Measurements of SICE Matrices

2.2.1. Riemannian manifold geometry

The resulting SICE matrix $\mathbf{S}^*$ obtained by Eq. (1) is symmetric positive definite (SPD). Let Sym_d^+ denote the $d \times d$ SPD matrix set: $\text{Sym}_d^+ = \{\mathbf{A} | \mathbf{A} = \mathbf{A}^\top, \forall \mathbf{x} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{0}, \mathbf{x}^\top \mathbf{A} \mathbf{x} > 0\}$. It forms a closed, self-dual convex cone, which is a Riemannian manifold [27]. A set of SICE networks with various sparsity can be obtained for a single subject by varying λ (in Eq.(1)), as illustrated in Figure 2 (a). The operations and algorithms that are developed for Euclidean spaces can not be directly applied to SPD matrices residing on such a manifold [26]. For example, the average of two SPD matrices could result in an undesirable swelling effect [25]: the determinant of the Euclidean mean of two SPD matrices can be larger than the determinants of the original two matrices, which is physically unrealistic in many applications, such as diffusion tensor interpolation. To address this issue, methods that respect the manifold geometry have been developed [25]. Kernel method is one of the most effective methods and has been adopted in [29, 26, 30] to measure the similarity between SPD matrices. It implicitly maps the Riemannian manifold of SPD matrices onto a high-dimensional kernel-induced feature space $\mathcal{F}$, where linear algorithms developed for Euclidean spaces can be applied. At the same time, the manifold structure is well incorporated into the mapping by utilizing distance functions that are specially designed for SPD matrices. In addition, kernel methods are often computationally efficient.

2.2.2. Kernel functions for SICE networks

Various kernel functions have been proposed to map the Riemannian manifold of SPD matrices onto a feature space $\mathcal{F}$, such as Cholesky kernel [31], Log-Euclidean kernel [25, 26] etc. A recently proposed Stein kernel [27] has shown promising performance in a variety of applications [29, 32, 33, 34]. It is expressed as

$$k(\mathbf{X}, \mathbf{Y}) = \exp(-\theta \cdot S(\mathbf{X}, \mathbf{Y})) \quad (2)$$

where θ is a tunable positive scalar. $S(\mathbf{X}, \mathbf{Y})$ is called S-Divergence and it is defined as

$$S(\mathbf{X}, \mathbf{Y}) = \log \left(\det \left(\frac{\mathbf{X} + \mathbf{Y}}{2} \right) \right) - \frac{1}{2} \log(\det(\mathbf{X}\mathbf{Y})), \quad (3)$$

The S-Divergence has the following desirable properties: 1) It is invariant to affine transformations applied to the input SPD matrices, such as rotation, scaling; 2) The square-root of S-Divergence is proven to be a metric on Sym_d^+ [27]; 3) Stein kernel is guaranteed to be a Mercer kernel when θ varies within the range of $\Theta = \{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots, \frac{(d-1)}{d}\} \cup (\frac{(d-1)}{d}, +\infty)$. Readers are referred to [27] for more details.

Besides Stein kernel, there are other commonly used SPD kernels, including Cholesky kernel [31], Euclidean kernel [31], Log-Euclidean kernel [25, 26], and Power Euclidean kernel [31]. All the four kernels share the same form of:

$$k(\mathbf{S}_i, \mathbf{S}_j) = \exp(-\theta \cdot d^2(\mathbf{S}_i, \mathbf{S}_j)), \quad (4)$$

where $d(\cdot, \cdot)$ is a kind of distance between two SPD matrices. Different definitions of $d(\cdot, \cdot)$ lead to different kernels. The distance functions in the four kernels are introduced as follows.

- Cholesky distance. Cholesky distance measures the difference between $\mathbf{S}_i$ and $\mathbf{S}_j$ by

$$d(\mathbf{S}_i, \mathbf{S}_j) = \|\text{chol}(\mathbf{S}_i) - \text{chol}(\mathbf{S}_j)\|_F \quad (5)$$

where $\text{chol}(\mathbf{S})$ is a lower triangular matrix with positive diagonal entries obtained by the Cholesky decomposition of $\mathbf{S}$, that is, $\mathbf{S} = \text{chol}(\mathbf{S}) \text{chol}(\mathbf{S})^\top$ and $\|\cdot\|_F$ denotes the Frobenius matrix norm.

- Euclidean distance. Euclidean distance between $\mathbf{S}_i$ and $\mathbf{S}_j$ is given by

$$d(\mathbf{S}_i, \mathbf{S}_j) = \|\mathbf{X}_1 - \mathbf{X}_2\|_F \quad (6)$$

- Power Euclidean distance. Power Euclidean distance between $\mathbf{S}_i$ and $\mathbf{S}_j$ is given by

$$d(\mathbf{S}_i, \mathbf{S}_j) = \frac{1}{p} \|\mathbf{S}_i^p - \mathbf{S}_j^p\|_F \quad (7)$$

where p is a non-zero scalar. Note that $\mathbf{S}$, as a SPD matrix, can be eigen-decomposed as $\mathbf{S} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top$, and $\mathbf{S}^p$ can be easily computed by: $\mathbf{S}^p = \mathbf{U} \mathbf{\Lambda}^p \mathbf{U}^\top$.

- Log-Euclidean distance. Log-Euclidean distance is defined as

$$d(\mathbf{S}_i, \mathbf{S}_j) = \|\log(\mathbf{S}_i) - \log(\mathbf{S}_j)\|_F \quad (8)$$

where $\log(\mathbf{S}) = \mathbf{U} \log(\mathbf{\Lambda}) \mathbf{U}^\top$ and $\log(\mathbf{\Lambda})$ applies logarithm to each diagonal element of $\mathbf{\Lambda}$ to obtain a new diagonal matrix.

Table 1: Properties of SPD kernels.

Metric name	Cholesky [31]	Euclidean [31]	Log-Euclidean [25]	Power-Euclidean [31]	S-Divergence root [27]
Does $k = \exp(-\theta \cdot d^2)$ define a valid kernel?	Yes	Yes	Yes	Yes	Yes ($\theta \in \Theta$)
Kernel abbr. in the paper	CHK	EUK	LEK	PEK	SK
Time complexity	$O(d^3)$	$O(d^2)$	$O(d^3)$	$O(d^3)$	$O(d^{2.373})$

The properties of all these SPD kernels mentioned above are summarized in Table 1. All the kernel functions in this table will be investigated in the experimental study.

2.3. Sparse Representation

Sparse representation [35] of signals have received intensive attention over the last decade due to its effectiveness in modeling signals and the robustness in dealing with corrupted data. Sparse representation aims to search for the sparsest representation of a signal by using a linear combination of atoms in an overcomplete dictionary. Specifically, given a signal $\mathbf{x} \in \mathbb{R}^n$ and a set of atoms $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_N]$, $\mathbf{d}_i \in \mathbb{R}^n$ as a dictionary, the idea of sparse representation is to reconstruct signal $\mathbf{x}$ by using a few atoms in dictionary $\mathbf{D}$. The sparse representation coefficient of $\mathbf{x}$, denoted as $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_N]^T$, can be obtained by solving the following problem:

$$\boldsymbol{\alpha}^* = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \|\mathbf{x} - \sum_{i=1}^N \alpha_i \mathbf{d}_i\|_2^2 + \gamma \|\boldsymbol{\alpha}\|_1 \quad (9)$$

where the first term of Eq.(9) is the reconstruction error of $\mathbf{x}$, and the second term is used to control the sparsity of the coefficient $\boldsymbol{\alpha}$. The parameter γ is the tradeoff parameter used to balance the reconstruction error and sparsity. A larger γ leads to a sparser solution. Kernel-based sparse representation (K-SR) [36] extends the concept of sparse representation from a vector space to a kernel-induced feature space $\mathcal{F}$. KSR shares the same form of Eq.(9) except that both the signal and the atoms become the mapped features in $\mathcal{F}$.

3. Proposed method

3.1. Problem Formulation

As previously mentioned, the SICE representation naturally leads to a set of networks with different sparsity levels. Each sparsity level of the network captures specific connectivity patterns, which are different from that of other sparsity levels. Therefore, when the final aim is to classify different groups of subjects, such as subjects with ADHD versus healthy controls, multiple sparsity levels of connectivity networks should be jointly considered and integrated. By doing so, the complementary and discriminative information of the connectivity patterns from multiple sparsity levels could be fully explored to boost the classification performance.

To integrate multiple sparsity levels of SICE networks, a natural approach is linear combination. However, as previously

mentioned, linearly combining manifold data in a Euclidean space fails to respect the non-linear structure of the Riemannian manifold and could lead to spurious result, such as the swelling effect [25]. Accounting for the manifold structure of SICE networks, we propose a method that conducts such an integration in a kernel-induced feature space $\mathcal{F}$. Firstly, multiple sparsity levels of SICE networks from the same subject are mapped into $\mathcal{F}$ by a non-linear mapping function $\phi(\cdot) : \text{Sym}_d^+ \rightarrow \mathcal{F}$. This $\phi(\cdot)$ mapping brings at least two advantages. 1) The Riemannian manifold geometry has been well considered by using distance functions specially designed for SPD matrices, such as the Stein divergence; 2) The images of SICE matrices can be linearly processed in the kernel-induced feature space $\mathcal{F}$. After the mapping, we utilize a convex combination to integrate multiple ϕ -mapped SICE networks in $\mathcal{F}$. Convex combination constrains the resulting network to lie in the convex hull of these ϕ -mapped SICE networks, implicitly making the solution better comply to the local distribution of these networks. Specifically, let $\mathbb{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M\}$, $\mathbf{X}_i \in \text{Sym}_d^+$, denote the SICE networks of one subject at M different sparsity levels and $f(\mathbb{X})$ denote the combined network of $\mathbb{X}$ with implicit function f . The convex combination is carried out in the following manner:

$$\begin{aligned} \Phi &= \sum_{j=1}^M \beta_j \phi(\mathbf{X}_j) \\ \text{subject to: } &\sum_{j=1}^M \beta_j = 1; \quad \beta_j \geq 0. \end{aligned} \quad (10)$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_M]^T$ is the combination coefficient of function f . Instead of applying a uniform $\boldsymbol{\beta}$ to all the subjects, we assign a subject-adaptive $\boldsymbol{\beta}$ for each subject. We argue that the discriminative power of SICE networks at different sparsity levels may not necessarily be uniform across all the subjects. This can be intuitively understood by the following considerations. 1) Although sharing the same class label, the patient subjects could be experiencing different stages of the disease. As a result, disease-induced alterations in the brain networks will be different across subjects. 2) The brain networks may demonstrate differences across subjects due to different age, gender, etc. Thus, we propose to adaptively optimize $\boldsymbol{\beta}$ for each subject to handle such variation as follows.

To solve $\boldsymbol{\beta}$, this paper extends the kernel-based sparse representation (KSR) [36]. The original KSR is used to obtain the sparse representation coefficient $\boldsymbol{\alpha}$ only. In this paper, we develop it to optimize the integration coefficients $\boldsymbol{\beta}$ and the

Figure 2: Illustration of the proposed SICE network integration method. (a) a set of SICE networks with various sparsity (denoted by green dots) residing on a Riemannian manifold can be obtained for a single subject; (b) the SICE networks are mapped to a kernel-induced feature space $\mathcal{F}$ by using a SPD kernel. Then the mapped networks in $\mathcal{F}$ are convexly combined as a unified representation Φ (denoted by a red dot); (c) the unified representation is sparsely represented by the atoms (denoted by blue dots) in $\mathcal{F}$.

sparse representation coefficient α simultaneously. The proposed method is called sample-adaptive SICE network integration (SASNI).

Let $\mathbb{D} = \{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_N\}$, $\mathbf{D}_j \in \text{Sym}_d^+$, denote a dictionary with N SPD matrices. Note that $\mathbf{X}_i$ and $\mathbf{D}_j$ are both SICE matrices, however, $\mathbf{X}_i$ denotes the i -th SICE matrix of a test subject while $\mathbf{D}_j$ denotes the j -th atom of the dictionary, which is formed by SICE matrices from training subjects in this paper. The objective function of SASNI is formulated as follows.

$$\begin{aligned} \arg \min_{\alpha \in \mathbb{R}^N, \beta} \quad & \|\Phi - \sum_{i=1}^N \alpha_i \phi(\mathbf{D}_i)\|_2^2 + \gamma \|\alpha\|_1 \\ \text{subject to:} \quad & \sum_{j=1}^M \beta_j = 1; \quad \beta_j \geq 0. \end{aligned} \quad (11)$$

The first term of Eq.(11) is the reconstruction error of Φ , and the second term is used to control the sparsity of α .

Figure 2 illustrates the proposed method: (a) a set of SICE networks with various sparsity (denoted by green dots) residing on a Riemannian manifold can be obtained for a single subject; (b) the SICE networks are mapped to a kernel-induced feature space $\mathcal{F}$ by using SPD kernels. Then the mapped networks in $\mathcal{F}$ are convexly combined as a unified representation (denoted by a red dot); (c) the unified representation is sparsely represented by the atoms (denoted by blue dots) in $\mathcal{F}$. How to solve the combination coefficient β and the sparse representation coefficient α and how they can be used for classification will be elaborated in the following parts.

3.2. Solving The Optimization Problem

As known in kernel methods, the kernel mapping ϕ is usually too complicated to be explicitly computed. Therefore, we need to bypass the mapping ϕ in Eq.(11). With the kernel property of $k(\mathbf{A}, \mathbf{B}) = \phi(\mathbf{A})^\top \phi(\mathbf{B})$, the first term in Eq.(11) can be expanded as:

$$\begin{aligned} & \|\Phi - \sum_{i=1}^N \alpha_i \phi(\mathbf{D}_i)\|_2^2 \\ &= \Phi^\top \Phi - 2 \sum_{i=1}^N \alpha_i \Phi^\top \phi(\mathbf{D}_i) \\ &+ \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \phi(\mathbf{D}_i)^\top \phi(\mathbf{D}_j) \\ & \quad (\text{Substitute } \Phi \text{ with Eq.(10)}) \\ &= \beta^\top \mathcal{K}(\mathbb{X}, \mathbb{X}) \beta - 2\alpha^\top \mathcal{K}(\mathbb{D}, \mathbb{X}) \beta + \alpha^\top \mathcal{K}(\mathbb{D}, \mathbb{D}) \alpha \end{aligned} \quad (12)$$

where $\mathcal{K}(\mathbb{X}, \mathbb{X}) = [k(\mathbf{X}_i, \mathbf{X}_j)]_{M \times M}$, $\mathcal{K}(\mathbb{D}, \mathbb{X}) = [k(\mathbf{D}_i, \mathbf{X}_j)]_{N \times M}$ and $\mathcal{K}(\mathbb{D}, \mathbb{D}) = [k(\mathbf{D}_i, \mathbf{D}_j)]_{N \times N}$. Note that, as shown in Eq.(12), the objective in Eq.(11) is a function of both α and β . It can be proved that the objective function is jointly convex on both of α and β . To find the optima for α and β , we devise an iterative procedure that monotonically decreases the objective by alternately optimizing one variable at a time while keeping the other fixed. Specifically, at the t -th iteration, we first optimize the objective in Eq.(11) over α with β fixed. In this case, $\beta^\top \mathcal{K}(\mathbb{X}, \mathbb{X}) \beta$ reduces to a constant and Eq.(11) can be rewritten as follows:

$$\alpha^{(t)} = \arg \min_{\alpha \in \mathbb{R}^N} -2\alpha^\top \mathcal{K}(\mathbb{D}, \mathbb{X}) \beta^{(t)} + \alpha^\top \mathcal{K}(\mathbb{D}, \mathbb{D}) \alpha + \gamma \|\alpha\|_1 \quad (13)$$

The objective of Eq.(13) is a quadratic function and convex with respect to α . This optimization problem can be efficiently solved via existing packages, such as SPAMS [37]. After obtaining $\alpha^{(t)}$, β can be updated in a similar manner by optimizing the following problem:

$$\begin{aligned} \beta^{(t)} = \arg \min_{\beta} \quad & \beta^\top \mathcal{K}(\mathbb{X}, \mathbb{X}) \beta - 2\alpha^{(t)\top} \mathcal{K}(\mathbb{D}, \mathbb{X}) \beta \\ \text{subject to:} \quad & \sum_{j=1}^M \beta_j = 1; \quad \beta_j \geq 0. \end{aligned} \quad (14)$$

This optimization problem can be solved by the reduced gradient descent method as in [38] to handle the convex constrain. α and β are optimized alternately until a stopping criterion is satisfied. There are two commonly used stopping criteria: i) a predefined threshold on the difference of consecutive objective function values is met; and ii) the maximum iteration number is reached.

Note that as the original KSR [36], there is no training stage in the proposed method, and the training subjects are used as the atoms in the dictionary $\mathbb{D}$. α and β are optimized individually for each test subject, so they are adaptive for different test subjects. This subject-adaptivity of β enhances the flexibility of SASNI to integrate multiple networks across subjects, and in turn improves the effectiveness of SASNI.

3.3. Classification Based on SASNI

With the obtained sparse representation coefficient α , a test sample can be classified in multiple manners. If the atoms in the dictionary are not associated with class labels, the sparse representation coefficient can be treated as a feature vector, and a classifier, such as support vector machine (SVM), can be trained with the feature vectors of training samples. The class label of a test sample is then predicted by feeding its feature vector to the trained classifier. If the atoms in the dictionary are associated with class labels, i.e., the atoms are just the training samples, the sparse representation coefficient can be directly used to classify a test sample through calculating class-specific residue. This approach is denoted as sparse representation based classification (SRC) in [35] and also used in this paper. SRC takes the assumption that the samples from the same class do lie in a subspace. In this case, any new (test) sample from the same class shall approximately lie in the span of the training samples and can therefore be sufficiently represented by the training samples from the same class. Then the class-specific reconstruction residue can be calculated by using the sparse representation coefficients and the training samples associated with each class. Eventually, the test sample is assigned to the class with the minimal class-specific reconstruction residue. SRC has been used in some generic image classification tasks, such as face recognition [35] and object categorization [29, 30], and has demonstrated its effectiveness and high efficiency.

Specifically, in this paper, once the combination coefficients β and the sparse representation coefficient α are obtained by optimizing Eq.(11), a new (test) sample $\mathbb{X}$ will be assigned to the class with the minimum reconstruction residue. The reconstruction residue of Φ for class i can be computed as follows:

$$\begin{aligned} \epsilon_i(\Phi) &= \|\Phi - \sum_{j=1}^N \alpha_j \delta(l(j) - i) \phi(\mathbf{D}_j)\|_2^2 \\ &= \beta^T \mathcal{K}(\mathbb{X}, \mathbb{X}) \beta - 2\tilde{\alpha}_i^T \mathcal{K}(\mathbb{D}, \mathbb{X}) \beta + \tilde{\alpha}_i^T \mathcal{K}(\mathbb{D}, \mathbb{D}) \tilde{\alpha}_i. \end{aligned} \quad (15)$$

where $l(j)$ is the label of the j -th atom $\mathbf{D}_j$, $\delta(k)$ is the Kronecker delta function, which is one when k is zero, and zero otherwise, and $\tilde{\alpha}_i = \alpha \odot [\delta(l(1) - i), \delta(l(2) - i), \dots, \delta(l(N) - i)]^T$, with $\odot$ indicating element-wise product. Based on Eq. (15), the final label of $\mathbb{X}$ is determined as follows:

$$l(\mathbb{X}) = \arg \min_i \epsilon_i(\Phi) \quad (16)$$

Algorithm 1 outlines the proposed method SASNI.

Algorithm 1 Proposed subject-adaptive SICE network integration (SASNI).

Input: A training sample set $\mathbb{D} = \{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_N\}$; A test sample $\mathbb{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M\}$; γ , the maximum iteration number T and the threshold τ .

Output: $l(\mathbb{X}), \beta^*, \alpha^*$;

- 1: Initialize $\beta_0, \alpha_0, t = 0, Obj_0 = inf$;
- 2: **repeat**
- 3: $t = t + 1$;
- 4: Update $\beta^{(t)}$ according to Eq.(14) with $\alpha^{(t-1)}$;
- 5: Update $\alpha^{(t)}$ according to Eq.(13) with $\beta^{(t)}$;
- 6: Update Obj_t according to Eq.(11);
- 7: **until** $t \geq T$ or $|Obj_{t-1} - Obj_t| \leq \tau$;
- 8: **return** $l(\mathbb{X}) = \arg \min_i \epsilon_i(\mathbb{X}),$
 $\beta^* = \beta^{(t)}, \alpha^* = \alpha^{(t)}.$

3.4. Projection Back to The Original Space

The integrated network provides a unified network representation for each subject and therefore it is worthy to visualize for future analysis. However, as shown above, the integration takes place in a kernel-induced feature space. As a result, it needs to be projected back into the original space. For this purpose, we use a kernel pre-image estimation method to recover the integrated network in the original input space. The idea is to model the pre-image, denoted as $\hat{\mathbf{S}}$, of the integrated network Φ by a linear combination of its neighboring SICE matrices in Sym_d^+ , i.e., $\hat{\mathbf{S}} = \sum_{\mathbf{S}_j \in \Omega} w_j \mathbf{S}_j$, where Ω denotes a set of neighboring SICE networks and w_j denotes the combination weight. Note that the neighboring SICE networks can be found by measuring similarities between Φ and $\phi(\mathbf{S}_i)$ in the kernel-induced feature space. The optimal $\mathbf{w}$ can be obtained by solving the following optimization problem using gradient descent based algorithms.

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \geq 0; \mathbf{w}^T \mathbf{1} = 1} \|\Phi - \phi\left(\sum_{\mathbf{S}_j \in \Omega} w_j \mathbf{S}_j\right)\|_2^2 \quad (17)$$

The objective function in Eq.(17) can be further reorganized into terms of kernel matrices as in Eq.(18).

$$\begin{aligned} &\|\Phi - \phi\left(\sum_{\mathbf{S}_j \in \Omega} w_j \mathbf{S}_j\right)\|_2^2 \\ &= \beta^T \mathcal{K}(\mathbb{X}, \mathbb{X}) \beta - 2\mathcal{K}\left(\sum_{\mathbf{S}_j \in \Omega} w_j \mathbf{S}_j, \mathbb{X}\right) \beta + k \left(\sum_{\mathbf{S}_j \in \Omega} w_j \mathbf{S}_j, \sum_{\mathbf{S}_j \in \Omega} w_j \mathbf{S}_j\right) \end{aligned} \quad (18)$$

3.5. Discussion

3.5.1. Comparison between MKL and SASNI

Multiple kernel learning (MKL) has been regarded as a promising technique for integrating multiple data sources [39].

Table 2: Summary of the ADHD-200 data set.

Training Data Set								Test Data Set							
	KKI	NeuroImage	NYU	OHSU	Peking	Pittsburgh	WashU	KKI	NeuroImage	NYU	OHSU	Peking	Pittsburgh	Brown	
# sub	83	48	216	79	194	89	59	11	25	41	34	51	9	26	
Control	61	23	98	42	116	89	59	8	14	12	28	27	5	N.A.	
ADHD	22	25	118	37	78	0	0	3	11	29	6	24	4	N.A.	
Age	8-13	11-22	7-18	7-12	8-17	10-20	7-22	8-12	13-26	7-17	7-12	8-15	14-17	8-18	
Male	46	31	140	43	144	46	32	10	12	28	17	32	7	9	
Female	37	17	76	36	50	43	27	1	13	13	17	19	2	17	

It can also be used to integrate multiple sparsity levels of brain networks. Here we discuss the similarity and difference between MKL and the proposed SASNI. The idea of MKL is to search for an integration of base kernel functions that maximizes a generalized performance measure, such as structural risk minimization [39]. Formally, let $\mathbb{X} = \{X_1, X_2, \dots, X_M\}$ denote M data sources. MKL uses the concatenation of mapped features as the integrated feature representation:

$$\Phi_{\mathbb{X}} = [\sqrt{\omega_1}\phi(X_1), \sqrt{\omega_2}\phi(X_2), \dots, \sqrt{\omega_M}\phi(X_M)]^T$$

where $\omega = [\omega_1, \omega_2, \dots, \omega_M]^T$ is the integration coefficient. The dot product in the integrated feature space gives the integrated kernel:

$$\begin{aligned} \langle \Phi_{\mathbb{X}^p}, \Phi_{\mathbb{X}^q} \rangle &= [\sqrt{\omega_1}\phi(X_1^p), \sqrt{\omega_2}\phi(X_2^p), \dots, \sqrt{\omega_M}\phi(X_M^p)] \cdot \\ &\quad [\sqrt{\omega_1}\phi(X_1^q), \sqrt{\omega_2}\phi(X_2^q), \dots, \sqrt{\omega_M}\phi(X_M^q)]^T \\ &= \sum_{j=1}^M \omega_j k(X_j^p, X_j^q) \end{aligned} \quad (19)$$

In contrast, the integration of SASNI is carried out in the following manner:

$$\Phi_{\mathbb{X}} = \sum_{j=1}^M \beta_j \phi(X_j)$$

and the dot product in the integrated feature space is:

$$\langle \Phi_{\mathbb{X}^p}, \Phi_{\mathbb{X}^q} \rangle = \sum_{i=1}^M \sum_{j=1}^M \beta_i \beta_j k(X_i^p, X_j^q) \quad (20)$$

By cross-referring Eq.(19) and Eq.(20), we can identify that:

1. The similarity of MKL and SASNI is that both of them adopt multiple data sources in a kernel-induced feature space $\mathcal{F}$.
2. The difference between MKL and SASNI is that MKL only considers the similarity between samples from the same data source, i.e. X_j^p and X_j^q , $j \in [1, M]$. In contrast, SASNI explores all pairs of data sources, i.e. X_i^p and X_j^q , $i, j \in [1, M]$ to measure the similarity between two samples. Thus SASNI admits more flexibility to incorporate cross-source information.

At the same time, note that the proposed method specially caters for brain network data where cross-source data can be compared, and MKL is a more general method for combining different data sources.

3.5.2. Convergence Analysis and Computational Complexity

As outlined in Algorithm 1, the optimization problem defined in Eq.(11) is solved by a commonly used alternate optimization strategy which alternately minimizes the objective function with respect to one of the two variables, i.e. the combination coefficient β and the sparse representation coefficient α . Considering the two facts 1) these two variables are iteratively optimized to decrease the same objective function monotonically; and 2) the objective function is lower bounded by zero, the optimization problem defined in Eq.(11) is guaranteed to converge. Also, it can be proved that the objective function is jointly convex over both β and α . Therefore, the final solution is guaranteed to be a global optimum. The experimental results show that employing this optimization strategy has already been able to achieve promising performance.

The main computational cost of the proposed method is twofold: kernel computation and parameter optimization.

1. *kernel computation.* As indicated in Eq.(12), $\mathcal{K}(\mathbb{X}, \mathbb{X})$, $\mathcal{K}(\mathbb{D}, \mathbb{X})$ and $\mathcal{K}(\mathbb{D}, \mathbb{D})$ can be precomputed before the optimization. The time complexity of computing different kernel functions over a pair of $d \times d$ SICE matrices is listed in Table 1. It can be seen that when the Stein kernel (SK) is used, the time complexity involved in calculating the kernel matrices $\mathcal{K}(\mathbb{X}, \mathbb{X})$, $\mathcal{K}(\mathbb{D}, \mathbb{X})$ and $\mathcal{K}(\mathbb{D}, \mathbb{D})$ is $O(M^2 d^{2.373})$, $O(MN d^{2.373})$ and $O(N^2 d^{2.373})$, respectively, where M indicates the number of sparsity levels of connectivity networks and N indicates the number of atoms used to reconstruct $\mathbb{X}$.
2. *Parameter optimization.* In terms of the optimization, either β or α is obtained by solving a quadratic programming (QP) problem. Let $O(QP(n))$ denote the computational complexity to solve a convex QP problem with n variables and T denote the number of iterations performed in the optimization process. The optimization complexity can be expressed as $O(T(QP(M) + QP(N)))$. Once β and α are optimized, the class label of $\mathbb{X}$ can be quickly assigned according to Eq.(15) and Eq.(16).

Therefore, the overall time complexity of SASNI for each test subject is $O(T(QP(M) + QP(N)) + (M^2 + MN + N^2) d^{2.373})$. In our experiment, a single subject can be classified in less than one second on a desktop computer with 3.6 GHz CPU and 32 GB memory.

Figure 3: Comparison of classification performance between different methods using a single sparsity level of SICE network on ADHD-200 data set. LCC indicates that local clustering coefficient is extracted for classification. CHK, EUK, LEK, PEK, SK denote the five SPD kernels summarized in Table 1.

3.5.3. Non-subject-adaptive variants of SASNI

In SASNI, the integration coefficient β is adaptively assigned to each subject to handle the variation of the utility of different sparsity levels across subjects. Then a question arises naturally: how the classification performance will be if a uniform β is applied for all the subjects to integrate multiple SICE networks at different sparsity levels? To answer this question, we explore two non-subject-adaptive integration variants and conduct a comparison between them and the proposed subject-adaptive counterpart in our experiment (see Section 4.2.5). The first variant is to treat each sparsity level equally and set each entry of β to $1/M$, where M is the total number of sparsity levels to integrate. This method, called ‘MeanC’, can be intuitively understood as using the geometric center of the networks in the kernel-induced feature space $\mathcal{F}$ as the representation of the network set. The second method is to learn a fixed β based on the training data set. To do this, the training data set is split into two folds, one fold as dictionary and the other fold as the validation set. A fixed β is learned in a similar manner as introduced in Section 3. The only difference is the reconstruction error defined in Eq. (11) is calculated over the subjects in the validation set only. Therefore, the objective function becomes:

$$\begin{aligned} \arg \min_{\alpha \in \mathbb{R}^N, \beta} & \sum_{p=1}^P \| \phi(f_{\beta}(\mathbb{X}^p)) - \sum_{i=1}^N \alpha_i \phi(\mathbf{D}_i) \|_2^2 + \gamma \| \alpha \|_1 \\ \text{subject to: } & \sum_{j=1}^M \beta_j = 1; \quad \beta_j \geq 0. \end{aligned} \quad (21)$$

where P denotes the total number of subjects in the validation set. Eq. (21) can be solved in the same way as introduced in Section 3. Once a fixed β is learned, it will be applied to all the subjects in the test sets to perform classification. This method is called ‘FixedTrainC’.

4. Experimental Study

4.1. Data Preprocessing and Experimental Settings

Two rs-fMRI data sets are used to verify the effectiveness of the proposed method SASNI. One data set is ADHD-200 provided by the Neuro Bureau for differentiating Attention Deficit Hyperactivity Disorder (ADHD) from healthy control subjects. ADHD-200 consists of 768 training subjects and 197 test subjects¹ collected from eight independent imaging sites. The summary of this data set is provided in Table 2. The rs-fMRI data are processed with Athena pipeline. Specifically, the first four echo-planar imaging (EPI) volumes are removed for signal equilibrium and then slice timing, orientation and motion correction are performed. Each rs-fMRI image is co-registered to T1 image and warped into MNI space at $4 \times 4 \times 4 \text{ mm}^3$ resolution. The time series of 90 brain nodes in gray matter are extracted from the preprocessed data using the automated anatomical labeling (AAL) [40] atlas. Detailed preprocessing descriptions and the processed time series are available at Neuro Bureau website².

The other data set is ADNI data set downloaded from the following website <http://adni.loni.usc.edu> with the aim of identifying Mild Cognitive Impairment (MCI), which is very early stage of Alzheimer’s disease, from healthy controls. There are 38 healthy controls and 44 MCIs. The data are acquired on a 3 Tesla (Philips) scanner with TR/TE set as 3000/30 ms and flip angle of 80° . Each series has 140 volumes, and each volume consists of 48 slices of 64×64 dimensional image matrices at $3.31 \times 3.31 \times 3.31 \text{ mm}^3$ resolution. The preprocessing is carried out using SPM8³ and DPARSFA [41]. The first 10 volumes of each series are discarded for signal equilibrium. Similar with ADHD-200, slice timing, head motion correction and MNI space normalization are performed. Participants with too

¹The labels of 26 subjects from Brown University in the test set are not released yet. They are not included in our performance evaluation.

²<http://neurobureau.projects.nitrc.org/ADHD200/Introduction.html>

³<http://www.fil.ion.ucl.ac.uk/spm/software/>

much head motion are excluded. The normalized brain images are warped into AAL atlas to obtain 90 ROIs as nodes. The ROI mean time series are extracted by averaging the time series from all voxels within each ROI and then band-pass filtered to obtain the most discriminative frequency band as in [42].

For both of the ADHD-200 and ADNI data sets, the functional connectivity networks are obtained by the SICE method using SLEP [28]. Nine sparsity levels of SICE matrices are obtained for each subject by setting $\lambda = [0.1 : 0.1 : 0.9]$. For the ADHD-200 data set, the predefined training/test sets are used while a leave-one-out procedure is used for ADNI data set to make full use of the limited subjects. For both of the data sets, the dictionary is made up by all the SICE matrices of the training subjects. The parameters used in the classification tasks of these two data sets, including θ in the SPD kernels, γ , and the regularization parameter of SVM are tuned by using five-fold cross-validation on the training set.

4.2. Experimental Result

This experiment consists of the following parts:

1. Evaluation of using a single sparsity level on ADHD-200 data set;
2. Evaluation of the proposed SASNI on ADHD-200 data set;
3. Comparison with other state-of-the-art methods on ADHD-200 data set;
4. Comparison with other integration schemes on ADHD-200 data set;
5. Comparison with non-subject-adaptive integration variants on ADHD-200 data set;
6. Evaluation of the proposed SASNI on ADNI data set;
7. Visualization of the integrated brain networks;
8. Convergence Evaluation;
9. Extension of Application.

4.2.1. Evaluation of using a single sparsity level on ADHD-200 data set

Before applying the proposed method to integrating different sparsity levels of brain networks, we first evaluate the performance of a single sparsity level of brain network. In this case, the brain network of one subject is represented as a single SPD matrix, which can be used to train a classifier in multiple manners. The following three common manners are evaluated in this experiment. A straightforward way is to extract graphical features from the network, such as the local clustering coefficient (LCC) feature in [42], and then train SVM with these features. LCC, as a measure of local neighborhood connectivity for a node, is defined as the ratio of the number of existing edges between the neighbors of the node and the number of possible connections between these neighbors [43]. In this case, LCC can map a network, represented by a $d \times d$ adjacency matrix, to a d -dimensional vector, where d is the number of nodes in the network. The second manner is employing one of the five SPD kernels in Table 1 to directly evaluate the similarity between SICE matrices and adopt SVM classifier with

these kernels to perform classification. The third manner is using SRC instead of SVM as the classifier with the same SPD kernel. Figure 3 shows the classification results of these methods on each of nine sparsity levels of SICE networks. As seen, the feature of LCC (labeled by ‘SVM (LCC)’) produces poor classification with most sparsity levels. When SPD kernels are used, either with SVM or SRC, the classification performance can be improved. In particular, the highest accuracy is achieved by SRC with the Stein kernel (SK), reaching over 69% on the 7th sparsity level.

This demonstrates that: 1) graphical feature of LCC does not sufficiently convey the discriminative information contained in SICE matrices; 2) SPD kernels can achieve reasonably good classification performance by considering the manifold property; 3) using SRC as the classifier and SK as the SPD kernel admits promising classification performance. The good performance verifies the effectiveness of SRC when dealing with brain network classification.

It is also worth noting that the discriminative power of different sparsity levels are not same. The sparser (4th-9th) levels are generally more discriminative than denser (1st-3rd) ones. This consolidates our motivation that combining the complementary information of different sparsity levels could benefit the classification performance.

4.2.2. Evaluation of the proposed SASNI on ADHD-200 data set

SRC with SK achieves the best classification performance when a single sparsity level is used. It is considered as a baseline and compared to the proposed SASNI method. In this experiment, we would like to investigate whether the classification performance can be improved by integrating multiple sparsity levels via SASNI. Note that, SK is also used as the SPD kernel in SASNI. As seen in Figure 4, the yellow bars indicate the classification performance of SASNI when different sparsity levels of brain networks are integrated. The tick ‘[1, n]’ on the x-axis means that the n densest levels of brain networks are integrated. Compared with SRC (with SK) using a single sparsity level, SASNI can consistently boost the classification performance for all integration settings. When the top four dense levels are integrated, SASNI achieves an accuracy more than 71%. When the top seven dense levels of SICE matrices are integrated, SASNI achieves an accuracy of over 72.5%, obtaining an improvement of three percentage points over SRC with the best single sparsity level.

This demonstrates that integration of multiple sparsity levels could attain more discriminative power and in turn improve the classification performance. As for which sparsity levels to be integrated, they can be selected by cross-validation on the training data set. An alternative way is just integrating all the sparsity levels since, as seen in Figure 4, the performance of SASNI is insensitive to the combination range when more than six sparsity levels are integrated.

Figure 4: Comparison of classification performance between single sparsity level SRC and the proposed SASNI on ADHD-200 data set.

4.2.3. Comparison with the state-of-the-art methods on ADHD-200 data set

To the best of our knowledge, in the literature, the highest classification accuracy on ADHD-200 data set is reported in [44]. That work computed the pairwise correlation of the time series between brain voxels to model the brain network and extracted multiple kinds of features from the network. Then a PCA-LDA based classifier is trained with the extracted features. Table 3 provides the classification performance obtained in the literature [44, 45] and by our proposed SASNI on the whole test set from all imaging sites. As seen, both the proposed SASNI and [44] outperform [45] by a large margin in terms of the overall classification performance. Our proposed SASNI method further exceeds [44] by about three percentage points on the overall test set. In particular, SASNI achieves better performance on OHSU and Peking test sets and perform equally on NeuroImage and NYU test sets in comparison with [44]. On the KKI and Pittsburgh test sets, the performance of SASNI is worse than that of [44]. Note that the number of subjects in KKI and Pittsburgh test sets is only 11 and 9, respectively. The absolute difference is only 1~2 subjects. On other larger test sets, SASNI consistently achieve better or equal performance in comparison with [44].

Table 3: Comparison of classification performance (in%) between the-state-of-the-art methods and the proposed SASNI on ADHD-200 data set.

	# subject	[45]	[44]	SASNI
OHSU	34	82.4	73.5	79.4
Peking	51	58.8	62.7	74.5
NeuroImage	25	48.0	72	72
NYU	41	-	70.7	70.7
KKI	11	54.6	72.7	63.6
Pittsburgh	9	-	77.8	55.6
Overall	171	62.8	69.6	72.5

★The classification performance of [45] on NYU and Pittsburgh data sets are not reported. Also note that the overall accuracy is the accuracy directly calculated over the 171 subjects from all the six data sets.

4.2.4. Comparison with other integration schemes on ADHD-200 data set

An experiment is carried out to compare the classification performance of the proposed SASNI method with other integration methods on ADHD-200 data set. These include MKL and a straightforward concatenation of LCC features from different sparsity levels (denoted as LCC in Figure 5). As reported in Figure 5, SASNI consistently outperforms both LCC feature concatenation method and five MKL methods (using each of the five SPD kernel functions in turn) once the top three or more sparsity levels are integrated.

Another integration approach is a late fusion method where a mixture of classifiers are used and each one of them looks at only one source (a single value of λ). Although extensive literature [46] has demonstrated that early fusion, e.g. feature selection and feature combination, often outperforms late fusion methods, e.g. major voting, we conduct additional experiments to compare the proposed SASNI, as an early fusion method, with two late fusion methods, including:

1. Majority Voting. An SRC classifier is trained for each value of λ and the final predicted label is determined by major voting;
2. 2-layer SVM. An SVM classifier is trained for each value of λ and the prediction scores for multiple λ s from the same subject are concatenated as a feature vector to train a second layer SVM for final classification.

Table 4: Comparison of classification performance (in%) between late fusion methods and the proposed SASNI on ADHD-200 data set.

Late fusion methods	Major Voting	65.5
	2-layer SVM	63.2
Early fusion	SASNI (proposed)	72.5

Top seven sparsity levels are used in this experiment. As seen in Table 4, the proposed SASNI outperforms the other two late fusion methods by a large margin.

Figure 5: Comparison of classification performance between different integration schemes on ADHD-200 data set.

4.2.5. Comparison with non-subject-adaptive integration variants on ADHD-200 data set

In this experiment, the proposed SASNI is compared with two non-subject-adaptive variants of SASNI, i.e. ‘MeanC’ and ‘FixedTrainC’ introduced in Section 3.5.3, to investigate the effectiveness of subject-adaptive integration. For ‘FixedTrainC’, the training data set is split into two equal-sized folds as the dictionary and the validation set, respectively. And the partition of the training data set is repeated 10 times to accumulate statistics. The averaged classification performance is reported in Figure 6 in comparison with that of the ‘MeanC’ and the proposed SASNI method. As seen, the SASNI method achieves the best classification performance in comparison with the two non-subject-adaptive integration methods. This verifies the advantage of the subject-adaptive mechanism. Note that each of the non-subject-adaptive integration methods outperforms SRC using the best single sparsity level, which indicates that the improved performance of the proposed SASNI method is attributed to both integration of multiple sparsity levels and the subject-adaptive mechanism.

4.2.6. Evaluation of the proposed method on ADNI data set

As previously mentioned, a leave-one-out procedure is used in the classification on ADNI data set, and the averaged accuracy is reported. As seen in Figure 7, the superiority of the proposed SASNI over a single sparsity level based SRC is confirmed again on ADNI data set. Specifically, when all the nine sparsity levels are integrated, SASNI outperforms the best single sparsity level based SRC by a large margin of over five percentage points.

4.2.7. Visualization of the integrated brain networks

An example of recovered pre-image is shown in Figure 8. The 90 ROIs of AAL atlas are grouped into eight lobes according to the anatomical structure. As seen, evident block-wise structures are presented. This is expected since the ROIs in the same lobe have higher chances to be anatomically and functionally connected. Also, most of the functional abnormalities

identified in the literature [45, 8] are within lobes. Especially, as indicated in Figure 8, the most strongest connections include connections within frontal gyrus, e.g. Superior frontal gyrus dorsolateral-Middle frontal gyrus (4, 8), Middle frontal gyrus orbital part-Inferior frontal gyrus opercular part (9, 11) and connections between Hippocampus (37~40), Amygdala (41~42) and Calcarine (43~44). An interesting thing this figure presented is that the same parts in left and right brain are often closely functionally connected. For example, the connections between Inferior frontal gyrus orbital part in left brain and Inferior frontal gyrus orbital part in right brain (15, 16), and the connection between Rolandic operculum in left brain and Rolandic operculum in right brain (17, 18). This is probably due to the extensive cooperation between the left and right brain in many functions. Since the experimental study has demonstrated that the integrated network possesses more discriminative power in comparison with a network corresponding to a single sparsity level, the recovered pre-image in the original space may reveal more disease-related connectivity patterns. In this sense, the visualization of the integrated brain networks provides medical specialist a new perspective to conduct analysis of the brain networks and this may promote understanding the underlying pathophysiology of brain diseases.

4.2.8. Convergence Evaluation

As discussed in Section 3.5.2, the alternate optimization of the sparse representation coefficient α and the combination coefficient β is guaranteed to converge. Here, we would like to verify the evolution of the objective function defined in Eq.(11). The evolution of the objective values averaged over all samples in the test set is plotted in Figure 9. As seen, the objective is monotonically decreased by optimizing β and α alternately. Moreover, the objective value decreases significantly in the first few iterations and quickly becomes convergent. This result experimentally demonstrates that the proposed optimization method can be effectively and efficiently solved.

Figure 6: Comparison between non-subject-adaptive integration methods and the proposed SASNI on ADHD-200 data set. ‘MeanC’ indicates average combination while ‘FixedTrainC’ indicates that a set of fixed integration coefficients is learned with the training data set and uniformly applied to all test subjects.

Figure 7: Comparison of classification performance between a single sparsity level based SRC and the proposed SASNI on ADNI data set.

Figure 8: An example of the recovered pre-image of the integrated network.

4.2.9. Extension of Application

In order to demonstrate the generality of the proposed SASNI method, we apply it to another data set on action recognition, which is called MSR-Action3D human action data set. MSR-Action3D contains 20 categories of actions from ten subjects. Each action is performed two or three times by each subject. For this data set, only skeleton data are used. Following the literature on human action recognition [47], the cross-subject test setting is used, i.e., the odd-indexed subjects for training and the even-indexed ones for test. Figure 10 presents the comparison of classification performance between single sparsity level SVM/SRC, MKL and the proposed SASNI on this data set. As seen, SRC achieves higher classification performance than SVM on most single sparsity levels. When multiple levels are combined, SASNI can further boost the accuracy of SRC and outperform MKL. Note that, this paper focuses on analyzing brain network data, as indicated by its title. Therefore, the results on human action data here are only used to demonstrate

Figure 9: Evolution of the objective function with respect to the number of iterations.

the generality of our proposed methods.

5. Conclusion

Recently, sparse inverse covariance estimation (SICE) has been widely employed to model brain connectivity networks and naturally leads to a set of connectivity networks with different sparsity for each subject. To explore the complementary information in the set of networks, we propose a learning framework that integrates brain networks and respects the underlying manifold structure of the SPD-based network representations. The proposed framework conducts a subject-adaptive integration via a kernel sparse learning scheme, and the obtained integrated network representation can be projected back in the original space for medical-related exploration. The effectiveness of the proposed method is verified on both ADHD and ADNI data sets. The experimental results demonstrate that the proposed integration method considerably outperforms a single network based methods and other commonly used integration methods. Our future work will follow two directions. The first direction is to apply the proposed method to other brain disease diagnosis and the second direction is to explore nonlinear integration of a set of networks.

- [1] V. D. Heuvel, M. P. H. E. Hulshoff Pol, Exploring the brain network: a review on resting-state fMRI functional connectivity, *European Neuropsychopharmacology* 20 (8) (2010) 519–534.
- [2] R. L. Buckner, J. L. Vincent, Unrest at rest: default activity and spontaneous network correlations, *Neuroimage* 37 (4) (2007) 1091–1096.
- [3] M. D. Greicius, B. Krasnow, A. L. Reiss, V. Menon, Functional connectivity in the resting brain: a network analysis of the default mode hypothesis, *Proceedings of the National Academy of Sciences* 100 (1) (2003) 253–258.
- [4] Y. Kong, Y. Deng, Q. Dai, Discriminative clustering and feature selection for brain mri segmentation, *IEEE Signal Processing Letters* 22 (5) (2015) 573–577.
- [5] G. Varoquaux, F. Baronnet, A. Kleinschmidt, P. Fillard, B. Thirion, Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling, in: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, Springer, 2010, pp. 200–208.
- [6] G. Deshpande, P. Wang, D. Rangaprakash, B. Wilamowski, Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data, *Cybernetics, IEEE Transactions on PP* (99) (2015) 1–1. doi:10.1109/TCYB.2014.2379621.
- [7] X. Li, A. Sroubek, M. S. Kelly, I. Lesser, E. Sussman, Y. He, C. Branch, J. J. Foxe, Atypical pulvinar–cortical pathways during sustained attention performance in children with attention-deficit/hyperactivity disorder, *Journal of the American Academy of Child & Adolescent Psychiatry* 51 (11) (2012) 1197–1207.
- [8] B. de Celis Alonso, S. H. Tobón, P. D. Suarez, J. G. Flores, B. de Celis Carrillo, E. B. Pérez, A multi-methodological mr resting state network analysis to assess the changes in brain physiology of children with adhd, *PLoS ONE* 6. doi:10.1371/journal.pone.0099119.
- [9] R. Gould, B. Arroyo, R. Brown, A. Owen, E. Bullmore, R. Howard, Brain mechanisms of successful compensation during learning in alzheimer disease, *Neurology* 67 (6) (2006) 1011–1017.
- [10] V. Calhoun, T. Adali, G. Pearson, J. Pekar, A method for making group inferences from functional mri data using independent component analysis, *Human brain mapping* 14 (3) (2001) 140–151.
- [11] J. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. Stam, S. M. Smith, C. Beckmann, Consistent resting-state networks across healthy subjects, *Proceedings of the National Academy of Sciences* 103 (37) (2006) 13848–13853.
- [12] M. van den Heuvel, R. Mandl, H. H. Pol, Normalized cut group clustering of resting-state fmri data, *PloS one* 3 (4) (2008) e2001.
- [13] S. M. Smith, D. Vidaurre, C. F. Beckmann, M. F. Glasser, M. Jenkinson, K. L. Miller, T. E. Nichols, E. C. Robinson, G. Salimi-Khorshidi, M. W. Woolrich, et al., Functional connectomics from resting-state fMRI, *Trends in cognitive sciences* 17 (12) (2013) 666–682.
- [14] S. M. Smith, The future of fMRI connectivity, *Neuroimage* 62 (2) (2012) 1257–1266.
- [15] H. Lee, D. S. Lee, H. Kang, B.-N. Kim, M. K. Chung, Sparse brain network recovery under compressed sensing, *Medical Imaging, IEEE Transactions on* 30 (5) (2011) 1154–1165.
- [16] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* 9 (3) (2008) 432–441.
- [17] J. Kim, J. R. Wozniak, B. A. Mueller, W. Pan, Testing group differences in brain functional connectivity: using correlations or partial correlations?, *Brain connectivity* 5 (4) (2015) 214–231.
- [18] A. Qiu, A. Lee, M. Tan, M. K. Chung, Manifold learning on brain functional networks in aging, *Medical image analysis* 20 (1) (2015) 52–60.
- [19] B. Ng, G. Varoquaux, J. B. Poline, B. Thirion, A novel sparse group gaussian graphical model for functional connectivity estimation, in: *Information Processing in Medical Imaging*, Springer, 2013, pp. 256–267.
- [20] R. P. Monti, P. Hellyer, D. Sharp, R. Leech, C. Agnostonopoulos, G. Montana, Estimating time-varying brain connectivity networks from functional mri time series, *Neuroimage* 103 (2014) 427–443.
- [21] S. Huang, J. Li, L. Sun, J. Ye, A. Fleisher, T. Wu, K. Chen, E. Reiman, Learning brain connectivity of alzheimer’s disease by sparse inverse covariance estimation, *NeuroImage* 50 (3) (2010) 935–949.
- [22] J. W. Bohland, S. Saperstein, F. Pereira, J. Rapin, L. Grady, Network, anatomical, and non-imaging measures for the prediction of adhd diagnosis in individual subjects, *Frontiers in systems neuroscience* 6.
- [23] M. J. Rosa, L. Portugal, T. Hahn, A. J. Fallgatter, M. I. Garrido, J. Shawe-Taylor, J. Mourao-Miranda, Sparse network-based models for patient classification using fmri, *NeuroImage* 105 (2015) 493–506.
- [24] L. Zhou, L. Wang, P. Ogunbona, Discriminative sparse inverse covariance matrix: Application in brain functional network classification, in: *Computer Vision and Pattern Recognition (CVPR)*, 2014 IEEE Conference on, 2014, pp. 3097–3104. doi:10.1109/CVPR.2014.396.
- [25] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Log-euclidean metrics for fast and simple calculus on diffusion tensors, *Magnetic Resonance in Medicine* 56 (2) (2006) 411–421. doi:10.1002/mrm.20965. URL <http://dx.doi.org/10.1002/mrm.20965>
- [26] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, M. Harandi, Kernel methods on the riemannian manifold of symmetric positive definite matrices, in: *Computer Vision and Pattern Recognition*, IEEE Computer Society Conference on, 2013.
- [27] S. Sra, A new metric on the manifold of kernel matrices with application to matrix geometric means, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* 25, Curran Associates, Inc., 2012, pp. 144–152.
- [28] J. Liu, S. Ji, J. Ye, SLEP: Sparse Learning with Efficient Projections, Arizona State University (2009). URL <http://www.public.asu.edu/~jye02/Software/SLEP>
- [29] M. Harandi, R. Hartley, B. Lovell, C. Sanderson, Sparse coding on symmetric positive definite manifolds using bregman divergences, *Neural Networks and Learning Systems*, IEEE Transactions on PP (99) (2015) 1–1. doi:10.1109/TNNLS.2014.2387383.
- [30] P. Li, Q. Wang, W. Zuo, L. Zhang, Log-euclidean kernels for sparse representation and dictionary learning, in: *Computer Vision (ICCV)*, 2013 IEEE International Conference on, IEEE, 2013, pp. 1601–1608.
- [31] I. L. Dryden, A. Koloydenko, D. Zhou, Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging, *The Annals of Applied Statistics* (2009) 1102–1123.
- [32] H. Salehian, G. Cheng, B. C. Vemuri, J. Ho, Recursive estimation of the stein center of spd matrices & its applications, *ICCV2013, International conference on computer vision*, 2013 IEEE Transactions on . (.) (2013) .
- [33] A. Alavi, M. T. Harandi, C. Sanderson, Relational divergence based classification on riemannian manifolds, in: *Applications of Computer Vision (WACV)*, 2013 IEEE Workshop on, IEEE, 2013, pp. 111–116.
- [34] M. Harandi, M. Salzmann, F. Porikli, Bregman divergences for infinite dimensional covariance matrices, *arXiv preprint arXiv:1403.4334*.

Figure 10: Comparison of classification performance between single sparsity level SVM/SRC, MKL and the proposed SASNI on MSR-Action3D data set.

- [35] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31 (2) (2009) 210–227. doi:10.1109/TPAMI.2008.79.
- [36] S. Gao, I. W. Tsang, L.-T. Chia, Sparse representation with kernels, *Image Processing, IEEE Transactions on* 22 (2) (2013) 423–434.
- [37] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding, *The Journal of Machine Learning Research* 11 (2010) 19–60.
- [38] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, Simplemkl, *Journal of Machine Learning Research* 9 (2008) 2491–2521.
- [39] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, *The Journal of Machine Learning Research* 12 (2011) 2211–2268.
- [40] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, *Neuroimage* 15 (1) (2002) 273–289.
- [41] Y. Chao-Gan, Z. Yu-Feng, DPARSF: a matlab toolbox for pipeline data analysis of resting-state fMRI, *Frontiers in systems neuroscience* 4.
- [42] C.-Y. Wee, P.-T. Yap, D. Zhang, L. Wang, D. Shen, Constrained sparse functional connectivity networks for MCI classification, in: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, Springer, 2012, pp. 212–219.
- [43] M. Kaiser, A tutorial in connectome analysis: topological and spatial features of brain networks, *Neuroimage* 57 (3) (2011) 892–907.
- [44] S. Dey, A. R. Rao, M. Shah, Exploiting the brain’s network structure in identifying adhd subjects, *Frontiers in systems neuroscience* 6.
- [45] S. Dey, A. R. Rao, M. Shah, Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects, *Frontiers in neural circuits* 8.
- [46] H. Gunes, M. Piccardi, Affect recognition from face and body: early fusion vs. late fusion, in: *2005 IEEE international conference on systems, man and cybernetics*, Vol. 4, IEEE, 2005, pp. 3437–3443.
- [47] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: *CVPRW, IEEE*, 2010, pp. 9–14.