Semi-Supervised Discriminative Classification Robust to Sample-Outliers and Feature-Noises

Ehsan Adeli, *Member, IEEE,* Kim-Han Thung, Le An, Guorong Wu, *Member, IEEE,* Feng Shi, Tao Wang, and Dinggang Shen*, *Fellow, IEEE*

Abstract—Discriminative methods commonly produce models with relatively good generalization abilities. However, this advantage is challenged in real-world applications (*e.g.*, medical image analysis problems), in which there often exist outlier data points (*sample-outliers*) and noises in the predictor values (*feature-noises*). Methods robust to both types of these deviations are somewhat overlooked in the literature. We further argue that denoising can be more effective, if we learn the model using all the available labeled and unlabeled samples, as the intrinsic geometry of the sample manifold can be better constructed using more data points. In this paper, we propose a semi-supervised robust discriminative classification method based on the least-squares formulation of linear discriminant analysis to detect sample-outliers and feature-noises simultaneously, using both labeled training and unlabeled testing data. We conduct several experiments on a synthetic, some benchmark semi-supervised learning, and two brain neurodegenerative disease diagnosis datasets (for Parkinson's and Alzheimer's diseases). Specifically for the application of neurodegenerative diseases diagnosis, incorporating robust machine learning methods can be of great benefit, due to the noisy nature of neuroimaging data. Our results show that our method outperforms the baseline and several state-of-the-art methods, in terms of both accuracy and the area under the ROC curve.

Index Terms—Linear discriminant analysis, semi-supervised learning, robust classification, feature selection, sample outlier detection, Alzheimer's disease, Parkinson's disease, biomarker identification, disease diagnosis, nuclear norm, regularization.

1 Introduction

Discriminative methods learn a mapping from the input feature space to the output label space for a task of classification (or regression). Such methods usually achieve good classification (or regression) results, compared to the generative methods, when there is enough number of training samples. But they carry out limited abilities when there are a small number of labeled data [1]. On the other hand, when noise contaminates the data, discriminative models usually fail to find an optimal mapping. In many real-world applications, the data are usually contaminated by different levels of noise. In some cases, a whole bunch of samples are affected (*e.g.*, deviations in neuroimaging data due to radiation or patient movements during the imaging process), and therefore not useful for the learning task. These types of deviations are often denoted as *sample-outliers*. On the other hand, sometimes only some specific predictor values or features are infected, known as intra-sample-outliers (or *feature-noises*).

Various efforts have been made to add robustness to different learning methods. For instance, Suzumaura *et al*. [2] and Xu *et al*. [3] introduced robustness to the conventional support vector machine formulation by proposing various regularization terms or suppressing the influence of the outliers. In other works, Kim *et al*. [4] and Croux *et al*. [5] proposed robust variations of Fisher/Linear Discriminant Analysis (LDA) method, and Li *et al*. [6] introduced a worst-case LDA, by minimizing the upper bound of the LDA cost function. These methods are all robust to *sample-outliers*. On the other hand, some methods were proposed to deal with the *feature-noises*, such as [7, 8]. Many previous methods use Robust Principal Component Analysis (RPCA) [9], to deal with feature-noises in an unsupervised manner. Furthermore, many robust approaches that denoise the data while training the model do not offer straightforward strategies to deal with the testing data. Often, the denoising procedure of the training and the testing data are conducted separately (*e.g.*, in [10]), which might induce a bias to the whole learning process. One solution is to denoise the training and the testing data together, provided that the testing data are available. Therefore, we propose to take advantage of them as unlabeled data during the training phase. Under such semi-supervised setting, the constructed discriminative model can be more reliable, particularly for the cases with the small-sample-size problem. This could be attributed to the fact that more samples are being used to model the intrinsic geometry of the sample manifold.

The main application we are anticipating in this paper

E. Adeli is with Stanford University, Stanford, CA 94305, USA.
E. Adeli, K.-H. Thung, L. An, G. Wu, and D. Shen are with the Biomedical Research Imaging Center (BRIC) and Department of Radiology, University of North Carolina at Chapel Hill, NC, 27599, USA.
F. Shi is with the Biomedical Imaging Research Institute, Cedars Sinai Medical Center, Los Angeles, CA 90048, USA.
T. Wang is with the Department of Geriatric Psychiatry, Shanghai Mental Health Center and the Alzheimer's Disease and Related Disorders Center, Shanghai Jiao Tong University, Shanghai, China.
D. Shen is also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea.
**Corresponding author: D. Shen (email: dgshen@med.unc.edu).*
Parts of the data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu), and Parkinson's Progression Markers Initiative (PPMI) database (http://www.ppmi-info.org). The investigators within the ADNI or PPMI contributed to the design and implementation of the datasets and/or provided data but did not participate in analysis or writing of this paper. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/howtoapply/ADNIAcknowledgementList.pdf

is the diagnosis of neurodegenerative diseases, based on neuroimaging data. This is a challenging problem, as the data is pretty much prone to noise and often there is a limited number of samples. Hence, there is a calling need for robust machine learning methods for such applications. Neurodegenrative diseases are debilitating and incurable conditions caused by progressive degeneration or death of the cells in the brain nervous system. These diseases affect millions of people around the world. Alzheimer's Disease (AD) and Parkinson's Disease (PD) are among the most common types. Although neurodegenerative diseases manifest with diverse pathological features, the cellular level processes resemble similar structures. For instance, in AD, deposits of tiny protein plaques result into brain damage and progressive loss of memory [11], while PD is mainly initiated by a selective loss of dopaminergic neurons in the Substantia Nigra (SN) brain region, leading to declining in the generation of a chemical messenger, dopamine. Lack of dopamine yields loss of ability to control body movements, along with several non-motor problems (*e.g.*, depression, and anxiety) [12]. These diseases are often incurable; thus, early diagnosis and treatment are crucial to slow down their progression in the initial stages.

The challenges for building reliable diagnosis models include: (1) It is usually burdensome to acquire noise-free imaging data from the patients. Different sources of noise may affect the acquired data, including a wide variety of noises in the neuroimage acquisition procedure, the imposed artifacts due to preprocessing, and the large amount of inter-subject variabilities; (2) To build a good diagnosis model, through learning a classifier, we need a sufficiently large number of labeled subjects. However, acquiring reliably enough labeled data is costly and time-consuming. Therefore, models that can take advantage of unlabeled data (subjects that we are not certain about their disease) could be of great interest; (3) Different neurodegenerative diseases often affect different regions of the brain, *i.e.*, only certain regions of the brain are associated with the disease. Thus, using all features can undermine the diagnosis performance, and we need to identify the imaging biomarkers for each specific disease while learning the diagnosis model.

To deal with the aforementioned challenges, we propose a semi-supervised discriminative classifier, to take advantage of the available unlabeled testing data. This leads to a more substantial number of samples, which can yield better modeling of the intrinsic geometry of the sample manifold. As a result, our model jointly estimates the noise model (both sample-outliers and feature-noises) on the whole labeled training and unlabeled testing data and simultaneously builds a discriminative model upon the denoised training data. Unlike many previous works on denoising medical images, we do not define the problem of denoising separately from the analysis part. In the sense that if a sample (or a feature value) does not act in accordance with others in building the model, it should be counted as a sample-outlier (or a feature-noise). This observation suggests that intertwining the denoising procedure with the learning framework will help to identify the sample-outliers and feature-noises more efficiently while learning a robust classification model. It is important to note that denoising and outlier detection has a long history in the area of medical image analysis and computing. The inter- and intra-subject variabilities, the noise sourced from the images devices, and the pre-processing errors emerge the study of robust methods for analyzing medical imaging data. For instance, in the recent years, several attempts have been made for denoising the medical images [13–16] or detecting outliers [17, 18], as a preprocessing step to any analysis on medical images.

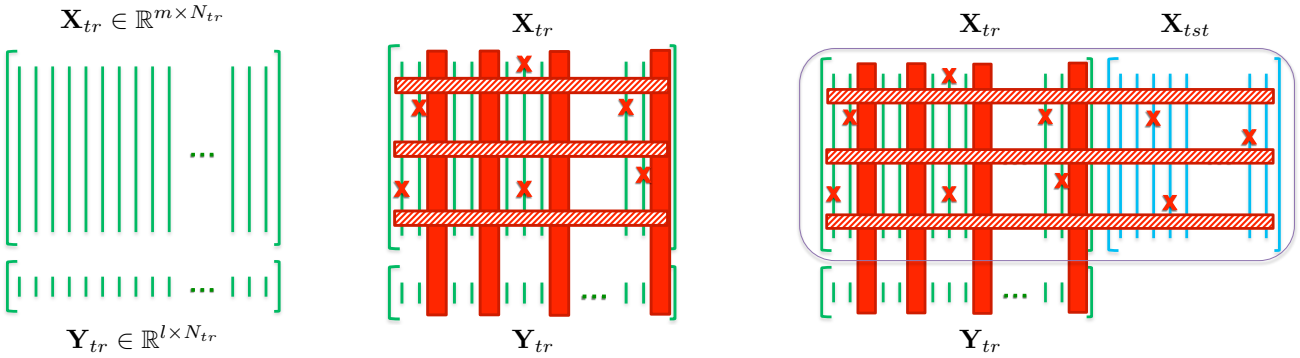
1.1 Background and Overview of the Proposed Method

In this paper, we introduce a novel classification model based on LDA, which is robust against both sample-outliers and feature-noises, referred to as robust feature-sample linear discriminant analysis (RFS-LDA). The original LDA formulation finds the mapping between the sample space and the label space through a linear transformation matrix, maximizing a so-called Fisher discriminant ratio [4]. In practice, the major drawback of LDA is the small-sample-size problem, which arises when the number of available training samples is much less than the dimensionality of the feature space [19]. Original LDA finds the mapping by incorporating covariance matrices of the input feature matrices. In cases where the number of samples is much less than the number of features, these matrices are probably rank-deficient [20]. A reformulation of LDA based on the reduced-rank least-squares problem (known as LS-LDA) [20] tackles this problem. LS-LDA finds the mapping $\boldsymbol{\beta} \in \mathbb{R}^{l\times m}$ by solving the following problem:

$$\min_{\boldsymbol{\beta}} \|(\mathbf{Y}_{tr}\mathbf{Y}_{tr}^{\top})^{-1/2}(\mathbf{Y}_{tr} - \boldsymbol{\beta}\mathbf{X}_{tr})\|_{\mathrm{F}}^2, \tag{1}$$

where $\mathbf{Y}_{tr} \in \mathbb{R}^{l\times N_{tr}}$ is a binary class label indicator matrix, for l different classes (or labels), and $\mathbf{X}_{tr} \in \mathbb{R}^{m\times N_{tr}}$ is the matrix containing N_{tr} m-dimensional training samples. $(\mathbf{Y}_{tr}\mathbf{Y}_{tr}^{\top})^{-1/2}$ is a normalization factor that compensates for the different number of samples in each class [20]. As a result, the mapping $\boldsymbol{\beta}$ is a reduced rank transformation matrix [8, 20], which could be used to project a test data $\mathbf{x}_{tst} \in \mathbb{R}^{m\times 1}$ onto an l-dimensional space. Note that directly minimizing (1) avoids the small-sample-size problem by not using the covariance matrices. After it projects the samples to the output space, we need a simple step to infer the class labels. LDA maximizes inter-class variance, while minimizing the intra-class variance, in the mapped space. Thus, we expect that in the mapped space, same-class samples to be closer to each other. The class labels could, therefore, be simply determined using a k-NN strategy.

To make LDA robust against noisy data, Fidler *et al*. [7] estimate a robust basis, which consists all the discriminative information for classification or regression. In the testing phase, the estimated basis identifies the outliers in samples (images in their case) and then calculates the coefficients using a subsampling approach. On the other hand, Huang *et al*. [8] proposed a general formulation for Robust Regression (RR) and classification (*i.e.*, Robust LDA or RLDA), where, they first denoise the training feature values using a strategy similar to RPCA [9], and then build the above LS-LDA model using the denoised data. In the testing stage, they denoise the testing samples using the denoised training data. This separate denoising procedure could not effectively form the underlying geometry of sample space to



(a) A traditional supervised learning task, defined as finding a mapping from training samples, $\mathbf{X}_{tr}$, to their respective labels, $\mathbf{Y}_{tr}$.

(b) A robust classifier needs to deal with both sample-outliers (red bars) and feature-noises (red crosses). In some applications (like ours), discarding non-discriminant features (shaded bars) is also crucial.

(c) In order to form the underlying geometry of the sample manifold with largely enough number of samples, we incorporate the unlabeled testing data. This could contribute to denoising both the training and testing samples, while learning the mapping from the training samples to their respective labels.

Fig. 1: Overview of the proposed semi-supervised learning framework, robust to both sample-outliers and feature-noises.

denoise the data. Furthermore, RR [8] only accounts for feature-noises by imposing a sparse noise model constraint on the features matrix, despite the fact that the least-squares data fitting term in (1) is vulnerable to large sample-outliers.

Recently, in robust statistics, it is found that ℓ_1 functions are able to make more reliable estimations [21] than ℓ_2 least-squares fitting functions. This has been previously adopted in many applications, including robust face recognition [22] and robust dictionary learning [23]. Reformulating the objective in (1) with ℓ_1 loss entails the following problem:

$$\min_{\boldsymbol{\beta}} \|(\mathbf{Y}_{tr}\mathbf{Y}_{tr}^{\top})^{-1/2}(\mathbf{Y}_{tr} - \boldsymbol{\beta}\mathbf{X}_{tr})\|_1. \tag{2}$$

We incorporate this fitting function to deal with the sample-outliers, in this paper. We also adopt a strategy to simultaneously denoise the data from feature-noises. This is done through a semi-supervised setting to take advantage of all labeled and unlabeled data, and build the structure of the sample space more robustly. Figure 1 illustrates this idea, in which Fig. 1a shows a traditional learning problem. However, if the data contains sample-outliers or some samples suffer from noise in their feature values (Fig. 1b), traditional methods usually fail to build reliable models.

Semi-supervised learning has long been of great interest in different fields, because it can make use of unlabeled or poorly labeled data to achieve better prediction models [24, 25]. For instance, Joulin and Bach [26] introduced a convex relaxation and used their model in different semi-supervised learning scenarios. In another work, Cai *et al*. [27] proposed a semi-supervised discriminant analysis, where the separation between different classes is maximized using the labeled data points, while the unlabeled data points estimate the structure of the data. Belkin *et al*. [28] similarly used the unlabeled data for regularization. In contrast, we incorporate the unlabeled testing data in our formulation to better estimate the intrinsic geometry of the sample manifold and denoise the data, while building the discriminative model upon the labeled training data. By incorporating the unlabeled testing data (Fig. 1c), we learn the classification model, while denoising both training and testing data and detecting sample-outliers.

We apply our method for the diagnosis of neurodegenerative brain disorders. Specifically, in this study, we use two popular databases: PPMI [29] and ADNI [30]. The former aims at investigating PD and its related disorders, while the latter is designed for diagnosing AD and its prodromal stage, known as Mild-Cognitive Impairment (MCI). In addition, to validate the proposed method, we further conduct experiments on synthetic data, as well as some benchmark datasets for semi-supervised learning.

1.2 Contributions

The contributions of this paper are multi-fold: (1) We propose an approach to dealing with the sample-outliers and feature-noises simultaneously and build a robust discriminative classifier. The sample-outliers are penalized through an ℓ_1 fitting function. (2) Our proposed model operates under a semi-supervised setting, where the whole data (*i.e.*, labeled training, and unlabeled testing samples) are incorporated to build the intrinsic geometry of the sample space, which leads to better data denoising. (3) We further select the most discriminative features for the learning process through regularizing the weights matrix with an ℓ_1 norm. This is especially of great interest for the neurodegenerative disease diagnosis, where the features from different regions of the brain are extracted, but not all the regions are associated with a certain disease. Thus, the most discriminative regions associated with the disease would be identified, leading to a more reliable diagnosis model.

2 The Proposed Method: RFS-LDA

Suppose we have N_{tr} training and N_{tst} testing samples, each with a m dimensional feature vector, which leads to a set of $N = N_{tr} + N_{tst}$ total samples. Let $\mathbf{X} \in \mathbb{R}^{m \times N}$ denote the set of all samples (both training and testing), in which each column indicates a single sample, and also let $\mathbf{y}_i \in \mathbb{R}^{1 \times N}$ their corresponding i^{th} labels. In general, with l different labels, we can define $\mathbf{Y} \in \mathbb{R}^{l \times N}$. Thus, $\mathbf{X}$ and $\mathbf{Y}$ are composed by stacking up the training and testing data

as: $\mathbf{X} = [\mathbf{X}_{tr}\ \mathbf{X}_{tst}]$ and $\mathbf{Y} = [\mathbf{Y}_{tr}\ \mathbf{Y}_{tst}]$. Our goal is to determine the labels of the test samples, $\mathbf{Y}_{tst} \in \mathbb{R}^{l \times N_{tst}}$.

Note that, throughout the paper, bold capital letters denote matrices (*e.g.*, $\mathbf{A}$), while bold lowercase letters denote vectors (*e.g.*, $\mathbf{a}$). All non-bold letters denote scalar variables. a_{ij} is the scalar in the row i and column j of $\mathbf{A}$. $\langle \mathbf{a}_1, \mathbf{a}_2 \rangle$ denotes the inner product between $\mathbf{a}_1$ and $\mathbf{a}_2$. $\|\mathbf{a}\|_2^2 = \langle \mathbf{a}, \mathbf{a} \rangle = \sum_i a_i^2$ and $\|\mathbf{a}\|_1 = \sum_i |a_i|$ represent the squared Euclidean norm and the ℓ_1 norm of $\mathbf{a}$, respectively. $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^\top \mathbf{A}) = \sum_{ij} a_{ij}$, $\|\mathbf{A}\|_{1,1} = \sum_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_*$ designate the squared Frobenius norm, $\ell_{1,1}$ norm and the nuclear norm (sum of singular values) of $\mathbf{A}$, respectively. $\mathbf{I}_K \in \mathbb{R}^{K \times K}$ denotes the identity matrix.

2.1 Formulation

All the available samples, both labeled and unlabeled, are arranged into a matrix, $\mathbf{X} \in \mathbb{R}^{m \times N}$, each of whose columns represents the feature vector of a sample. To achieve a robust classifier, we seek to denoise this matrix. Following [31, 32], this could be done by assuming that $\mathbf{X}$ can be spanned on a low-rank subspace and therefore should be rank-deficient. This assumption supports the fact that samples from same classes are more correlated [8, 32] and linearly-dependent. Accordingly, the original matrix $\mathbf{X}$ is decomposed into the summation of two counterparts, $\mathbf{D} \in \mathbb{R}^{m \times N}$ and $\mathbf{E} \in \mathbb{R}^{m \times N}$. The former represents the denoised data matrix, while the latter is the error matrix. This is similar to RPCA [9], used in many computer vision applications. With this decomposition, we can assume that the denoised data matrix shall be rank-deficient and the error matrix sparse.

But as one can easily infer, this process of denoising does not incorporate the label information and is, therefore, unsupervised. Nevertheless, recall that we are also seeking a mapping between the denoised training samples and their respective labels. So, matrix $\mathbf{D}$ should be spanned on a low-rank subspace that would lead to a good classification model of its sub-matrix, $\mathbf{D}_{tr}$. We incorporate the regression model in (2) as the fitting function to compute a mapping $\boldsymbol{\beta}$. A schematic illustration of the proposed method is depicted in Fig. S1 of the supplementary material.

To ensure the rank-deficiency of the matrix $\mathbf{D}$, like many previous works [9, 31, 32], we approximate the rank function using the nuclear norm (*i.e.*, the sum of the singular values of the matrix). The noise is modeled using the ℓ_1 norm of the matrix, which ensures a sparse noise model on the feature values. Accordingly, the objective function for RFS-LDA under a semi-supervised setting would be formed as:

$$\begin{aligned} \min_{\boldsymbol{\beta},\mathbf{D},\hat{\mathbf{D}},\mathbf{E}} \quad & \frac{\eta}{2}\|\mathbf{Y}_{tr} - \boldsymbol{\beta}\hat{\mathbf{D}}\|_1 + \|\mathbf{D}\|_* + \lambda_1\|\mathbf{E}\|_1 + \lambda_2\mathcal{R}(\boldsymbol{\beta}), \\ s.t. \quad & \mathbf{X} = \mathbf{D} + \mathbf{E}, \hat{\mathbf{D}} = [\mathbf{D}_{tr}; \mathbf{1}^\top], \end{aligned} \tag{3}$$

where the first term is the ℓ_1 regression model introduced in (2). This term only operates on the denoised training samples from matrix $\mathbf{D}$ with a row of all 1's added to it (denoted as $\hat{\mathbf{D}}$), to counter for the bias in the linear model. The second and third terms, together with the first constraint, are similar to the RPCA formulation [9]. They denoise the labeled training and unlabeled testing data together, and in combination with the first term, we ensure that the denoised data also specifies a favorable regression. The last term is a regularization on the learned mapping coefficients, to avoid trivial or unexpectedly large values. The hyperparameters η, λ_1 and λ_2 are the scalar regularization hyperparameters, which will be discussed in detail later.

The regularization on the coefficients could be posed as a simple norm of the matrix, $\boldsymbol{\beta}$. But, in many applications, like ours (disease diagnosis), many of the features in the feature vectors are redundant. This is because we extract features from different brain regions, but not all the regions contribute to a certain disease. Therefore, it is desirable to determine which features are the most relevant and the most discriminative for the task. Following [11, 22, 33], we are seeking a sparse set of weights that ensures incorporating the most discriminative features. Therefore, we propose a regularization on the weights matrix as a combination of the ℓ_1 and Frobenius norms:

$$\mathcal{R}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_{1,1} + \gamma\|\boldsymbol{\beta}\|_{\mathrm{F}}. \tag{4}$$

Evidently, the solution to the objective function in (3) is not easy to achieve. This is because it contains a quadratic term, and the minimization of the ℓ_1 fitting function is not straightforward, due to its indifferentiability. To this end, we formalize the solution with a similar strategy as in Iteratively Re-weighted Least Squares (IRLS) [21]. The ℓ_1 fitting term is approximated by a conventional ℓ_2 least-squares, in which each of the samples in the $\hat{\mathbf{D}}$ matrix is weighted with the reverse of their regression residual. Additionally, since we regularize the weights $\boldsymbol{\beta}$ using a combination of ℓ_1 and ℓ_2 norms, the non-zero elements would represent the selected features by the algorithm. In order to reflect this to feature denoising scheme, we define a projection operator $\mathcal{P}_{\boldsymbol{\beta}}(.)$. This operator projects the values of the non-selected features (respective to zero values in $\boldsymbol{\beta}$) to zero, to decrease their effect in minimizing the rank of the matrix $\mathbf{D}$ (in the second term). Therefore, the new problem would be:

$$\begin{aligned} \min_{\boldsymbol{\beta},\mathbf{D},\hat{\mathbf{D}},\mathbf{E}} \quad & \frac{\eta}{2}\|(\mathbf{Y}_{tr} - \boldsymbol{\beta}\hat{\mathbf{D}})\hat{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\boldsymbol{\beta}}(\mathbf{D})\|_* + \lambda_1\|\mathbf{E}\|_1 + \lambda_2\mathcal{R}(\boldsymbol{\beta}), \\ s.t. \quad & \mathbf{X} = \mathbf{D} + \mathbf{E}, \hat{\mathbf{D}} = [\mathbf{D}_{tr}; \mathbf{1}^\top]. \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\alpha}}$ is a diagonal matrix, the i^{th} diagonal element of which is the i^{th} sample's weight:

$$\hat{\boldsymbol{\alpha}}_{ii} = \frac{1}{\sqrt{(\mathbf{y}_i - \boldsymbol{\beta}\hat{\mathbf{d}}_i)^2 + \delta}}, \forall\ i, j \in \{0, \ldots, N_{tr}\}, i \neq j, \hat{\boldsymbol{\alpha}}_{ij} = 0. \tag{6}$$

Hyperparameter δ is a small positive number (10^{-4} in our experiments), to prevent from any division by zeros in (6). In the next subsection, we introduce an algorithm to solve this optimization problem.

Our work is closely related to the RR formulations in [8], where the authors impose a low-rank assumption on the training data feature values and an ℓ_1 assumption on the noise model. The discriminant model is learned similarly to LS-LDA, as described in (1). Whereas, we observed that to have a more robust regression model, we need to establish a strategy where we can weight the samples. This is because the ℓ_1 noise model in [8] can only discard a controlled amount of sparse noise in the feature values, not the whole samples. On the other hand, our model operates under a semi-supervised setting, where both labeled training and unlabeled testing samples are denoised simultaneously, leading to a more robust denoising model. Also, our model

further selects the most discriminative features to learn the regression model, by regularizing the learned weights and enforcing a sparsity condition on them.

To optimize the objective function in (5), we use the Alternating Direction Method of Multipliers (ADMM) [34]. The detailed optimization steps, along with the comprehensive analysis of the algorithm, its convergence properties and an upper bound for the time complexity of the proposed algorithm are provided in the supplementary material.

3 Experiments

To evaluate the proposed approach, we compare our method against several baselines and state-of-the-art methods in different scenarios. The first experiment evaluates our method on a synthetic set of data, which highlights how the proposed method is robust against sample-outliers or feature-noises separately, or when they occur at the same time. Then we employ some benchmark semi-supervised learning datasets and report results in comparisons with some baseline and state-of-the-art methods. The results of these two experiments (*i.e.*, on **synthetic** and **benchmark data**) are reported in the **supplementary material**. We then apply the proposed RFS-LDA method to the problem of neurodegenerative brain disorder and disease diagnosis.

For the choice of hyperparameters, a set of possible values are first predefined, and the best hyperparameters are selected through 10-fold cross-validation, for all the competing methods. The RFS-LDA hyperparameters (as in Eq. (5)) are set with the same strategy as in [8]:

$$\lambda_1 = \frac{\Lambda_1}{(\sqrt{\min(m,N)})},\ \lambda_2 = \frac{\Lambda_2}{\sqrt{m}},\ \eta^k = \frac{\Lambda_3\|\mathbf{X}\|_*}{\|\mathbf{Y}_{tr} - \boldsymbol{\beta}^k\hat{\mathbf{D}}^k\|_{\mathrm{F}}^2}, \tag{7}$$

and ρ (controlling the $\{\mu\}$s in the iterative optimization algorithm) is set to 1.01. We have set $\Lambda_1, \Lambda_2, \Lambda_3$ and γ through inner-cross-validation grid-search in the range $[10^{-4}, 10]$.

3.1 Datasets

In this study, we use two real-world databases for two different brain neurodegenerative diseases, namely PD and AD. The first set of data is obtained from the Parkinson's Progression Markers Initiative (PPMI) database [29], with the MRI data from 374 PD and 169 normal control (NC) subjects. The second dataset comes from the Alzheimer's disease neuroimaging initiative (ADNI) database, which includes MRI and FDG-PET data. We used 93 AD patients, 202 MCI patients, and 101 NC subjects, each with complete MRI and FDG-PET data. The subjects' brain images are preprocessed and regions of interest (ROI) features are extracted for each subject. For more detailed information about these two datasets and the preprocessing steps for feature extraction refer to the **supplementary material**.

3.2 Baseline Methods

We compare our proposed method with different baseline methods, including the conventional LS-LDA [20], RLDA [8], and linear Support Vector Machine (SVM). Another baseline method can be defined as running the same procedures as in the proposed method but disjointly. Therefore, we apply RPCA on the matrix $\mathbf{X}$ separately to first denoise, and then classify the denoised data using LS-LDA (denoted as RPCA+LS-LDA) [8]. To analyze the effectiveness of the feature selection strategy of the proposed method, we also include baseline methods which use sparse feature selection (SFS) together with SVM (SFS+SVM), and RLDA (SFS+RLDA). Except for RPCA+LDA, the other methods in comparison do not incorporate the testing data. In order to have a fair set of comparisons, we also compare against the transductive Matrix Completion (MC) approach [32] and the semi-supervised formulation of SVM (S^3VM) [35]. These two methods incorporate the unlabeled testing data in the process of training their models. Additionally, in order to further evaluate the effect of the ℓ_1 norm regularization on the weights matrix $\boldsymbol{\beta}$, we also report results for RFS-LDA when regularized by only $\gamma\|\boldsymbol{\beta}\|_{\mathrm{F}}$ (denoted as RFS-LDA*), rather than the regularization term introduced in (4). Finally, we report results using the supervised version of our proposed method, which is denoted as supervised RFS-LDA (S-RFS-LDA). In S-RFS-LDA, we train our model using only the training data, where $\mathbf{X}$ in (5) is replaced with $\mathbf{X}_{tr}$. In this way, we can examine the effect of using unlabeled testing data in the prediction model.

3.3 Disease Diagnosis

We evaluate our method with two popular datasets for neurodegenerative disease diagnosis, PPMI and ADNI, for diagnosis of PD and AD, respectively. These datasets, subject information, preprocessing steps, and feature extraction are explained in Section C of the supplementary material.

Results: The first row in Table 1 shows the diagnostic accuracy of the proposed technique (RFS-LDA) in comparisons with different baseline and state-of-the-art methods using 10-fold cross-validation. The results show that the proposed method outperforms all others. This can be attributed to the fact that our method better deals with feature-noises and sample-outliers. Recall that samples and their corresponding feature vectors extracted from the neuroimaging data are quite prone to noise, as discussed earlier. Therefore, some of the samples might not be useful, and some might be contaminated by a certain amount of noise. Our method can deal with both types of noises, as supported by the results. The second disease diagnosis experiment is conducted on ADNI, in which the goal is to discriminate normal controls (NC) from mild cognitive impairment (MCI) and AD subjects. Therefore, NC subjects form our negative class, while the positive class is defined as AD in one experiment, and MCI in the other. The diagnosis results of the AD *vs*. NC and MCI *vs*. NC are reported in the second and third rows in Table 1, respectively. As it can be seen, in comparison with the state-of-the-art, our method achieves better results in terms of both accuracy and the area under ROC curve.

It is worth noting that running the model using a 10-fold cross-validation for the PD *vs*. NC (543 subjects), AD *vs*. NC (194 subjects), and MCI *vs*. NC (303 subjects) experiments on a PC (Intel® Core™ i7 @ 2.30 GHz and 8.00 GB of memory), with a parallel implementation in MATLAB® (*i.e.*, using `parfor` for 4 workers) took approximately 6, 2 and 3.5 hours, respectively. Additionally, to test the statistical significance of the obtained results, we further conducted a Fisher exact test [36] on the accuracy score achieved by each of the

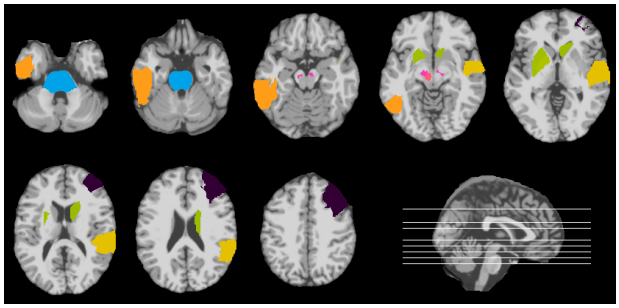
TABLE 1: Diagnosis accuracy of the proposed method (RFS-LDA) and the baseline methods on both PPMI and ADNI datasets. The † sign indicates a p-value > 0.05 in a Fisher exact test.

	RFS-LDA	RFS-LDA*	S-RFS-LDA	RLDA	SFS+RLDA	RPCA+LS-LDA	LS-LDA	SVM	SFS+SVM	S^{3}VM	MC
PD *vs.* NC	**79.8**	76.1	74.1	68.3	70.5	61.0†	58.5†	66.1†	69.2	71.5	70.6
AD *vs.* NC	**92.1**	89.8	88.3	87.8	90.0	87.6	82.7	85.4	87.1	**90.5**	88.7
MCI *vs.* NC	**81.9**	80.6	79.1	79.5	**80.9**	76.9	72.3†	74.1	78.3	**80.8**	76.1

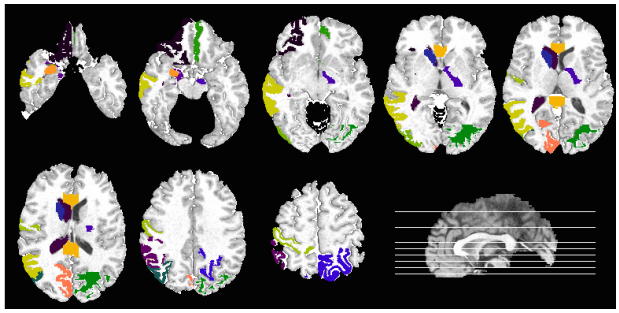
methods. This test verifies that the method is significantly more accurate (with a p-value of $p < 0.05$) than randomly assigning the samples to the two classes. The results of this statistical test indicated that the proposed method achieves a p-value of even less than 0.001. This shows that there are no random associations with the obtained results. However, for some of the compared baseline methods, a p-value of $p > 0.05$ was observed, which is not appealing. These methods are marked with a † sign in Table 1. It is important to note that the comparisons between the supervised (S-RFS-LDA) and the semi-supervised (RFS-LDA) versions of the proposed algorithm in both Table 1 and Figure S5 of the supplementary material show that including the unlabeled testing data improves the results by a relatively notable margin. This can be because including more samples gives us a better representation of the sample manifold, leading to better denoising of simultaneous training and testing data, in a way that a better classifier is built.

Although the studies on Parkinson's disease using modern machine learning techniques are scarce, there are quite a few studies in the literature for Alzheimer's disease. State-of-the-art machine learning approaches for this purpose either aim at developing feature selection techniques or focus on designing delicate classifiers. The first type usually use sophisticated techniques for feature selection [37, 38], feature learning [39], or feature extraction[40–42] and then an straightforward classification technique (like SVM) is utilized. The second type develops task-specific classifiers to enhance the classification accuracies, *e.g.* [43–45]. In contrast, our method constructs the sample manifold using all labeled and unlabeled data to denoise the features and also selects the best features for classification, with a classification loss robust to sample-outliers. In Table 2, we compare our method with several state-of-the-art methods for Alzheimer's disease diagnosis. The table includes all the information about the dataset and the methods they used for obtaining those results. This is only to show where our method stands among the previous works in the same field.

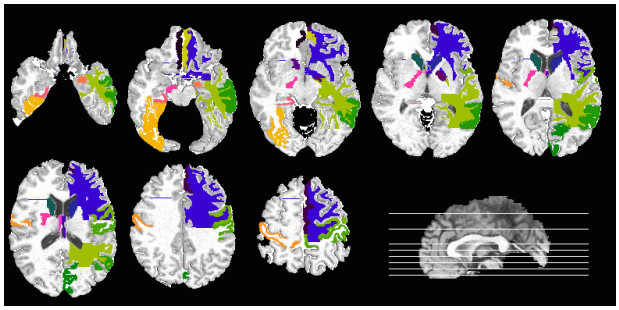
As discussed earlier, in medical imaging applications many sources of noise contribute to the acquired data, and therefore methods that can deal with noise and outliers are of great interest. Our method enjoys from a single optimization objective that can simultaneously suppress sample-outliers and feature-noises, which, compared to other methods, exhibits a good performance. One of the interesting functions of the proposed method is the regularization on the mapping coefficients with the ℓ_1 norm, which would select a compact set of features to contribute to the learned mapping. The magnitude of the coefficients would show the relevance of the specific features for building the prediction model. In our application, the features from the whole brain regions are extracted, but not all the ROIs are associated with the disease (*e.g.*, AD, MCI or PD). By exploring the learned coefficients by our method, we can determine which brain regions are highly associated with a certain disease.



(a) PD *vs.* NC



(b) AD *vs.* NC



(c) MCI *vs.* NC

Fig. 2: Top selected regions for each experiment. Selected regions are shown with different colors for clarity.

Identification of Disease Biomarkers: To extract these most relevant ROIs, we select the ROIs that were given larger weights in 50% of the ten repetitions of the 10-fold cross-validation tests. Fig. 2(a) visualizes the most relevant regions for PD on a raw brain template, including the middle frontal gyrus right, pons, substrata nigra left and right, red nucleus left, pallidum left, pautmen left, caudate right, inferior temporal left, and superior temporal gyrus right. As in the previous studies in the literature [46, 47], deep brain and striatum areas are known to play crucial roles for PD. Our study also confirms these clinical findings. Same experimental settings for AD and MCI identifies

TABLE 2: Comparisons of the proposed method with state-of-the-art methods for diagnosis of AD and MCI. N/A: indicates that the methods did not report results for that experiment; CSF: cerebrospinal fluid; Gen: categorical genetic information.

Method	Subjects			Methodology	Modalities	AD *vs.* NC (%)	MCI *vs.* NC (%)
	AD	MCI	NC				
Liu *et al.* [45]	198	N/A	229	Voxel GM+SVM Ensemble	MRI	**92.0**	N/A
Cuingnet *et al.* [42]	137	N/A	162	Voxel Direct D+SVM	MRI	88.58	N/A
Eskildsen *et al.* [41]	194	N/A	226	Cortical Thickness+SVM	MRI	84.50	N/A
Duche. *et al.* [40]	75	N/A	75	Tensor-based Morphometry+SVM	MRI	**92.0**	N/A
Min *et al.* [37]	97	N/A	128	Multi-Atlas ROI Features+SVM	MRI	91.6	N/A
Gary *et al.* [44]	37	75	35	Random Forest	MRI+PET+CSF+Gen	89.0	74.6
Tong *et al.* [43]	35	75	77	Graph Fusion	MRI+PET+CSF+Gen	91.8	79.5
Liu *et al.* [39]	85	169	77	Deep Feature Learning	MRI+PET	91.4	**82.1**
Ours	93	202	101	RFS-LDA	MRI+PET	**92.1**	**81.9**

the top regions selected by our algorithm in AD *vs.* NC and MCI *vs.* NC classification scenarios (Figs. 2(b) and (c), respectively). These regions, including middle temporal gyrus, medial front-orbital gyrus, postcentral gyrus, caudate nucleus, cuneus, and amygdala, have also been reported to be associated with AD and MCI in the literature [11, 48]. The analysis of such selection of brain regions can be further incorporated for future clinical studies.

Method Discussions: To analyze the effect of the sample-outlier detection in the proposed framework, we employ a dimensionality reduction technique to facilitate the visualization of the data points. We project the samples of the AD *vs.* NC experiment into the 2-D space using t-SNE [49]. The t-SNE projection technique visualizes high-dimensional data by giving each sample a location in a two-dimensional map. The map created by the t-SNE reveals the neighborhood structure of the sample manifold at many different scales [49]. This is particularly important for our application, in which the high-dimensional neuroimaging data lie on several different low-dimensional manifolds since the samples come from different subjects with or without the neurodegenerative disease. Fig. 3 shows the t-SNE projection in the 2-D space. In this figure, the samples, which received the smallest weights in their respective elements in the $\hat{\boldsymbol{\alpha}}$ weight matrix (as in Equation (5)), are shown in the top part of the figure. We also depict the samples detected as outliers using the RANSAC [50] algorithm in the bottom part of the figure. Notably, as it is obvious in the figure, the samples detected as sample-outliers by our algorithm are those which are more controversial for the task of classification and lie outside the main neighborhood of each class. This is attributed to the fact that we detect them jointly with the classifier learning framework. On the other hand, the outliers detected by RANSAC are not always the best in terms of discriminability. This suggests that unsupervised outlier detection methods might not perform well when the aim is to learn a classifier or a regression model. In other words, in many learning tasks, the definition for sample-outliers might be different based on what the goal is.

One of the important hyperparameters in the proposed RFS-LDA is λ_1, as in Eq. (5), which controls the noise term. Modifying this hyperparameter leads to altered noise levels, detected by our algorithm. To analyze its effect on the learning performance, we fix all other hyperparameters and run the algorithm with different values of Λ_1, and therefore λ_1 (as discussed at the beginning of Section 3). The changes in the AUC for each of our experiments are illustrated in Fig. 4. As can be seen, the proposed method achieves reasonably good results with a wide range of the values of the hyperparameter.

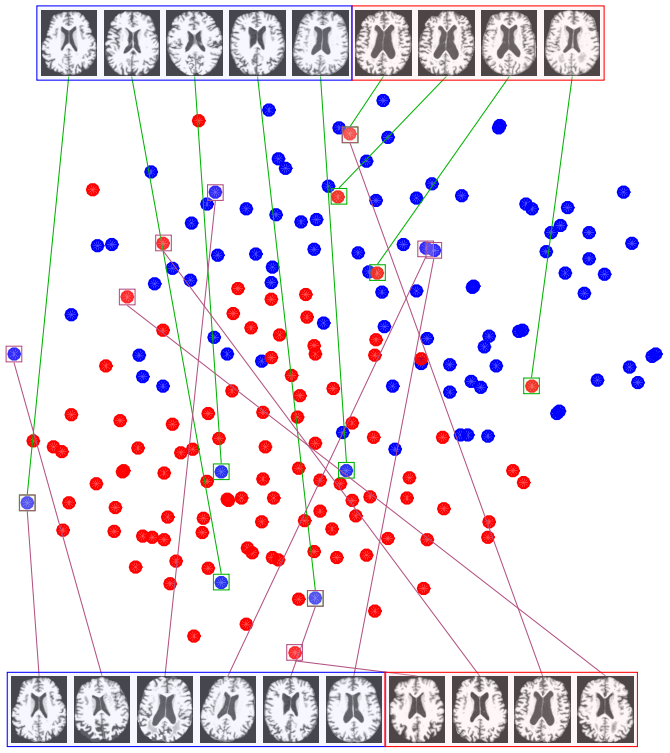


Fig. 3: t-SNE projection of AD *vs.* NC samples (better viewed in color). Top: Samples detected as outliers by our method. Bottom: Samples detected as outliers using RANSAC [50].

It is worth noting that the proposed method works under a semi-supervised setting, which can be quite interesting for the application of disease diagnosis. When performing the diagnosis for new patients, all subjects whose clinical diagnosis has not been finalized (*i.e.*, they are still in the process of evaluations and clinical monitoring) can yet be included in model building as unlabeled samples, to build a potentially more reliable classifier.

4 Conclusion

In this paper, we proposed a novel approach for discriminative classification, which is robust against both sample-

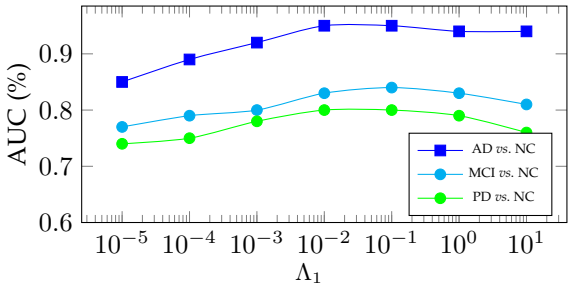


Fig. 4: Area under the ROC curve (AUC) as a function of the RFS-LDA hyperparameter Λ_1, related to λ_1 in (3).

outliers and feature-noises. Our method enjoys a semi-supervised setting, where all the labeled training and the unlabeled testing data are used to detect outliers and are denoised simultaneously. We have applied our method to several datasets, including synthetic, semi-supervised learning benchmark, and neurodegenerative brain disease diagnosis datasets, specifically for Parkinson's disease and Alzheimer's disease. The results showed that our method outperformed all competing techniques. As a direction for the future works, one can develop a multi-task learning reformulation of the proposed method to incorporate diagnosis from multiple modalities of neuroimaging data or extend the approach for the case of incomplete data.

REFERENCES

[1] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *NIPS*, vol. 14, p. 841, 2002.

[2] S. Suzumura, K. Ogawa, M. Sugiyama, and I. Takeuchi, "Outlier path: A homotopy algorithm for robust svm," in *ICML*, pp. 1098–1106, 2014.

[3] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *JMLR*, vol. 10, pp. 1485–1510, 2009.

[4] S.-J. Kim, A. Magnani, and S. Boyd, "Robust fisher discriminant analysis," in *NIPS*, pp. 659–666, 2005.

[5] C. Croux and C. Dehon, "Robust linear discriminant analysis using s-estimators," *Canadian J. of Statistics*, vol. 29, no. 3, pp. 473–493, 2001.

[6] H. Li, C. Shen, A. van den Hengel, and Q. Shi, "Worst-case linear discriminant analysis as scalable semidefinite feasibility problems," *IEEE TIP*, vol. 24, no. 8, 2015.

[7] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE TPAMI*, vol. 28, no. 3, pp. 337–350, 2006.

[8] D. Huang, R. Cabral, and F. De la Torre, "Robust regression," in *ECCV*, pp. 616–630, 2012.

[9] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[10] D. Huang, R. Cabral, and F. De la Torre, "Robust regression," *IEEE TPAMI*, 2015.

[11] K.-H. Thung, C.-Y. Wee, P.-T. Yap, and D. Shen, "Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion," *NeuroImage*, vol. 91, pp. 386–400, 2014.

[12] D. Ziegler and J. Augustinack, "Harnessing advances in structural MRI to enhance research on Parkinson's disease," *Imag. in med.*, vol. 5, no. 2, pp. 91–94, 2013.

[13] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot, "An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images," *IEEE TMI*, vol. 27, no. 4, pp. 425–41, 2008.

[14] I. Rodrigues, J. Sanches, and J. Bioucas-Dias, "Denoising of medical images corrupted by poisson noise," in *ICIP*, pp. 1756–1759, 2008.

[15] H. Bhadauria and M. Dewal, "Medical image denoising using adaptive fusion of curvelet transform and total variation," *Computers and Electrical Engineering*, vol. 39, no. 5, pp. 1451 – 1460, 2013.

[16] J. Manjón, P. Coupé, A. Buades, D. Louis Collins, and M. Robles, "New methods for mri denoising based on sparseness and self-similarity," *Med. Image Anal.*, vol. 16, no. 1, pp. 18–27, 2012.

[17] V. Fritsch, G. Varoquaux, B. Thyreau, J.-B. Poline, and B. Thirion, "Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators," *Med. Image Anal.*, vol. 16, no. 7, pp. 1359 – 1370, 2012.

[18] S. Mriaux, A. Roche, B. Thirion, and G. Dehaene-Lambertz, "Robust statistics for nonparametric group analysis in fmri," in *ISBI*, pp. 936–939, 2006.

[19] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," in *NIPS*, pp. 97–104, 2003.

[20] F. De la Torre, "A least-squares framework for component analysis," *IEEE TPAMI*, vol. 34, no. 6, pp. 1041–1055, 2012.

[21] N. Bissantz, L. Dümbgen, A. Munk, and B. Stratmann, "Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces," *SIAM J. on Optimization*, vol. 19, no. 4, pp. 1828–1845, 2009.

[22] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *CVPR*, pp. 597–604, 2009.

[23] C. Lu, J. Shi, and J. Jia, "Online robust dictionary learning," in *CVPR*, pp. 415–422, June 2013.

[24] O. Chapelle, B. Schölkopf, and A. Zien, eds., *Semi-Supervised Learning*. MIT Press, 2006.

[25] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[26] A. Joulin and F. Bach, "A convex relaxation for weakly supervised classifiers," in *ICML*, 2012.

[27] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *CVPR*, 2007.

[28] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *Learning Theory*, vol. 3120 of *LNCS*, pp. 624–638, 2004.

[29] K. Marek and *et al.*, "The parkinson progression marker initiative (PPMI)," *Progress in Neurobiology*, vol. 95, no. 4, pp. 629 – 635, 2011.

[30] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (adni)," *Alzheimer's & dementia: J. of the Alzheimer's Association*, vol. 1, pp. 55–66, 07 2005.

[31] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE TPAMI*, vol. 35, no. 1, pp. 171–184, 2013.

[32] A. Goldberg, X. Zhu, B. Recht, J.-M. Xu, and R. Nowak, "Transduction with matrix completion: Three birds with one stone," in *NIPS*, pp. 757–765, 2010.

[33] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *CVPR*, 2011.

[34] S. Boyd and *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[35] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *NIPS*, pp. 368–374, 1998.

[36] R. A. Fisher, "The logic of inductive inference," *Journal of the Royal Statistical Society*, vol. 98, no. 1, pp. 39–82, 1935.

[37] R. Min, G. Wu, J. Cheng, Q. Wang, and D. Shen, "Multi-atlas based representations for alzheimer's disease diagnosis," *Human brain mapping*, vol. 35, no. 10, pp. 5052–5070, 2014.

[38] M. Liu, D. Zhang, E. Adeli-Mosabbeb, and D. Shen, "Inherent structure based multi-view learning with multi-template feature representation for alzheimer's disease diagnosis," *IEEE TBME*, 2016.

[39] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, and M. J. Fulham, "Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 62, no. 4, pp. 1132–1140, 2015.

[40] S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins, "Mri-based automated computer classification of probable ad versus normal controls," *IEEE TMI*, vol. 27, no. 4, pp. 509–520, 2008.

[41] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, D. L. Collins, A. D. N. Initiative, *et al.*, "Prediction of alzheimer's disease in subjects with mild cognitive impairment from the adni cohort using patterns of cortical thinning," *Neuroimage*, vol. 65, pp. 511–521, 2013.

[42] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative, *et al.*, "Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database," *Neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[43] T. Tong, K. Gray, Q. Gao, L. Chen, and D. Rueckert, "Nonlinear graph fusion for multi-modal classification of alzheimer's disease," in *Machine Learning in Medical Imaging*, pp. 77–84, Springer, 2015.

[44] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert, A. D. N. Initiative, *et al.*, "Random forest-based similarity measures for multi-modal classification of alzheimer's disease," *NeuroImage*, vol. 65, pp. 167–175, 2013.

[45] M. Liu, D. Zhang, and D. Shen, "Hierarchical fusion of features and classifier decisions for alzheimer's disease diagnosis," *Human brain mapping*, vol. 35, no. 4, pp. 1305–1319, 2014.

[46] H. Braak, K. Tredici, U. Rub, R. de Vos, E. J. Steur, and E. Braak, "Staging of brain pathology related to sporadic parkinson's disease," *Neurobio. of Aging*, vol. 24, no. 2, pp. 197 – 211, 2003.

[47] A. Worker and *et al.*, "Cortical thickness, surface area and volume measures in parkinson's disease, multiple system atrophy and progressive supranuclear palsy," *PLOS ONE*, vol. 9, no. 12, 2014.

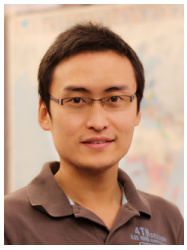
[48] B. Pearce, A. Palmer, D. Bowen, G. Wilcock, M. Esiri, and A. Davison, "Neurotransmitter dysfunction and atrophy of the caudate nucleus in alzheimer's disease," *Neurochem Pathol.*, vol. 2, no. 4, pp. 221–32, 1985.

[49] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *JMLR*, vol. 9, no. 2579-2605, p. 85, 2008.

[50] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

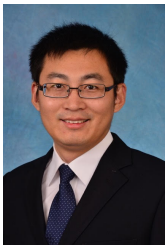
Ehsan Adeli is a postdoctoral research fellow at Stanford University. He received his Ph.D. from Iran University of Science and Technology, after which he worked as a postdoctoral researcher at the University of North Carolina at Chapel Hill. He has previously worked as a visiting research scholar at the Robotics Institute, Carnegie Mellon University in Pittsburgh, PA. Dr. Adeli's research interests include machine learning, computer vision, medical image analysis and computational neuroscience.

Kim-Han Thung obtained his PhD degree from the University of Malaya, Malaysia, in 2012. He is currently a post-doctoral research associate at the University of North Carolina at Chapel Hill, USA. He was a research assistant at University of Malaya, Malaysia and a software engineer at Motorola Penang, Malaysia. His research interests include sparse learning, multi-task learning, deep learning, matrix completion, machine learning using incomplete data and medical image analysis.



Le An received the B.Eng. degree in telecommunications engineering from Zhejiang University in China in 2006, the M.Sc. degree in electrical engineering from Eindhoven University of Technology in Netherlands in 2008, and the Ph.D. degree in electrical engineering from University of California, Riverside in USA in 2014. His research interests include image processing, computer vision, pattern recognition, and machine learning.

Guorong Wu received the Ph.D. degree in computer science and engineering from Shanghai Jiao Tong University, Shanghai, China. He is currently with the Department of Radiology, The University of North Carolina, Chapel Hill, as an Assistant Professor. His research interests focus on fast and robust analysis of large population data, computer-assisted diagnosis, and image-guided radiation therapy.



Feng Shi received his Bachelors Degree in Electronics Engineering from Peking University, Beijing, China in 2002, and obtained his PhD degree in Computer Science from Institute of Automation, Chinese Academy of Sciences in 2008. He was then trained as a postdoctoral research associate in Medical Image Analysis at the University of North Carolina at Chapel Hill, NC, and later appointed as an Assistant Professor. Currently, he is an Assistant Professor in Cedars-Sinai Medical Center, Los Angeles, CA since 2016. His research interests include machine learning, neuro and cardiac imaging, multimodal image analysis and computer-aided early diagnosis.

Tao Wang graduated from the Shanghai Jiaotong University School of Medicine with a doctoral degree in psychogeriatrics in 2012. He completed training as visiting scholar in UNC IDEA Group, Biomedical Research Imaging Center, UNC Chapel Hill, NC, USA in 2015-2016, and Banner Alzheimers Institute, Arizona, USA from June 2013 to July 2013. Since 2013 he has been an associate chief psychiatrist in the department of geriatric psychiatry, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine. Now he is the vice director of the department of geriatric psychiatry at the Shanghai Mental Health Center. His main research interest is neurocognitive disorder.

Dinggang Shen is Jeffrey Houpt Distinguished Investigator, and a Professor of Radiology, Biomedical Research Imaging Center (BRIC), Computer Science, and Biomedical Engineering in the University of North Carolina at Chapel Hill (UNC-CH). He is currently directing the Center for Image Analysis and Informatics, the Image Display, Enhancement, and Analysis (IDEA) Lab in the Department of Radiology, and also the medical image analysis core in the BRIC. He was a tenure-track assistant professor in the University of Pennsylvanian (UPenn), and a faculty member in the Johns Hopkins University. Dr. Shen's research interests include medical image analysis, computer vision, and pattern recognition. He has published more than 800 papers in the international journals and conference proceedings. He serves as an editorial board member for eight international journals. He has also served in the Board of Directors, The Medical Image Computing and Computer Assisted Intervention (MICCAI) Society, in 2012-2015. He is Fellow of IEEE, and also Fellow of The American Institute for Medical and Biological Engineering (AIMBE).

Semi-Supervised Discriminative Classification Robust to Sample-Outliers and Feature-Noises (Supplementary Material)

Ehsan Adeli, *Member, IEEE,* Kim-Han Thung, Le An, Guorong Wu, *Member, IEEE,* Feng Shi, Tao Wang, and Dinggang Shen*, *Fellow, IEEE,*

A Optimization

We optimize the objective function in (5) using the Alternating Direction Method of Multipliers (ADMM) [1]. To this end, we introduce the Lagrangian multipliers, $\mathscr{L}_1 \in \mathbb{R}^{m\times N}$, $\mathscr{L}_2 \in \mathbb{R}^{(m+1)\times N_{tr}}$ and $\mathscr{L}_3 \in \mathbb{R}^{l\times(m+1)}$, along with an auxiliary variable, $\mathbf{B} \in \mathbb{R}^{l\times(m+1)}$, and write the Lagrangian function $\mathcal{L}(\boldsymbol{\beta}, \mathbf{B}, \mathbf{D}, \hat{\mathbf{D}}, \mathbf{E}, \{\mu_i^k\}_{i=1}^3, \{\mathscr{L}_i^k\}_{i=1}^3)$ as:

$$\begin{aligned}&\frac{\eta}{2}\|(\mathbf{Y}_{tr} - \boldsymbol{\beta}\hat{\mathbf{D}})\hat{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\boldsymbol{\beta}}(\mathbf{D})\|_* + \lambda_1\|\mathbf{E}\|_1 \\ &+ \lambda_2(\|\mathbf{B}\|_{1,1} + \gamma\|\boldsymbol{\beta}\|_{\mathrm{F}}) + \langle\mathscr{L}_1, \mathbf{X} - \mathbf{D} - \mathbf{E}\rangle \\ &+ \frac{\mu_1}{2}\|\mathbf{X} - \mathbf{D} - \mathbf{E}\|_{\mathrm{F}}^2 + \langle\mathscr{L}_2, \hat{\mathbf{D}} - [\mathbf{D}_{tr}; \mathbf{1}^\top]\rangle \\ &+ \frac{\mu_2}{2}\|\hat{\mathbf{D}} - [\mathbf{D}_{tr}; \mathbf{1}^\top]\|_{\mathrm{F}}^2 + \langle\mathscr{L}_3, \boldsymbol{\beta} - \mathbf{B}\rangle + \frac{\mu_3}{2}\|\boldsymbol{\beta} - \mathbf{B}\|_{\mathrm{F}}^2,\end{aligned} \tag{S.1}$$

where μ_1, μ_2 and μ_3 are the penalty hyperparameters. There are five variables ($\boldsymbol{\beta}, \mathbf{B}, \mathbf{D}, \hat{\mathbf{D}}$ and $\mathbf{E}$) contributing to the problem. We alternatively optimize for each variable, while fixing the others. Except for the matrix $\boldsymbol{\beta}$, all the variables have straightforward or closed-form solutions. $\boldsymbol{\beta}$ is calculated through IRLS [2] by iteratively setting $\hat{\boldsymbol{\alpha}}$ weights and solving the conventional least-squares problem, until convergence.

To utilize ADMM for solving the above problem, we first initialize all the optimization variables, and then iteratively minimize the objective function with respect to each of the optimization variables, while keeping the others fixed. Note that, solving the subproblems for each optimization variable leads to a convex subproblem, which can be simply solved, and will be explained in more details in the following.

A.1 Initialization

Since the objective function with respect to each optimization variable is convex, the algorithm will converge, and the initialization would not influence it. But a good initialization will help to converge in a shorter amount of time. The variables are initialized by:

$$\begin{aligned}&\mathbf{D}^0 = [\mathbf{X}_{tr}\ \mathbf{X}_{tst}], \hat{\mathbf{D}}^0 = [\mathbf{X}_{tr}; \mathbf{1}^\top], \\ &\boldsymbol{\beta}^0 = \mathbf{Y}_{tr}(\hat{\mathbf{D}}^0)^\top(\hat{\mathbf{D}}^0(\hat{\mathbf{D}}^0)^\top + \gamma\mathbf{I}), \mathbf{E}^0 = \mathbf{0},\end{aligned} \tag{S.2}$$

the Lagrangian multipliers as:

$$\mathscr{L}_1^0 = \frac{\mathbf{X}}{\|\mathbf{X}\|_2}, \mathscr{L}_2^0 = \frac{\mathbf{X}_{tr}}{\|\mathbf{X}_{tr}\|_2}, \mathscr{L}_3^0 = \frac{\boldsymbol{\beta}^0}{\|\boldsymbol{\beta}^0\|_2}, \tag{S.3}$$

Algorithm 1 Optimization step for $\boldsymbol{\beta}$.

1: $t \leftarrow 0, \hat{\boldsymbol{\beta}}^0 = \boldsymbol{\beta}^k$
2: **repeat**
3: $\quad\forall\, i, j \in \{0, \ldots, N_{tr} - 1\}, i \neq j, \hat{\boldsymbol{\alpha}}_{ij}^t \leftarrow 0$
4: $\quad\hat{\boldsymbol{\alpha}}_{ii}^t \leftarrow 1/\sqrt{(\mathbf{y}_i^k - \hat{\boldsymbol{\beta}}^t\hat{\mathbf{a}}_i^k)^2 + 0.0001}$
5: $\quad\hat{\boldsymbol{\beta}}^{t+1} \leftarrow (\mathbf{Y}_{tr}\hat{\boldsymbol{\alpha}}^t\hat{\boldsymbol{\alpha}}^{t\top}(\hat{\mathbf{D}}^k)^\top + \mu_3(\mathbf{B}^k - \mathscr{L}_3^k))(\hat{\mathbf{D}}^k\hat{\boldsymbol{\alpha}}^t\hat{\boldsymbol{\alpha}}^{t\top}(\hat{\mathbf{D}}^k)^\top + \gamma\mathbf{I})$
6: $\quad t \leftarrow t + 1$
7: **until** $\|\hat{\boldsymbol{\beta}}^{t-1} - \hat{\boldsymbol{\beta}}^t\|_{\mathrm{F}}/(\|\hat{\boldsymbol{\beta}}^{t-1}\|_{\mathrm{F}} \times \|\hat{\boldsymbol{\beta}}^t\|_{\mathrm{F}}) < 0.001$ or $t > 100$
8: $\boldsymbol{\beta}^{k+1} \leftarrow \hat{\boldsymbol{\beta}}^t$.

and the penalty hyperparameters:

$$\mu_1^0 = \frac{mN}{4}\|\mathbf{X}\|_1, \mu_2^0 = \frac{mN_{tr}}{4}\|\mathbf{X}_{tr}\|_1, \mu_3^0 = \frac{ml}{4}\|\boldsymbol{\beta}^0\|_1. \tag{S.4}$$

A.2 Optimization step for $\boldsymbol{\beta}$

In each iteration k, the mapping $\boldsymbol{\beta}$ is calculated through IRLS [2], as an approximation for the ℓ_1 fitting function. For this purpose, all other variables associated with the optimization are fixed, and therefore the problem would be similar to solving a regularized least-squares problem, iteratively. The iterative algorithm to solve this subproblem is provided in Algorithm 1. As can be seen in Step 7 of the algorithm, the stopping criteria include the convergence of the iterative process (*i.e.*, $\boldsymbol{\beta}$ is converged), or the number of iterations exceeds a threshold (which is set to 100 in our experiments).

A.3 Optimization step for $\hat{\mathbf{D}}$

Fixing all other variables, and solving (S.1) for $\hat{\mathbf{D}}$, in the k^{th} iteration, would yield the following closed-form solution:

$$\begin{aligned}\hat{\mathbf{D}}^{k+1} \leftarrow\ &(\eta\hat{\boldsymbol{\alpha}}^\top(\boldsymbol{\beta}^{k+1})^\top\boldsymbol{\beta}^{k+1}\hat{\boldsymbol{\alpha}} + \mu_2^k\mathbf{I})^{-1} \\ &\times(\eta\hat{\boldsymbol{\alpha}}^\top(\boldsymbol{\beta}^{k+1})^\top\mathbf{Y}_{tr} - \mathscr{L}_2^k + \mu_2^k[\mathbf{D}_{tr}^k; \mathbf{1}^\top]),\end{aligned} \tag{S.5}$$

where $\mathbf{I}$ is the identity matrix.

A.4 Optimization step for $\mathbf{D}$

This step requires minimization of the nuclear norm of the matrix, $\mathbf{D}$. This is done using the popular singular value thresholding (SVT) algorithm [3], in which a thresholding

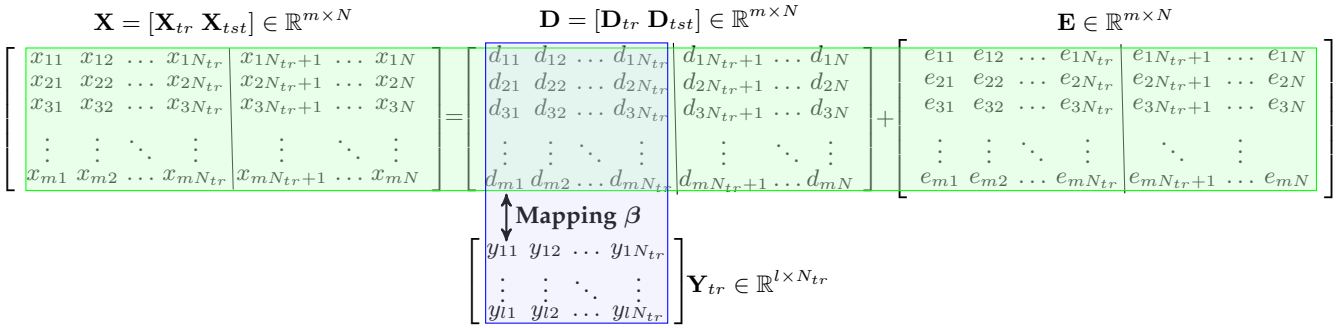


Fig. S1: Outline: The original data matrix, $\mathbf{X}$, is composed of both labeled training and unlabeled testing data. Our method decomposes this matrix to a denoised data matrix, $\mathbf{D}$, and an error matrix, $\mathbf{E}$, to account for *feature-noises*. Simultaneously, we learn a mapping, $\boldsymbol{\beta}$, from the denoised training samples in $\mathbf{D}$ ($\mathbf{D}_{tr}$) through a robust ℓ_1 fitting function, dealing with the *sample-outliers*. The same learned mapping on the testing data, $\mathbf{D}_{tst}$, leads to the test labels.

operator, $\mathcal{D}_\tau(.)$, is applied to the singular values of the matrix:

$$\begin{aligned}&\mathbf{D}^{k+1} \leftarrow \mathcal{D}_{\frac{1}{(\mu_1^k+\mu_2^k)}}(\mathbf{A}),\\ &\mathbf{A} = \mathscr{L}_1^k + \mu_1^k(\mathbf{X}-\mathbf{E}^k) + [[\mathscr{L}_2^k + \mu_2^k\mathcal{P}_{\boldsymbol{\beta}}(\hat{\mathbf{D}}^{k+1})]_{(1:N_{tr},:)}\ \mathbf{0}].\end{aligned} \tag{S.6}$$

Note that $\mathcal{D}_\tau(\mathbf{A}) = \mathbf{U}\mathcal{D}_\tau(\mathbf{\Sigma})\mathbf{V}^*$ applies SVT on the intermediate matrix $\mathbf{\Sigma}$, as $\mathcal{D}_\tau(\mathbf{\Sigma}) = \text{diag}(\{(\sigma_i-\tau)_+\})$, where $\mathbf{U\Sigma V}^*$ is the singular value decomposition (SVD) of $\mathbf{A}$, σ_i's are the singular values, and s_+ is defined as $s_+ = \max(0, s)$.

A.5 Optimization step for $\mathbf{E}$

This step involves minimization of the ℓ_1 norm of the matrix, that can be simply achieved using soft thresholding operator or the proximal operator for the ℓ_1 norm [1], $\mathcal{S}_\kappa(a) = (a-\kappa)_+ - (-a-\kappa)_+$. Therefore, the update state would be formed as:

$$\mathcal{P}_{\boldsymbol{\beta}}(\mathbf{E}^{k+1}) \leftarrow \mathcal{S}_{\lambda_1/\mu_1^k}(\mathbf{X} - \mathcal{P}_{\boldsymbol{\beta}}(\mathbf{D}^{k+1}) + \mathscr{L}_1^k/\mu_1^k). \tag{S.7}$$

$$\mathcal{P}_{\bar{\boldsymbol{\beta}}}(\mathbf{E}^{k+1}) \leftarrow \mathbf{0}. \tag{S.8}$$

A.6 Optimization step for $\mathbf{B}$

This step requires minimization of the ℓ_1 norm of each row i of the matrix:

$$\mathbf{B}_i^{k+1} \leftarrow \mathcal{S}_{\lambda_2/\mu_3^k}(\boldsymbol{\beta}_i^{k+1} + \mathscr{L}_3^k),\ \forall 1 \le i \le l. \tag{S.9}$$

A.7 Update steps for the Lagrangian multipliers

These update steps, as suggested by the ADMM approach, would be:

$$\begin{aligned}&\mathscr{L}_1^{k+1} \leftarrow \mathscr{L}_1^k + \mu_1^k(\mathbf{X} - \mathbf{D}^{k+1} - \mathbf{E}^{k+1}),\\ &\mathscr{L}_2^{k+1} \leftarrow \mathscr{L}_2^k + \mu_2^k(\hat{\mathbf{D}} - [\mathbf{D}_{tr}^{k+1}; \mathbf{1}^\top]),\\ &\mathscr{L}_3^{k+1} \leftarrow \mathscr{L}_3^k + \mu_3^k(\boldsymbol{\beta} - \mathbf{B}).\end{aligned} \tag{S.10}$$

A.8 Update steps for the penalty hyperparameters

As discussed in many previous works [1, 4, 5], the number of iterations of the whole algorithm until convergence is dependent on the choice of the penalty hyperparameters, the $\{\mu\}$s. They serve as step sizes on how fast to move towards the optima. If we select very small values for them, the solution will converge very slowly, whereas, if they have large values, the steps will be very big and might make us keep jumping over the optimum. If μ penalty hyperparameters are increasing smoothly in each iteration, the overall algorithm would be Q-linearly (quotient-linearly) [4] convergent. A reasonable choice for the sequence of all $\{\mu\}$s yields in a decrease in the number of required iterations. Similar to [4], we first initialize these hyperparameters to a small value, to take small steps at the beginning. In each next iteration, we increase them smoothly, so that we take larger steps towards the optima. This is because, at the beginning, the optimization process starts with randomly initialized variables and, if we take larger steps, we might be misleading the direction to the optimum point; hence increasing the convergence time. But, after a number of iterations, larger steps would lead to faster convergence.

$$\begin{aligned}&\mu_1^{k+1} \leftarrow \min(\rho\mu_1^k, 10^9),\\ &\mu_2^{k+1} \leftarrow \min(\rho\mu_2^k, 10^9),\\ &\mu_3^{k+1} \leftarrow \min(\rho\mu_3^k, 10^9).\end{aligned} \tag{S.11}$$

A.9 The stopping criteria

The stopping point of the algorithm is where all the constraints are met, and the algorithm has converged. As a result, they could be formulated as:

$$\begin{aligned}&\frac{\|\mathbf{X} - \mathbf{D}^k - \mathbf{E}^k\|_\text{F}}{\|\mathbf{X}\|_\text{F}} < 10^{-8}\\ \wedge &\frac{\|\hat{\mathbf{D}}^k - [\mathbf{D}_{tr}^k; \mathbf{1}^\top]\|_\text{F}}{\|\hat{\mathbf{D}}^k\|_\text{F}} < 10^{-8}\\ \wedge &\frac{\|\boldsymbol{\beta}^k - \mathbf{B}^k\|_\text{F}}{\|\boldsymbol{\beta}^k\|_\text{F}} < 10^{-8}.\end{aligned} \tag{S.12}$$

A.10 Obtaining the testing data labels

After the algorithm is converged and the mapping $\boldsymbol{\beta}$ is optimally calculated, the labels for the testing data would be achieved by applying $\boldsymbol{\beta}$ on the denoised testing samples: $\mathbf{Y}_{tst} = \boldsymbol{\beta}[\mathbf{D}_{tst}; \mathbf{1}^\top]$.

B Algorithm Analysis

The optimization step for each of the matrices $\mathbf{B}, \mathbf{D}, \hat{\mathbf{D}}, \mathbf{E}$ is convex, while all the other variables are fixed. For $\boldsymbol{\beta}$, the solution is achieved via the IRLS approach, in an iterative manner. Both the ℓ_1 fitting function and the approximated

re-weighted least-squares functions are convex. We need to ensure that the minimization of the latter is numerically better tractable than the minimization of the former. This is discussed in depth and the convergence is proved in [2], and adopted for our case in the following.

The IRLS framework updates the weights in an iterative manner, solving a least-squares problem in each iteration. The convergence property of Algorithm 1 would, therefore, be similar to those for IRLS [2, 6].

Theorem B.1. *The sequence $\{\boldsymbol{\beta}^t\}_{t=1}^{\infty}$ generated by Algorithm 1 converges to a stationary point, as t increases.*

Proof. Algorithm 1 solves the problem for $\boldsymbol{\beta}$, assuming all other variables are fixed, resulting in a set $\{\boldsymbol{\beta}^t\}_{t=1}^{\infty}$. Since the sub-problem for $\boldsymbol{\beta}$ (see Subsection A.2) is a convex problem, based on Theorem 1 in [6], the above sequence of $\{\boldsymbol{\beta}^t\}_{t=1}^{\infty}$, monotonically decreases the objective value of the problem (S.1):

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\beta}^{t+1},\mathbf{B},\mathbf{D},\hat{\mathbf{D}},\mathbf{E},\{\mu_i\}_{i=1}^3,\{\mathcal{Y}_i\}_{i=1}^3)\\&\leq \mathcal{L}(\boldsymbol{\beta}^{t},\mathbf{B},\mathbf{D},\hat{\mathbf{D}},\mathbf{E},\{\mu_i\}_{i=1}^3,\{\mathcal{Y}_i\}_{i=1}^3).\end{aligned}\tag{S.13}$$

Since the sequences of all $\mathcal{L}(\boldsymbol{\beta}^t,...)$ (short notation, for easier readability) is monotonic, and bounded (it will be proved in Lemmas B.2 and B.3 that all the associated variables in (S.1) are bounded, and therefore $\mathcal{L}(.)$ is also bounded), then $\mathcal{L}(\boldsymbol{\beta}^t,...)-\mathcal{L}(\boldsymbol{\beta}^{t+1},...)\rightarrow 0$ as $t\rightarrow\infty$. Then, as the solution to $\boldsymbol{\beta}^{t+1}$ in Step 5 of Algorithm 1 and its dependence to $\hat{\boldsymbol{\alpha}}^t$ (in Step 4) suggest, as a rule of thumb, we can conclude that $\boldsymbol{\beta}^{t+1}\leq\boldsymbol{\beta}^t$, and therefore $\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t\|\rightarrow 0$. The first stopping criterion in Step 7 of the Algorithm stops the iteration when a relatively stationary point is discovered. □

To study the convergence rates and properties of the whole algorithm, first, the boundedness of the associated variables (the Lagrangian multipliers and the optimization variables) should be investigated [1, 7].

Lemma B.2. *The sequences $\{\mathcal{Y}_i^k\}_{i=1}^3$ are bounded.*

Proof. Let $\{\boldsymbol{\beta}_*^k\}$, $\{\mathbf{B}_*^k\}$, $\{\mathbf{D}_*^k\}$, $\{\hat{\mathbf{D}}_*^k\}$ and $\{\mathbf{E}_*^k\}$ be the optimal values of the optimization problem (S.1), at iteration k. Then:

$$\begin{aligned}&0\in\partial_{\boldsymbol{\vartheta}}\mathcal{L}(\boldsymbol{\beta}_*^k,\mathbf{B}_*^k,\mathbf{D}_*^k,\hat{\mathbf{D}}_*^k,\mathbf{E}_*^k,\{\mu_i^k\}_{i=1}^3,\{\mathcal{Y}_i^k\}_{i=1}^3),\\&\forall\boldsymbol{\vartheta}\in\{\boldsymbol{\beta},\mathbf{B},\mathbf{D},\hat{\mathbf{D}},\mathbf{E}\},\end{aligned}\tag{S.14}$$

where $\partial_{\boldsymbol{\vartheta}}\mathcal{L}(.)$ is the sub-gradient of $\mathcal{L}(.)$ with respect to $\boldsymbol{\vartheta}$. Therefore,

$$\begin{aligned}&\{\mathcal{Y}_1^k\}\in\partial_{\mathbf{D}}\mathcal{L}\wedge\{\mathcal{Y}_1^k\}\in\partial_{\mathbf{E}}\mathcal{L},\\&\{\mathcal{Y}_2^k\}\in\partial_{\mathbf{D}}\mathcal{L}\wedge\{\mathcal{Y}_2^k\}\in\partial_{\hat{\mathbf{D}}}\mathcal{L},\\&\{\mathcal{Y}_3^k\}\in\partial_{\boldsymbol{\beta}}\mathcal{L}\wedge\{\mathcal{Y}_3^k\}\in\partial_{\mathbf{B}}\mathcal{L}.\end{aligned}\tag{S.15}$$

Consequently, since all the sub-gradients reduces to matrix norms, $\|.\|_*$, $\|.\|_1$ and $\|.\|_{\mathrm{F}}$ and their dual norms are $\|.\|_2$, $\|.\|_\infty$ and $\|.\|_{\mathrm{F}}$ respectively [3, 7, 8], based on Theorem E.1 in Section E, $\{\mathcal{Y}_i^k\}_{i=1}^3$ are bounded. □

Lemma B.3. *If $\{\mu_i^k\}_{i=1}^3$ satisfy the condition $\sum_{k=1}^{+\infty}(\mu_i^k)^{-2}\mu_i^{k+1}<+\infty,\forall i\in\{1,2,3\}$, the sequences $\{\boldsymbol{\beta}^{k*}\}$, $\{\mathbf{B}_*^k\}$, $\{\mathbf{D}_*^k\}$, $\{\hat{\mathbf{D}}_*^k\}$ and $\{\mathbf{E}_*^k\}$ are bounded.*

Proof. As far as we are solving a minimization problem, with the arguments in [3, 7] we can say:

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\beta}_*^{k+1},\mathbf{B}_*^{k+1},\mathbf{D}_*^{k+1},\hat{\mathbf{D}}_*^{k+1},\mathbf{E}_*^{k+1},\{\mu_i^{k+1}\}_{i=1}^3,\{\mathcal{Y}_i^{k+1}\}_{i=1}^3)\\&\leq\ \mathcal{L}(\boldsymbol{\beta}_*^{k+1},\mathbf{B}_*^{k+1},\mathbf{D}_*^{k+1},\hat{\mathbf{D}}_*^{k+1},\mathbf{E}_*^{k+1},\{\mu_i^{k}\}_{i=1}^3,\{\mathcal{Y}_i^{k}\}_{i=1}^3)\\&\leq\ \mathcal{L}(\boldsymbol{\beta}_*^{k},\mathbf{B}_*^{k},\mathbf{D}_*^{k},\hat{\mathbf{D}}_*^{k},\mathbf{E}_*^{k},\{\mu_i^{k}\}_{i=1}^3,\{\mathcal{Y}_i^{k}\}_{i=1}^3)\\&=\ \mathcal{L}(\boldsymbol{\beta}_*^{k},\mathbf{B}_*^{k},\mathbf{D}_*^{k},\hat{\mathbf{D}}_*^{k},\mathbf{E}_*^{k},\{\mu_i^{k-1}\}_{i=1}^3,\{\mathcal{Y}_i^{k-1}\}_{i=1}^3)\\&\quad+\frac{1}{2}\sum_{i=1}^3[\frac{\mu_i^{k-1}+\mu_i^k}{(\mu_i^{k-1})^2}\|\mathcal{Y}_i^k-\mathcal{Y}_i^{k-1}\|_{\mathrm{F}}^2].\end{aligned}\tag{S.16}$$

With the boundedness of $\{\mathcal{Y}_i^k\}_{i=1}^3$ and $\sum_{k=1}^{+\infty}(\mu_i^k)^{-2}\mu_i^{k+1}<+\infty,\forall i\in\{1,2,3\}$, we can say $\{\mathcal{L}(.)\}$ is bounded and as a result $\{\boldsymbol{\beta}^{k*}\}$, $\{\mathbf{B}_*^k\}$, $\{\mathbf{D}_*^k\}$, $\{\hat{\mathbf{D}}_*^k\}$ and $\{\mathbf{E}_*^k\}$ are bounded, as well. □

Theorem B.4. *The convergence rate to minimize the objective in (5) is of at least $O(\sup(\{\mu_i^k\}_{i=1}^3)^{-1})$, with $\sup(.)$ defined as the supremum or the least upper bound of the set.*

$$\left|F^k-F^*\right|=O(\sup(\{\mu_i^k\}_{i=1}^3)^{-1})\tag{S.17}$$

where F^k is the value of the objective in (5), and F^* is the overall optimal solution to the same problem.

Proof. Since F^* is the optimal solution, similar to [7], we have:

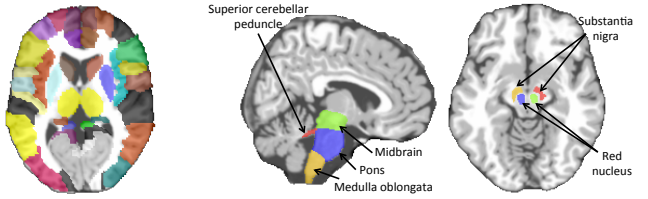
$$\begin{aligned}&\mathcal{L}(\boldsymbol{\beta}_*^k,\mathbf{B}_*^k,\mathbf{D}_*^k,\hat{\mathbf{D}}_*^k,\mathbf{E}_*^k,\{\mu_i^k\}_{i=1}^3,\{\mathcal{Y}_i^k\}_{i=1}^3)\\&=\min_{\boldsymbol{\beta},\mathbf{B},\mathbf{D},\hat{\mathbf{D}},\mathbf{E}}\mathcal{L}(\boldsymbol{\beta},\mathbf{B},\mathbf{D},\hat{\mathbf{D}},\mathbf{E},\{\mu_i^k\}_{i=1}^3,\{\mathcal{Y}_i^k\}_{i=1}^3)\\&\leq\min_{\mathbf{X}=\mathbf{D}+\mathbf{E},\hat{\mathbf{D}}=[\mathbf{D}_{tr};\mathbf{1}^\top]}\mathcal{L}(\boldsymbol{\beta},\mathbf{B},\mathbf{D},\hat{\mathbf{D}},\mathbf{E},\{\mu_i^k\}_{i=1}^3,\{\mathcal{Y}_i^k\}_{i=1}^3)\\&=F^*.\end{aligned}\tag{S.18}$$

So

$$\begin{aligned}F^k=&\ \mathcal{L}(\boldsymbol{\beta}_*^k,\mathbf{B}_*^k,\mathbf{D}_*^k,\hat{\mathbf{D}}_*^k,\mathbf{E}_*^k,\{\mu_i^k\}_{i=1}^3,\{\mathcal{Y}_i^k\}_{i=1}^3)\\=&\ \mathcal{L}(\boldsymbol{\beta}_*^k,\mathbf{B}_*^k,\mathbf{D}_*^k,\hat{\mathbf{D}}_*^k,\mathbf{E}_*^k,\{\mu_i^{k-1}\}_{i=1}^3,\{\mathcal{Y}_i^{k-1}\}_{i=1}^3)\\&-\sum_{i=1}^3[\frac{1}{2\mu_i^k}(\|\mathcal{Y}_i^k-\mathcal{Y}_i^{k-1}\|_F^2]\\\leq&\ F^*-\sum_{i=1}^3[\frac{1}{2\mu_i^k}(\|\mathcal{Y}_i^k-\mathcal{Y}_i^{k-1}\|_F^2]\\=&\ F^*-O(\sup(\{\mu_i^k\}_{i=1}^3)^{-1}).\end{aligned}\tag{S.19}$$

Due to the boundedness of all the variables, the theorem is proved. □

The above theorem shows the convergence rate relative to the convergence of each sub-problem. The convergence and the computational complexity of the whole algorithm would, therefore, be also dependent on each of the above sub-procedures. To calculate the computational complexity of the algorithm in every single iteration, the most computationally expensive steps should be investigated. The two most intensive steps are the iterative update of $\boldsymbol{\beta}$ and the SVT operation for updating $\mathbf{D}$. The former includes solving a least-squares iteratively, which is $O(m^2N)$ in each iteration, and the latter has the SVD operation as the most computationally intensive operation, which is of $O(m^2N+N^3)$. By considering the maximum number of iterations for the first sub-procedure equal to $t_{max}=100$ (as in Algorithm 1), the overall computational complexity of the algorithm in each iteration would be $O(100m^2N+N^3)$.



(a) AAL atlas with 90 cortical and subcortical ROIs (b) Four Brainstem ROIs (c) Substantia nigra and red nucleus (L & R)

Fig. S2: All 98 ROIs used for processing the PPMI data, in this study: 90 ROIs from the AAL atlas [13], 4 ROIs defined in brainstem, 2 ROIs in substantial nigra (L/R), and 2 ROIs in red nucleus (L/R).

As discussed earlier, the number of iterations until convergence is dependent on the choice of $\{\mu\}$s. From Theorem B.4, we see that if the penalty hyperparameters grow geometrically, the ADMM will converge Q-linearly.

C Neuroimaging Data Acquisition and Preprocessing

In this study, we use two databases for two different brain neurodegenerative diseases, namely PD and AD.

The first set of data is obtained from the Parkinson's Progression Markers Initiative (PPMI) database[1] [9]. PPMI is the first substantial study for identifying the PD progression biomarkers to advance the understanding of the disease. In this research, we use T1-weighted MR images acquired with MPRAGE sequence on 3T Siemens Magnetom TrioTim Syngo scanners. Imaging parameters are as below: 176 sagittal slices, repetition time = 2300 ms, echo time = 2.98 ms, flip angle = 9°, and voxel size = $1 \times 1 \times 1$ mm^3. A total of 374 PD and 169 normal control (NC) subjects are included. The demographic information of the subjects is shown in Table S1.

All the MR images were preprocessed by skull stripping [10, 11], cerebellum removal, and then segmentation into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) tissues [12]. As shown in Figure S2, we employ a brain atlas, which defines 98 ROIs that are clinically relevant regions for PD. Specifically, we first employed the Anatomical Automatic Labeling (AAL) atlas [13] (Figure S2a), which includes 90 cortical and subcortical ROIs. We further added eight more ROIs in the deep brain area (in basal ganglia and brainstem), which are clinically important regions for PD [14]. Specifically, these 8 regions are 'medulla oblongata', 'pons', 'midbrain', and 'superior cerebellar peduncle' in brainstem (Figure S2b), and 'substantia nigra', left and right, and 'red nucleus', left and right (Figure S2c). Through the corresponding intensity template, this atlas was propagated to each subject's native space using HAMMER[2] [15, 16] . We then computed WM, GM and CSF tissue volumes in all 98 regions, resulting in $98 \times 3 = 294$ features. The pipeline for processing the MR images is summarized in Figure S3.

The second dataset used in this study comes from the Alzheimer's disease neuroimaging initiative (ADNI) database[3], which includes MRI and FDG-PET data. For this experiment, we used 93 AD patients, 202 MCI patients, and 101 NC subjects, which had complete MRI and FDG-PET data. The demographic information of these subjects is included in Table S2. We used same tools as employed in the previous experiment to preprocess the data. Due to the disease differences, we use slightly different ROIs. Specifically, we obtained the labeled images based on a template with 93 manually labeled regions-of-interest (ROIs) [17]. Then, the volume of GM tissue in each ROI was calculated as the image feature. For FDG-PET images, a rigid transformation was employed to align it to the corresponding MR image, and the mean intensity of each ROI was measured as the feature, leading to a total of $93 \times 2 = 186$ features for each subject. These features are normalized, in a similar way as described in [18].

TABLE S1: Details of the subjects from the PPMI dataset used in our study. 'Age' indicates the mean±standard deviation (std) of the subject ages (in years) in that category. Similarly, 'Edu.' denotes the mean ± std of the amount of education (in years) of the subjects.

	Total	Gender		Age (years)	Edu. (years)
		F	M		
PD	374	132	242	61.50 ± 9.62	15.58 ± 2.93
NC	169	59	110	60.42 ± 11.43	16.09 ± 2.87

TABLE S2: Demographic information of the subjects from the ADNI dataset used in our study. 'Age' indicates the mean±standard deviation (std) of the subject ages (in years) in that category. Similarly, 'Edu.' denotes the mean ± std of the amount education (in years) of the subjects.

	Total	Gender		Age (years)	Edu. (years)
		F	M		
AD	93	36	57	75.38 ± 7.39	14.66 ± 3.20
MCI	202	66	136	75.06 ± 7.05	15.71 ± 2.86
NC	101	39	62	75.82 ± 4.84	15.82 ± 3.15

D Experiments on Synthetic and Benchmark Datasets

The first experiment evaluates our method on a synthetic set of data, which highlights how the proposed method is robust against sample-outliers or feature-noises separately, or when they occur at the same time. Then, we employ some benchmark semi-supervised learning datasets and report results in comparisons with some baseline and state-of-the-art methods.

D.1 Synthetic Data

To verify the method and analyze its performance, we first construct two independent subspaces with the dimensionality of 100, similar to what described in [4]. The two subspaces are constructed with bases $\mathbf{U}_1$ and $\mathbf{U}_2$, where $\mathbf{U}_1 \in \mathbb{R}^{100\times 100}$ is a random orthogonal matrix and $\mathbf{U}_2 = \mathbf{T}\mathbf{U}_1$, with $\mathbf{T}$ as a random rotation matrix. We then sample 500 vectors from each subspace: $\mathbf{X}_i = \mathbf{U}_i\mathbf{Q}_i, i = \{1, 2\}$,

1. http://www.ppmi-info.org/data
2. http://www.nitrc.org/projects/hammerwml
3. http://www.loni.ucla.edu/ADNI

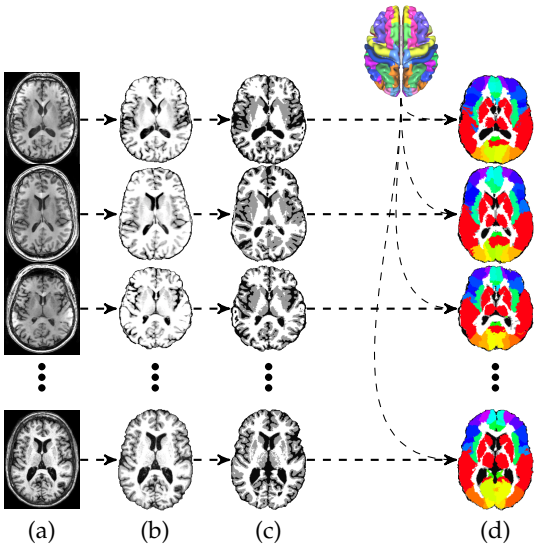


Fig. S3: MR image processing pipeline: The original MR images (a) are processed by skull stripping (b) and then tissue segmentation (c). Then, an atlas is non-linearly registered to each subject's original MR image (d), and tissue volumes of ROIs are calculated as features.

with $\mathbf{Q}_i$, a 100×500 matrix, being independent and identically distributed (*i.i.d.*) from $\mathcal{N}(0,1)$. This leads to a binary classification problem. We gradually add additional noisy samples and features to the data and evaluate the proposed method. The noise is also drawn *i.i.d* from $\mathcal{N}(0,1)$.

We generate five different sets of such synthetic data using the procedures described above and use them to evaluate our proposed method, compared to two of the most relevant baseline methods (RLDA and LS-LDA). The mean and the standard deviation of the accuracy for these evaluations are illustrated in Fig. S4. This experiment is conducted in three settings, first of which (Fig. S4a) shows the behavior of the methods against gradually added noise to the features (feature-noises). In the second scenario, we randomly add some completely random noisy samples to the aforementioned noise-free samples and evaluate the methods in the sole presence of sample-outliers. Results are depicted in Fig. S4b. Finally, we simultaneously add noise to the features and include some random noisy samples in the data. Fig. S4c shows the mean±std of accuracy as a function of the additional number of both noisy features and samples. All the reported results are obtained under 10-fold cross-validation settings. As can be seen, our method can achieve superior results compared to RLDA and conventional LS-LDA approaches. Furthermore, it behaves more robustly against the increase of the noise in data, especially when the task involves sample-outliers.

D.2 Semi-Supervised Learning Datasets

To substantiate the validity of the proposed method in solving SSL problems, we use the standard SSL settings and datasets as in Chapelle *et al*. [19]. Each of the sets of the data contains 1500 points, and either $l = 10$ or $l = 100$ of them are labeled. We compare the results with the results reported in [19], as well as those in [20]. The benchmarks

TABLE S3: The accuracy of the proposed method on SSL benchmark datasets, in comparisons to [19] and [20]. Bold numbers show relatively better results.

	Dataset	RFS-LDA	[19]	[19] (Entropy-Reg)	[20]
$l = 10$	Digit1	**85.12**	79.41	75.56	84.57
	BCI	**53.01**	49.96	52.29	52.22
	g241c	**87.05**	79.05	52.64	**87.15**
	g241d	**54.40**	53.65	54.19	**54.44**
	USPS	71.81	69.34	**79.75**	57.08
$l = 100$	Digit1	**93.10**	81.95	92.72	91.24
	BCI	73.90	57.33	71.11	**78.12**
	g241c	**87.15**	81.82	79.03	**86.02**
	g241d	**78.58**	76.24	74.64	77.11
	USPS	83.61	78.88	**87.79**	71.62

in [19] are established based on SVM formulations, and one of them uses an entropy regularization (Entropy-Reg). The method proposed in [20] enjoys from a convex relaxation of a general cost function, which is used for weakly supervised problems. Since our method is based on a linear loss function, we compare the results from the above benchmarks with linear settings.

As could be seen from the results, since the method presented in [20] has a convex formulation and is also robust to some extends of noise in the data (as claimed in the paper, due to the convex relation), we obtain roughly similar performance. But especially for the case with less labeled data ($l = 10$), our method is performing relatively better. This could be attributed to the fact that we successfully construct the intrinsic geometry of the sample manifold while learning the classifier.

D.3 Additional Results for the Disease Diagnosis Experiment

In addition to the results reported in the paper, Figure S5 shows the sensitivity, specificity and the area under the ROC curve for the proposed method compared to other baseline methods, on both PPMI and ADNI databases.

E Boundedness of the Dual Norms

Theorem E.1. *If $\|.\|^*$ is the dual norm of $\|.\|$ in a real Hilbert space $\mathcal{H}$ endowed with an inner product $\langle .,. \rangle$, and if $\mathbf{y} \in \partial\|\mathbf{x}\|$, where $\partial\|\mathbf{x}\|$ is the sub-gradient of $\|\mathbf{x}\|$, then:*

$$\|\mathbf{y}\|^* \begin{cases} = 1 & \text{if} \quad \mathbf{x} \neq \mathbf{0} \\ \leq 1 & \text{if} \quad \mathbf{x} = \mathbf{0} \end{cases} \tag{S.20}$$

Proof. Similar to the discussions in [7], since $\mathbf{y} \in \partial\|\mathbf{x}\|$, we can remark:

$$\|\mathbf{w}\| - \|\mathbf{x}\| \geq \langle \mathbf{y}, \mathbf{w} - \mathbf{x} \rangle, \forall \mathbf{w} \in \mathcal{H}. \tag{S.21}$$

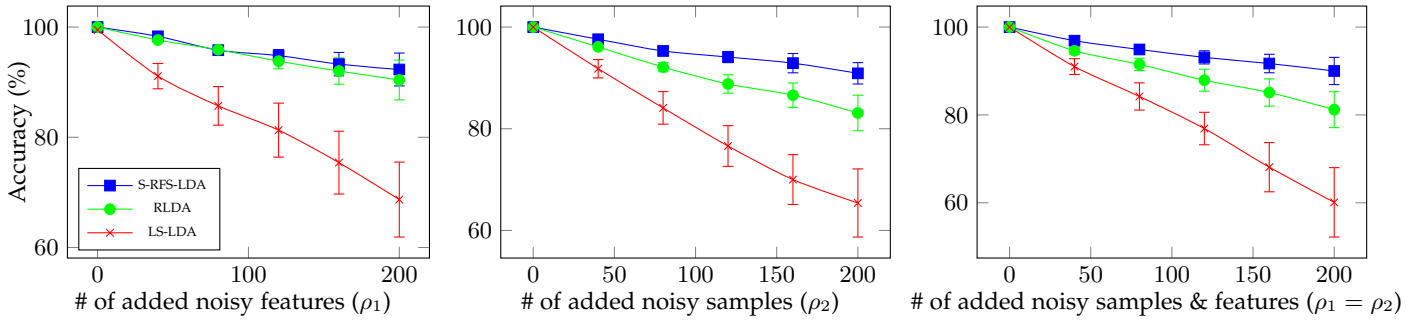
If $\mathbf{x} = \mathbf{0}$, the the above equation would be reduced to:

$$\langle \mathbf{y}, \mathbf{w} \rangle \leq 1, \forall \|\mathbf{w}\| = 1, \tag{S.22}$$

which means that $\|\mathbf{y}\|^* \leq 1$.

For the case $\mathbf{x} \neq \mathbf{0}$, we know that

$$\|\mathbf{w} - \mathbf{x}\| \geq \|\mathbf{w}\| - \|\mathbf{x}\|, \forall \mathbf{w} \in \mathcal{H}. \tag{S.23}$$



(a) Only added noisy features (b) Only added noisy samples (c) Added noisy samples & features

Fig. S4: Results comparisons on synthetic data for five different runs (mean±std).

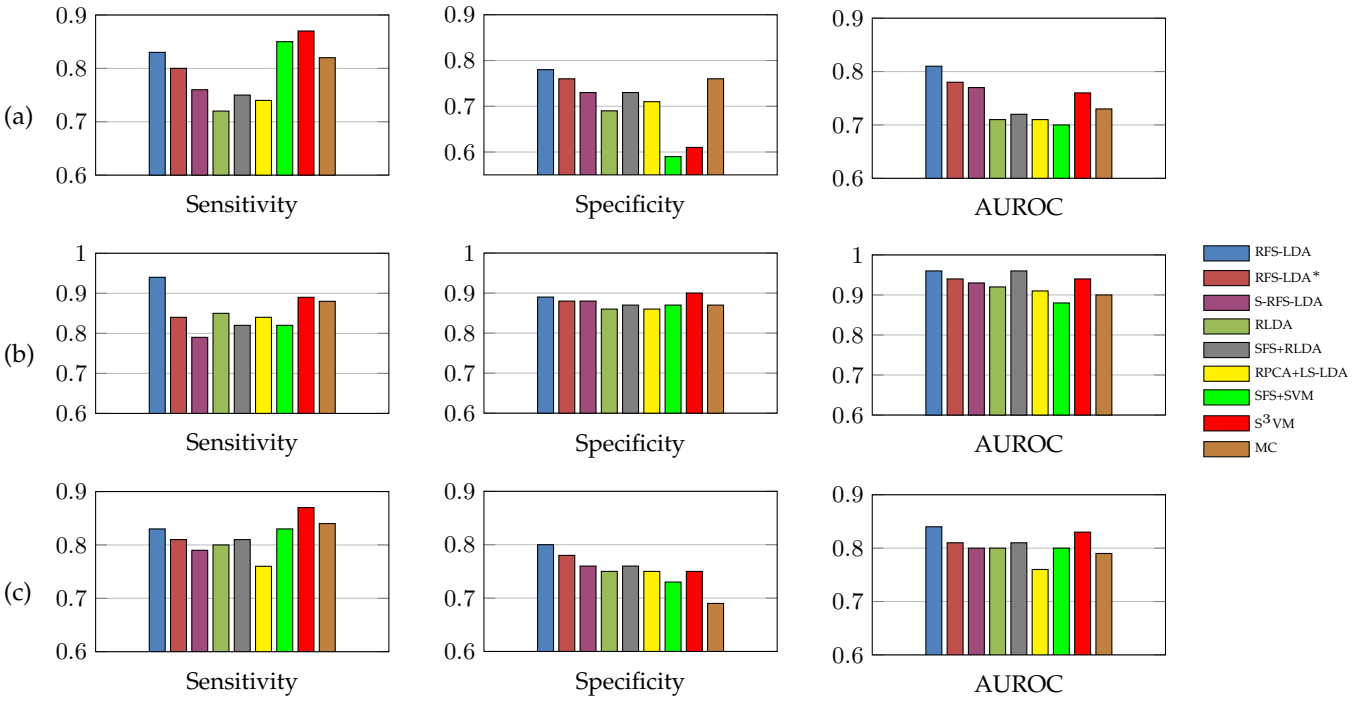


Fig. S5: Results comparisons between the proposed and the baseline methods: (a) PD *vs*. NC, (b) AD *vs*. NC, (c) MCI *vs*. NC.

By combining (S.21) and (S.23), we can conclude:

$$\langle \mathbf{y}, \frac{\mathbf{w}-\mathbf{x}}{\|\mathbf{w}-\mathbf{x}\|} \rangle \leq 1, \forall \mathbf{w} \neq \mathbf{x}, \tag{S.24}$$

which means that $\|\mathbf{y}\|^* \leq 1$. On the other hand, by setting $\mathbf{w} = \mathbf{0}$ or $2\mathbf{x}$, from (S.21) we can deduce:

$$\|\mathbf{x}\| = \langle \mathbf{y}, \mathbf{x} \rangle \leq \|\mathbf{x}\| \|\mathbf{y}\|^*. \tag{S.25}$$

This means that $\|\mathbf{y}\|^* \geq 1$. As a result, for $\mathbf{x} \neq \mathbf{0}$, we can conclude that $\|\mathbf{y}\|^* = 1$. □

REFERENCES

[1] S. Boyd and *et al*., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[2] N. Bissantz, L. Dümbgen, A. Munk, and B. Stratmann, "Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces," *SIAM J. on Optimization*, vol. 19, no. 4, pp. 1828–1845, 2009.

[3] J.-F. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[4] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE TPAMI*, vol. 35, no. 1, pp. 171–184, 2013.

[5] E. Adeli-Mosabbeb and M. Fathy, "Non-negative matrix completion for action detection," *Image and Vision Computing*, vol. 39, pp. 38 – 51, 2015.

[6] K. N. Chaudhury, "On the convergence of the irls algorithm in non-local patch regression," *Signal Processing Letters, IEEE*, vol. 20, no. 8, pp. 815–818, 2013.

[7] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices," *UIUC technical report 2215*, 2009.

[8] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, vol. 61, 2009.

[9] K. Marek and *et al*., "The parkinson progression marker initiative (PPMI)," *Progress in Neurobiology*, vol. 95, no. 4, pp. 629 – 635, 2011.

[10] F. Shi, L. Wang, Y. Dai, J. H. Gilmore, W. Lin, and D. Shen, "LABEL: Pediatric brain extraction using learning-based meta-algorithm," *NeuroImage*, vol. 62, no. 3, pp. 1975 – 1986, 2012.

[11] Y. Wang, J. Nie, P.-T. Yap, G. Li, F. Shi, X. Geng, L. Guo, D. Shen, ADNI, *et al*., "Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates," *PLOS ONE*, vol. 9, no. 1, p. e77810, 2014.

[12] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE TMI*, vol. 20, no. 1, pp. 45–57, 2001.
[13] N. Tzourio-Mazoyer and *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.
[14] H. Braak, K. Tredici, U. Rub, R. de Vos, E. J. Steur, and E. Braak, "Staging of brain pathology related to sporadic parkinson's disease," *Neurobio. of Aging*, vol. 24, no. 2, pp. 197 – 211, 2003.
[15] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE TMI*, vol. 21, pp. 1421–1439, 2002.
[16] Y. Wang, J. Nie, P.-T. Yap, F. Shi, L. Guo, and D. Shen, "Robust deformable-surface-based skull-stripping for large-scale studies.," in *MICCAI*, pp. 635–642, 2011.
[17] N. J. Kabani, D. J. Macdonald, C. J. Holmes, and A. C. Evans, "3D anatomical atlas of the human brain," in *OHBM*, 1998.
[18] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, ADNI, *et al.*, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.
[19] O. Chapelle, B. Schölkopf, and A. Zien, eds., *Semi-Supervised Learning*. MIT Press, 2006.
[20] A. Joulin and F. R. Bach, "A convex relaxation for weakly supervised classifiers," in *ICML*, pp. 1279–1286, 2012.