Joint Feature-Sample Selection and Robust Classification for Parkinson's Disease Diagnosis

Ehsan Adeli-Mosabbeb[1], Chong-Yaw Wee[1,2],
Le An[1], Feng Shi[1], and Dinggang Shen[1]

[1] Department of Radiology and BRIC,
University of North Carolina at Chapel Hill, NC, 27599, USA
[2] Department of Biomedical Engineering,
National University of Singapore, Singapore City, Singapore

Abstract. Parkinson's disease (PD) is an overwhelming neurodegenerative disorder caused by deterioration of a neurotransmitter, known as dopamine. Lack of this chemical messenger in the brain impairs several brain regions and yields to various movement and non-motor symptoms. The incidence of PD is considered to be doubled in the next two decades and this urges more researches on its early diagnosis and treatment. In this paper, we propose an approach to diagnose PD using magnetic resonance imaging (MRI) data. We first introduce a joint feature-sample selection method to select the optimal subset of samples and features for a reliable training process. This procedure selects the most discriminative features and discards poor sample (outliers). Then, a robust classification framework is proposed that can simultaneously de-noise the selected subset of features and samples, and learn a classification model. Our model can further denoise the test samples based on the cleaned training data. Experimental results on both synthetic and a publicly available PD dataset show promising results.

1 Introduction

Parkinson's disease (PD) is a neurodegenerative brain disorder characterized by the progressive impairment and deterioration of brain neurons. PD is caused when the brain gradually stops producing a vital endogenous chemical messenger known as dopamine. Dopamine is produced by the neurons that are concentrated in an area of brain recognized as substantia nigra. It is a neurotransmitter regulating the communication between the substantia nigra and the corpus striatum. This communication coordinates the balanced regular muscle movements. Lack of dopamine yields loss of ability to control body movements, along with some non-motor problems (*e.g.*, depression, anxiety, apathy/abulia, *etc.*) [1]. People with PD may lose up to 80% of dopamine before symptoms appear [2, 3]. Thus, early diagnosis and treatment are crucial to slow down the progression of PD in the initial stages.

Although there is no specific test for PD diagnosis, the diagnosis process includes analyzing a sequence of symptoms while eliminating other probable syndromes or causes of each single symptom. In practice, different imaging modalities could be incorporated. SPECT imaging is usually considered for the differential diagnosis of PD and often used for people with tremor [3, 4]. Recently, many researches exploit MRI

to analyze the changes in different regions of the brain in PD patients [1, 3]. After the production of dopamine is impaired, other parts of the brain, including cortical surfaces, are also affected, thus causing the movement and non-motor symptoms [2]. Literature studies show that these influences should be characterized by specific types of MRI data [3]. In this research, we investigate the diagnosis of PD from such effects on the brain by analyzing MR images, using machine learning techniques.

MR images can be noisy due to different factors, *e.g.*, patient movements or device limitations. Most existing works manually discard poor samples by checking the images one by one. This eventually induces undesirable bias to the learned model. Therefore, it is of great interest if we could automatically select the more reliable samples, boosting up the robustness of the method and its application in a clinical setting. On the other hand, for the purposes of diagnosis, we analyze the MR images by parcellating them into several pre-defined regions of interest (ROIs) and extracting features from each ROI. The disease might not be directly associated to all of the pre-defined regions in the brain [2]. Therefore, we also need to select the most important and relevant regions for our diagnosis procedure, like in [5–7].

Unlike many previous works, in which either feature selection [6, 7] or sample selection [8] is performed, or both are considered but in sequel [5], we perform a joint feature-sample selection. The two processes (or two sub-problems) affect one another, and performing one before the other does not guarantee the selection of the best overall subset of both features and samples. Thus, these two sub-problems are overlapping, but do not have optimal sub-structures [9]. In other words, optimal overall solution is not composed of optimal solution to each sub-problem. This motivates us to jointly search for the best subsets of features and samples and introduce a novel joint feature-sample selection (JFSS) method based on how well the training labels could be represented sparsely by different number of features and samples. After feature-sample selection, we further introduce a robust classification scheme, following the least-squares linear discriminant analysis (LS-LDA) [10] formulation and the robust regression scheme [11], specially designed to enhance robustness to noise. This is because MRI is prone to noise, due to many different factors in the imaging and processing stages.

2 Overview of the Proposed Method

The whole procedure is illustrated in Figure 1. After preprocessing the subjects' MRI scans, we extract features from their pre-defined brain ROIs, and select the best subsets of features and samples through our proposed JFSS. The joint feature-sample selection procedure is able to simultaneously discard poor samples and redundant features. Here, outlier samples and samples with non-reliable predictive power for the classification purpose are considered as poor. After JFSS, there may still exist some noise in the remaining data. This noise is usually reflected in the feature values, which are associated with the single ROIs and not the whole sample. To further clean the data, we decompose the data to a cleaned version and its noise component. This is done in conjugation with the classification process, in a supervised manner, to increase its robustness to noise. Importantly, the test data is also de-noised by representing it as a locally compact linear combination of the cleaned training data.

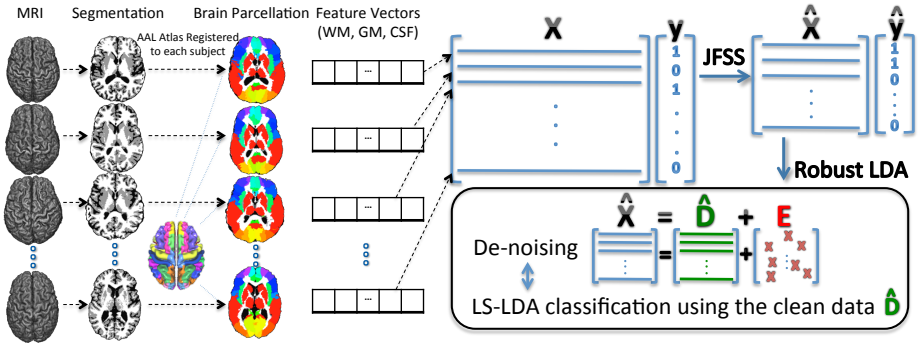


Fig. 1: Overview of our proposed method: First, the MR images are processed and tissue segmented images are obtained. Then, the anatomical automatic labeling (AAL) atlas is non-linearly registered to each subject's original space and the WM, GM and CSF volumes of each ROI are calculated as features. These features form $\mathbf{X}$, and the corresponding labels form $\mathbf{y}$. Through our proposed joint feature-sample selection (JFSS), we discard some uninformative features and samples, leading to $\hat{\mathbf{X}}$ and $\hat{\mathbf{y}}$. Then, we train a robust classifier (*i.e.*, Robust LDA) in which we jointly decompose $\hat{\mathbf{X}}$ into cleaned data $\hat{\mathbf{D}}$ and its noise component $\mathbf{E}$, and classify the cleaned data.

The key methodological contributions in our work are three-fold: (1) We propose a new joint feature-sample selection (JFSS) procedure, which jointly selects the best subset of most discriminative features and best samples to build a classification model, based on them. (2) We utilize the robust regression method [11] to develop a robust classification model and then proposed to de-noise the test data based on supervised cleaned training samples. (3) We applied our method for PD diagnosis, as PD-data driven methods are scarce, and achieved good results.

3 Data Acquisition, Processing and Notations

The data used in this paper were obtained from the Parkinson's progression markers initiative (PPMI) database[3] [12]. PPMI is the first substantial study for identifying the PD progression biomarkers to advance the understanding of the disease. In this research, we use the MRI data acquired by the PPMI study, in which a T1-weighted, 3D sequence (*e.g.*, MPRAGE or SPGR) is acquired for each subject using 3T SIEMENS MAGNETOM TrioTim syngo scanners. Note that we only use subjects who were scanned using MPRAGE sequence to minimize the effect of different scanning protocols. The T1-weighted images were acquired for 176 sagittal slices with the following parameters: repetition time (TR) = 2300 ms, echo time (TE) = 2.98 ms, flip angle = 9°, and voxel size = $1 \times 1 \times 1$ mm^3.

All the MR images were preprocessed by skull stripping [13], cerebellum removal, and tissue segmentation into white matter (WM), gray matter (GM), and cerebrospinal

[3] http://www.ppmi-info.org/data

fluid (CSF) [14]. The anatomical automatic labeling (AAL) atlas [15], parcellated with 90 predefined regions, was registered using HAMMER[4] [16, 17] to the native space of each subject. We then computed WM, GM and CSF tissue volumes in each region and used them as features, *i.e.*, obtaining 90 WM, 90 GM and 90 CSF features, for each subject. 56 PD and 56 normal control (NC) subjects are used in our experiments.

To formulate the problem, we consider N training samples, each with a $d = 270$ dimensional feature vector. Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ denote the training data, in which each row indicates a training sample, and $\mathbf{y} \in \mathbb{R}^N$ their corresponding labels. We seek to determine the labels for the test samples, $\mathbf{X}_{tst} \in \mathbb{R}^{N_{tst} \times d}$. After feature-sample selection, $\hat{N}$ samples and $\hat{d}$ features are selected, yielding to a new data matrix, $\hat{\mathbf{X}} \in \mathbb{R}^{\hat{N} \times \hat{d}}$, and train labels, $\hat{\mathbf{y}} \in \mathbb{R}^{\hat{N}}$, and same N_{tst} test samples but each with $\hat{d}$ features, $\hat{\mathbf{X}}_{tst} \in \mathbb{R}^{N_{tst} \times \hat{d}}$.[5]

4 Joint Feature-Sample Selection (JFSS)

Selecting good features and samples is critical in building reliable machine learning models. This process not only improves the generalization capability of the learned model, but also speeds up the learning process. In many applications, it is a cumbersome task to acquire the best samples and features for a learning task. Particularly in our application, feature vectors extracted from MRI data are quite prone to noise. Many researches are conducted on feature and sample selection in the recent years [5, 18–20], but few of them consider a joint formulation [21]. Feature selection and sample selection affect one another, and a separate selection might not lead to the best feature-sample subset, thus limiting the subsequent classification performance.

To this end, we aim to select only important samples and features to best describe a linear classification/regression model. In our method, we consider sparsity both in features and samples. The linear sparse regression model has been widely used for feature selection [18], where a sparse weight vector $\boldsymbol{\beta}$ is learned to best predict training labels. Accordingly, we are seeking to minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ by keeping $\boldsymbol{\beta}$ sparse at the same time. Nevertheless, doing this separate from sample selection leads to a set of features which were affected by the noisy samples already present in the data. Instead, we jointly select features and samples when constructing a linear classification/regression model, which accounts for redundant data in both domains, simultaneously. Specifically, we introduce two vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which are used to select samples and features, respectively. We impose a ℓ_1 regularization on both to ensure selection of the smallest subset in both domains. Joint feature-sample selection (JFSS) is thus performed by solving the following problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ \|\hat{\boldsymbol{\alpha}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_1, \quad \text{subject to } \hat{\boldsymbol{\alpha}} = \text{diag}(\boldsymbol{\alpha}). \tag{1}$$

The first term controls the overall data fitting error only for the selected samples indicated by $\boldsymbol{\alpha}$. The second and third terms are to ensure the selection of the smallest

[4] Could be downloaded at `http://www.nitrc.org/projects/hammerwml`

[5] Bold capital letters denote matrices (*e.g.*, $\mathbf{D}$). All non-bold letters denote scalar variables. d_{ij} denotes the scalar in the row i and column j of $\mathbf{D}$. $\|\mathbf{d}\|_2^2 = \langle \mathbf{d}, \mathbf{d} \rangle = \sum_i d_i^2$ denotes the squared Euclidean Norm of $\mathbf{d}$. $\|\mathbf{D}\|_*$ designates the nuclear norm (sum of singular values) of $\mathbf{D}$. $\|\mathbf{d}\|_1 = \sum_i |d_i|$ denotes the l_1 norm of the vector $\mathbf{d}$.

number of the most meaningful samples and features. To avoid the trivial solution for the optimization variable $\boldsymbol{\alpha}$, we further impose a condition so that at least a minimum number of the best representing samples will be selected. The iterative optimization procedure for optimizing $\boldsymbol{\alpha}$ discard the samples in each iteration. It is stopped, when the desired number of samples are selected. The number of the desired samples is determined through cross validation to avoid over-fitting. Some previous works have also used a common procedure to avoid trivial solutions [21, 22].

Solving the above problem is a cumbersome task, since the first term introduces a quadratic term. In order to solve problem (1), we break it down to two sub-problems and solve them iteratively. In each iteration, we optimize the objective by alternatively fixing one variable and optimizing for the other, until convergence. Specifically, optimizing for $\boldsymbol{\beta}$, while fixing $\boldsymbol{\alpha}$ and therefore $\hat{\boldsymbol{\alpha}}$, would reduce to:

$$\min_{\boldsymbol{\beta}} \quad \|\hat{\boldsymbol{\alpha}}\mathbf{y} - \hat{\boldsymbol{\alpha}}\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1. \tag{2}$$

Similarly, the optimization step for $\boldsymbol{\alpha}$, while fixing $\boldsymbol{\beta}$, would be:

$$\min_{\boldsymbol{\alpha}} \ \|\hat{\boldsymbol{\alpha}}\big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\big)\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1, \quad \text{subject to } \hat{\boldsymbol{\alpha}} = \text{diag}(\boldsymbol{\alpha}), \tag{3}$$

which is equivalent to solving:

$$\min_{\boldsymbol{\alpha}} \|\big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\big)^\top\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1. \tag{4}$$

Both sub-problems could be optimized using an Alternating Direction Method of Multipliers (ADMM) [23].

It is noteworthy that the above procedure could also be used for classification. The first term learns a linear regression model with weights $\boldsymbol{\beta}$ to reconstruct $\mathbf{y}$ from $\mathbf{X}$, which could be used as a classification tool by discretizing the $\mathbf{y}$ values into classes. We add a single column of 1s to $\mathbf{X}$ (*i.e.*, $\mathbf{X} = [\mathbf{X}\ \mathbf{1}]$) to create a linear classifier model with a bias term. We use this classification scheme as a baseline method in the experiments referred to as Sparse Regression (SR).

5 Robust Classification (Robust LDA)

Even with selection of the most discriminative features and best samples, there might still be some noise present in the data. These noise elements of data can adversely influence the classifier learning process. Therefore, we further model the noise in the features matrix, $\mathbf{X}$. In other words, after discarding some samples and features using our JFSS, we account for the intra-sample outliers (noises in feature values) in $\hat{\mathbf{X}}$ to further reduce the influences of noise elements in the data. For this purpose, following [24, 25], we assume that the data matrix $\hat{\mathbf{X}}$ could be spanned on a low-rank subspace and therefore should be rank-deficient. This assumption supports the fact that samples from same classes should be more correlated [11, 25]. In order to achieve a robust classifier, we use a same idea as in [11], which was proposed for robust regression. In our case, classification is posed as a binary regression problem, in which a transform $\mathbf{w}$

maps each sample in $\hat{\mathbf{X}}$ to a binary label in $\mathbf{y}$. In the linear case, this could be modeled with a Linear Discriminant Analysis (LDA), which learns a linear mapping to minimize the intra-class discrimination and maximize the inter-class variation. Furthermore, LS-LDA [10] models the LDA problem in a least-squares formulation: $\min_{\mathbf{w}} \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\mathbf{w}\|_2^2$, where $\mathbf{w}$ is a projection of $\hat{\mathbf{X}}$ to the space of labels, $\hat{\mathbf{y}}$.

Assume $\hat{\mathbf{X}}$ is the data matrix corrupted by noise. Therefore, we can say $\hat{\mathbf{X}} = \hat{\mathbf{D}} + \mathbf{E}$, where $\hat{\mathbf{D}} \in \mathbb{R}^{\hat{N} \times \hat{d}}$ is the underlying noise-free component and $\mathbf{E} \in \mathbb{R}^{\hat{N} \times \hat{d}}$ is the noise component. To model this noise in the above formulation and learn the mapping $\mathbf{w}$ from the clean data $\hat{\mathbf{D}}$, we utilize the scenario in [11] and rewrite our problem as:

$$\min_{\mathbf{w},\mathbf{D},\mathbf{E}} \frac{\eta}{2}\|\hat{\mathbf{y}} - \mathbf{D}\mathbf{w}\|_2^2 + \|\hat{\mathbf{D}}\|_* + \gamma\|\mathbf{E}\|_1, \quad \text{subject to } \hat{\mathbf{X}} = \hat{\mathbf{D}} + \mathbf{E}, \mathbf{D} = [\hat{\mathbf{D}}\ \mathbf{1}], \quad (5)$$

where the first term learns the mapping $\mathbf{w}$ from the clean data and projects the samples to the label space. The second and the third terms guarantee the rank-deficiency of the data matrix $\hat{\mathbf{D}}$ and also $\mathbf{E}$ to be sparse, respectively. These two terms are similar to the Robust Principle Component Analysis (RPCA) [26]. RPCA is an unsupervised method, while the above formulation cleans the data in a supervised manner. Particularly, the matrix $\hat{\mathbf{D}}$ retains the subspace of $\hat{\mathbf{X}}$, which is most correlated to the labels $\hat{\mathbf{y}}$. The solution to problem (5) could be achieved by writing the Lagrangian function and iteratively solving for $\mathbf{w}$, $\mathbf{D}$, $\hat{\mathbf{D}}$ and $\mathbf{E}$ one at a time, while fixing others [11].

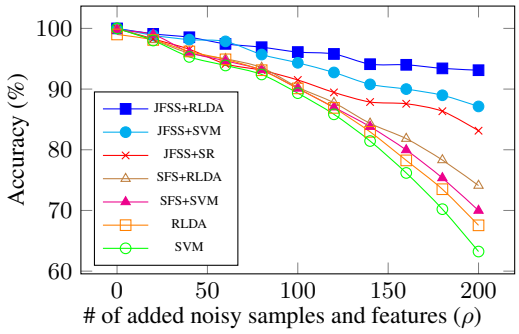
In most cases, due to the presence of noise in the test data, the classification accuracy may drop dramatically in the testing phase. For cleaning the test data, one can use RPCA [26], but again it is unsupervised. To this end, we utilize our data cleaned in , $\hat{\mathbf{D}}$, . We de-noise the test data, $\hat{\mathbf{X}}_{tst}$, by representing them as a combination of the training data: $\hat{\mathbf{D}}_{tst} = \hat{\mathbf{D}}\mathbf{Z}_{tst}$, where $\mathbf{Z}_{tst}$ is the coefficient matrix for the combination. To clean the data, we model the combination as: $\hat{\mathbf{X}}_{tst} = \hat{\mathbf{D}}\mathbf{Z}_{tst} + \mathbf{E}_{tst}$, where $\mathbf{E}_{tst} \in \mathbb{R}^{N_{tst} \times \hat{d}}$ is the noise component of the test data. In order for the linear combination to be locally compact, we further impose the low-rank constraint on the coefficients, as in [24, 27]:

$$\min_{\mathbf{Z}_{tst},\hat{\mathbf{E}}_{tst}} \|\mathbf{Z}_{tst}\|_* + \lambda\|\mathbf{E}_{tst}\|_1, \quad \text{subject to } \hat{\mathbf{X}}_{tst} = \hat{\mathbf{D}}\mathbf{Z}_{tst} + \mathbf{E}_{tst}. \quad (6)$$

This optimization problem could be solved using linearized ALM method as in [27]. After cleaning the test data, the prediction for the classification output is calculated as $\mathbf{y}_{tst} = [\hat{\mathbf{D}}\mathbf{Z}_{tst}\ \mathbf{1}]\mathbf{w}$. Same as in LS-LDA, $\mathbf{y}_{tst}$ is used as the decision value and the binary class labels are produced using the k-nearest neighbor method.

6 Experiments

In order to evaluate the proposed approach, we set up experiments on both synthetic and PPMI datasets. Baseline classifiers under comparison include linear support vector machines (SVM), sparse regression (SR), as described in Section 4, and the original LS-LDA [10]. To evaluate the JFSS procedure, we compare the results with separate feature and sample selections (FSS), sparse feature selection (SFS) and no feature sample selection (no FSS). Furthermore, we report results using other prominent methods for feature transforms like min-redundancy max-relevance (mRMR) [20] and principle



(a) Mean accuracy as a function of ρ

Method	ρ	mean $\pm$ std
JFSS+RLDA	100	96.1 $\pm$ 1.0
JFSS+SVM	100	94.3 $\pm$ 1.5
JFSS+SR	100	91.5 $\pm$ 2.4
SFS+RLDA	100	90.3 $\pm$ 2.3
SFS+SVM	100	90.1 $\pm$ 1.4
RLDA	100	90.1 $\pm$ 2.6
SVM	100	89.3 $\pm$ 2.1
JFSS+RLDA	200	93.1 $\pm$ 1.9
JFSS+SVM	200	87.2 $\pm$ 2.7
JFSS+SR	200	83.1 $\pm$ 4.1
SFS+RLDA	200	74.1 $\pm$ 3.9
SFS+SVM	200	70.0 $\pm$ 2.1
RLDA	200	67.6 $\pm$ 4.4
SVM	200	63.2 $\pm$ 4.3

(b) Mean and standard deviation

Fig. 2: Results comparisons on synthetic data, for three different runs.

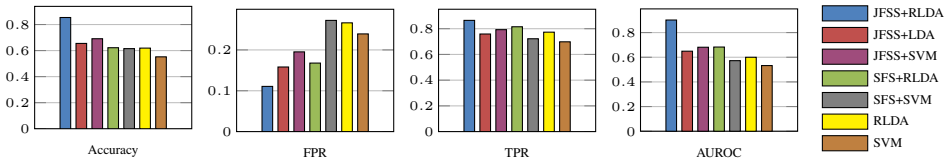


Fig. 3: Comparisons of results by the proposed (JFSS+RLDA) and the baseline methods.

component analysis (PCA). The results for the baseline methods were generated using 10-fold cross validation and best parameters for each of them were selected (like ours).

Synthetic Data: We construct two independent subspaces of dimensionality 100 (as described in [24]), and sample 500 samples from each subspace. This leads to a binary classification problem. We gradually add additional noisy samples and features to the data and evaluate the proposed JFSS and robust classification schemes using this data. Figure 2 shows the mean accuracy results of three different runs as a function of the additional number of noisy features and samples, with a 10-fold cross-validation strategy. Our JFSS coupled with the classifiers is able to select better subset of features and samples for improved results. Furthermore, it acts more robustly against the increase of noise elements. This is attributed to the de-noising process introduced by our RLDA.

PPMI Data: The details of the data are described in Section 3. For quantitative evaluations, the accuracy, true positive rate (TPR), false positive rate (FPR) and area under ROC are calculated and are shown in Figure 3, in comparisons with the baseline methods. The combination of JFSS and RLDA is outperforming all other methods by a significant margin. Table 1 also shows the diagnosis accuracy of the proposed technique (RLDA+JFSS) in comparisons with different feature/sample selection/transform methods, using a 10-fold cross-validation strategy. The optimization parameters are all chosen by a grid search for best performance using the same cross-validation strategy on training data. The proposed method outperforms all others. This could be because the data has many redundant samples and features and also suffers from noise corruption, and we deal with both issues properly.

Table 1: Accuracy results of the PD/NC classification, compared to the baseline classifiers and different feature-sample selection/transform techniques.

Classifier	**Selection/Transform Method**					
	JFSS	FSS	SFS	no FSS	mRMR	PCA
Robust LDA	**82.5**	78.1	72.5	61.9	70.8	65.5
LDA	65.5	62.5	62.2	56.6	62.1	57.0
SVM	69.1	62.4	61.5	55.2	59.8	58.0
SR	65.1	61.8	59.3	–	–	–

Discussions: Due to the huge amount of noise in the original data and redundant features and samples, we introduced JFSS framework to select the best subsets of both sample and feature spaces. Even with JFSS, portions of noise elements still exists in the data, which we de-noise using RLDA. On the other hand, RLDA alone does not provide good performance, since the amount of noisy and irrelevant feature values are high in the data. RLDA and in general most robust machine learning methods can deal with a controlled amount of noise, since they assume a sparse noise element in the data (with ℓ_1 regularization). Using JFSS, we discard a huge amount of redundant data, and RLDA is then utilized to de-noise the remaining, while classifying the data.

Furthermore, as confirmed by the results, we could distinguish PD from NC using only MRI. With the progression of PD, patients' brains are affected heavily through time. So, these data-driven methods could be of great use for early diagnosis, or prediction of the disease progression.

7 Conclusions

In this paper, we have introduced a joint feature-sample selection (JFSS) framework along with a robust classification approach for PD diagnosis. We have established robustness in both training and testing phases. We verified our method using subjects excerpted from the PPMI dataset, a first large-scale longitudinal study of PD. Our method outperforms several baseline methods on both synthetic data and the PD/NC classification problem. As a direction for future works, one can use clinical scores and other imaging modalities to predict the PD progress in subjects, or to improve the accuracy. More effective features can also be extracted to further improve the diagnosis accuracy.

References

[1] Ziegler, D.A., Augustinack, J.C.: Harnessing advances in structural MRI to enhance research on Parkinson's disease. Imaging in medicine **5**(2) (2013) 91–94

[2] Braak, H., Tredici, K., Rub, U., de Vos, R., Steur, E.J., Braak, E.: Staging of brain pathology related to sporadic parkinsons disease. Neurobio. of Aging **24**(2) (2003) 197 – 211

[3] Duchesne, S., Rolland, Y., Varin, M.: Automated computer differential classification in parkinsonian syndromes via pattern analysis on MRI. A. Radiology **16**(1) (2009) 61 – 70

[4] Prashanth, R., Roy, S.D., Mandal, P.K., Ghosh, S.: Automatic classification and prediction models for early parkinson's disease diagnosis from SPECT imaging. Expert Syst. Appl. **41**(7) (2014) 3333 – 3342

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

[6] Bron, E., Smits, M., van Swieten, J., Niessen, W., Klein, S.: Feature selection based on svm significance maps for classification of dementia. In: Machine Learning in Medical Imaging. Volume 8679. (2014) 272–279

[7] Oh, J.H., Kim, Y.B., Gurnani, P., Rosenblatt, K., Gao, J.: Biomarker selection for predicting alzheimer disease using high-resolution maldi-tof data. In: IEEE International Conference on Bioinformatics and Bioengineering. (Oct 2007) 464–471

[8] Rohlfing, T., Brandt, R., Menzel, R., Jr., C.R.M.: Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage **21**(4) (2004) 1428 – 1442

[9] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, Third Edition. 3rd edn. The MIT Press (2009)

[10] De la Torre, F.: A least-squares framework for component analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence **34**(6) (2012) 1041–1055

[11] Huang, D., Cabral, R., De la Torre, F.: Robust regression. In: European Conference on Computer Vision. (2012) 616–630

[12] Marek, K., *et al.*: The parkinson progression marker initiative (PPMI). Progress in Neurobiology **95**(4) (2011) 629 – 635

[13] Wang, Y., Nie, J., Yap, P.T., Li, G., Shi, F., Geng, X., Guo, L., Shen, D., for the Alzheimer's Disease Neuroimaging Initiative: Knowledge-guided robust mri brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PLOS ONE **9**(1) (2014) e77810

[14] Lim, K., Pfefferbaum, A.: Segmentation of MR brain images into cerebrospinal fluid spaces, white and gray matter. J. of Computer Assisted Tomography **13** (1989) 588–593

[15] Tzourio-Mazoyer, N., *et al.*: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage **15**(1) (2002) 273 – 289

[16] Shen, D., Davatzikos, C.: HAMMER: Hierarchical attribute matching mechanism for elastic registration. IEEE Trans. on Medical Imaging **21** (2002) 1421–1439

[17] Wang, Y., Nie, J., Yap, P.T., Shi, F., Guo, L., Shen, D.: Robust deformable-surface-based skull-stripping for large-scale studies. In: Medical Image Computing and Computer Assisted Intervention. Volume 6893. (2011) 635–642

[18] Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: Neural Information Processing Systems. (2010) 1813–1821

[19] Coates, A., Lee, H., Ng, A.: An analysis of single-layer networks in unsupervised feature learning. In: AI and Stat., JMLR. Volume 15. (2011) 215–223

[20] Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence **27**(8) (2005) 1226–1238

[21] Mohsenzadeh, Y., *et al.*: The relevance sample-feature machine: A sparse bayesian learning approach to joint feature-sample selection. IEEE T Cybernetics **43**(6) (2013) 2241–2254

[22] Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: Neural Information Processing Systems. (2007) 41–48

[23] Boyd, S., *et al.*: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1) (2011) 1–122

[24] Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. on Pattern Analysis and Machine Intelligence **35**(1) (2013) 171–184

[25] Goldberg, A.B., Zhu, X., Recht, B., Xu, J.M., Nowak, R.D.: Transduction with matrix completion: Three birds with one stone. In: Neural Information Processing Systems. (2010) 757–765
[26] Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Journal of the ACM **58**(3) (2011) 11:1–11:37
[27] Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low-rank representation. In: Neural Information Processing Systems. (2011) 612–620