



# Low-rank dimensionality reduction for multi-modality neurodegenerative disease identification

Xiaofeng Zhu<sup>1,2,3</sup> · Heung-Il Suk<sup>4</sup> · Dinggang Shen<sup>3,4</sup>

Received: 9 March 2018 / Revised: 22 October 2018 / Accepted: 31 October 2018 /  
Published online: 13 November 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

## Abstract

In this paper, we propose a novel dimensionality reduction method of taking the advantages of the variability, sparsity, and low-rankness of neuroimaging data for Alzheimer’s Disease (AD) classification. We first take the variability of neuroimaging data into account by partitioning them into sub-classes by means of clustering, which thus captures the underlying multi-peak distributional characteristics in neuroimaging data. We then iteratively conduct Low-Rank Dimensionality Reduction (LRDR) and orthogonal rotation in a sparse linear regression framework, in order to find the low-dimensional structure of high-dimensional data. Experimental results on the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset showed that our proposed model helped enhance the performances of AD classification, outperforming the state-of-the-art methods.

**Keywords** Alzheimer’s Disease (AD) · Feature selection · Subspace learning

## 1 Introduction

Alzheimer’s Disease (AD) is an irreversible and progressive brain disorder and typically begins with mild memory loss and later may seriously impair an individual’s ability of daily activities [12, 25, 26, 30, 44]. Due to limited knowledge about the cause of AD, there is no cure for AD and most drugs only help slow down the advance of AD.

---

This article belongs to the Topical Collection: *Special Issue on Deep vs. Shallow: Learning for Emerging Web-scale Data Computing and Applications*  
Guest Editors: Jingkuan Song, Shuqiang Jiang, Elisa Ricci, and Zi Huang

✉ Dinggang Shen  
dgshen@med.unc.edu

<sup>1</sup> Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, Guangxi, People’s Republic of China

<sup>2</sup> Institute of Natural and Mathematical Sciences, Massey University, Auckland 0745, New Zealand

<sup>3</sup> Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

<sup>4</sup> Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

As neuroimaging data provide in-vivo information to complement clinical evaluations, computer-aided neuroimaging analysis becomes a powerful tool to help complement the conventional clinical-based assessments and cognitive measurement for AD diagnosis [13]. Computer-aided neuroimaging analysis focuses on conducting AD study with multi-modality neuroimaging data, such as functional imaging data (e.g., Positron Emission Tomography (PET)) and structural imaging data (e.g., Magnetic Resonance Imaging (MRI)), because different types of imaging information could be complementary to investigate a human brain. In computer-aided neuroimaging analysis, neuroimaging data may be extracted with a number of features for characterizing the variability of data. However, both multi-modality data and high-dimensional features easily result in the issue of ‘curse-of-dimensionality’ [32–34, 47, 48].

To address this issue, many machine learning techniques (such as feature selection [3, 14, 32, 38, 50] and low-rank regression [5, 36, 37, 49]) have been designed to reduce the feature dimension. Feature selection methods, such as statistical  $t$ -test and sparse linear regression [42, 45], find informative feature subsets from original feature set to output interpretable results [7, 41], and thus is preferable in neuroimaging studies [23, 43]. For example, Liu et al. [15] and Zhu et al. [43] showed that feature selection improves the classification accuracy of AD diagnosis, and Chu et al. [2] demonstrated that feature selection does improve the classification accuracy, but depends on the method used. Besides, feature selection has been widely applied in other domains, such as gait analysis [31] and Parkinson’s disease diagnosis [22]. Low-rank regression considers the relations among the response variables to conduct subspace learning under the assumption that the rank of the coefficient matrix in regression is no large than each of its dimension [17], and is very popular in machine learning and statistics but used less in AD study. Furthermore, subspace learning methods have recently presented promising performances in various applications [23]. Hence, it is interesting to integrate them in a unified framework, in which we can complement the limitations of each of them.

In this work, we consider the variability, sparsity, and low-rankness of data into a unified framework by utilizing various relations inherent in data, to seek possible solutions to the problems of multi-modality and high-dimensionality of neuroimaging data, for improving the interpretability and predictability of multi-modality AD classification. We first employ existing clustering methods to divide each class (*i.e.*, each clinical status in this work) into sub-classes to form a multi-output representation of response variables, for reflecting the variability of neuroimaging data. Then we iteratively conduct a Low-Rank Dimensionality Reduction (LRDR) step and an orthogonal rotation step to seek the low-dimensional structure in data. In the LRDR step, low-rank constraints conduct subspace learning to extract low-dimensional latent factors, while a sparsity constraint via an  $\ell_{2,1}$ -norm regularizer allows to select class-discriminative features, which thus has the effect of feature selection. The orthogonal rotation step considers the relations among modalities such that the multi-modality data, which share the same response variables, are kept consistent in different modalities. Moreover, these two iteration steps adjust each of steps to output representative features in feature selection for improving the interpretability and predictability of AD diagnosis.

Different from the previous study for early AD diagnosis, the contributions of this paper have the following three folds.

First, this paper takes the advantages of the variability, sparsity, and low-rankness of data for AD classification. Second, in this paper, the column-wise low-rank constraints on coefficient matrices extract low-dimensional latent factors from all features for explaining the response variables, while the row-wise sparsity constraints on coefficient matrices impose to select class-discriminative features. This enables us to simultaneously conduct subspace

**Table 1** Demographic and clinical information of the subjects

	NC (N = 101)	AD (N = 93)	MCI (N = 202)	sMCI (N = 55)	pMCI (N = 63)
Male/Female	62/39	57/36	136/66	35/20	45/18
Age (Mean ± SD)	75.8±4.8 [62.0,86.2]	75.4±7.4 [55.1,87.7]	75.1±7.1 [55.1,88.8]	75.0±6.7 [57.7,88.3]	76.4±6.7 [61.4,86.7]
Education (Mean ± SD)	15.8±3.2 [7,20]	14.7±3.2 [4,20]	15.7±2.9 [7,20]	16.0±2.6 [12,20]	15.5±3.0 [7,20]
MMSE (Mean ± SD)	28.9±1.1 [25,30]	23.4±2.1 [18,27]	27.2±1.7 [24,30]	26.9±1.7 [24,30]	27.6±1.6 [24,30]
CDR (Mean ± SD)	0±0 [0,0]	0.8±0.2 [0.5,1]	0.50±0 [0.5,0.5]	0.50±0 [0.5,0.5]	0.50±0 [0.5,0.5]

learning and feature selection, and thus making up for the limitations of each dimensionality reduction method. Last but not least, unlike the existing multi-task feature selection methods [29, 40] that select the representative features across tasks, our method considers the complementary information among modalities to separately select representative features of each modality (*i.e.*, MRI and PET), since studies have manifested that MRI features have different informative brain regions with PET features for AD diagnosis [27]. Moreover, our method also considers the relations among modalities in the orthogonal rotation step to lead to a nonlinear dimensionality reduction framework. Such considerations among modalities are naturally designed to exploit the complex system of brain regions.

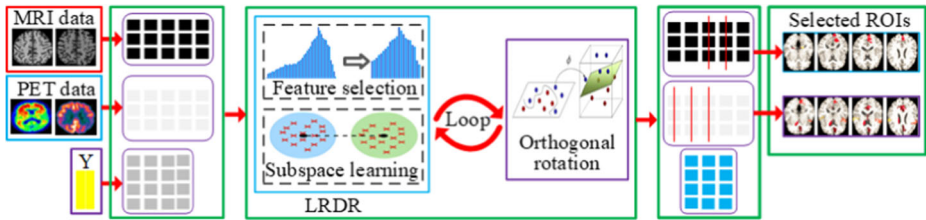
## 2 Materials and image preprocessing

We downloaded all neuroimaging data from the ADNI website,<sup>1</sup> where the used structural MR images were obtained from 1.5T scanners and the used PET images used were obtained 30–60 minutes post Fluoro-DeoxyGlucose (FDG) injection. Moreover, these MR images have been finished with the following preprocessing, including quality and automatically corrected, while PET images have been finished with the processes including averaging, spatially aligning, interpolating, intensity normalizing, and smoothing.

In this paper, we use 396 subjects of baseline MRI and 18F-FluoroDeoxyGlucose (18F-FDG) PET, including 93 AD, 202 MCI, and 101 NC subjects. We further partitioned MCI subjects into progressive MCI (pMCI), stable MCI (sMCI), and others. More specifically, 55 pMCI subjects converted from MCI to AD in 24 months, while 63 sMCI subjects did not convert to AD in both 24 months and 36 months<sup>2</sup> We summarize the demographics of the subjects in Table 1.

<sup>1</sup><http://www.loni.usc.edu/ADNI>.

<sup>2</sup>Besides, the remaining 84 MCI subjects include 33 subjects that did not convert in 24 months but converted in 36 months, and 51 subjects that were MCI at base line but were missed at any available time points among 0–96 months.



**Figure 1** A framework of our method. LRDR represents the low-rank dimensionality reduction step

## 2.1 Image analysis

We followed the literature [40] to conduct image processing for all MR images and PET images via the following steps:

- using MIPAV software<sup>3</sup> to conduct anterior commissure-posterior commissure correction.
- correcting intensity inhomogeneity.
- extracting the brain region on each structural MR image via the robust skull-stripping method.
- conducting manual edition (if needed) and intensity inhomogeneity correction.
- removing cerebellum based on registration.
- correcting intensity inhomogeneity by repeating N3 for three times.
- using FAST algorithm [35] in the FSL package to segment each structural MR image into three different tissues: Gray Matter (GM), White Matter (WM), and CerebroSpinal Fluid (CSF).
- using HAMMER [24] to conduct registration to obtain the Region-Of-Interests (ROIs) based on the Jacob template [11], which dissects a brain into 93 ROIs.

We computed the GM tissue volume in the ROI region by integrating the GM segmentation result of this subject. We further used affine registration to align the PET image to its responding MR T1 image, and then computed the average intensity of each ROI. As a result, we have 93 features for each MRI and 93 features for each PET.

## 3 Proposed method

In our framework, we first extract neuroimaging features from MRI images and PET images, and then select informative features by the proposed Orthogonal Low-Rank Dimensionality Reduction (OLRDR), that iteratively conducts a low-rank dimensionality reduction step and an orthogonal rotation step. The selected features are then fed into a Support Vector Machine (SVM), by which we finally identify clinical labels of testing data. Figure 1 shows the schematic diagram of our method.

<sup>3</sup><http://mipav.cit.nih.gov/clickwrap.php>.

### 3.1 Notations

In this paper, we denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters, respectively. For a matrix  $\mathbf{X} = [x_{ij}]$ , its  $i$ -th row and  $j$ -th column are denoted as  $\mathbf{x}^i$  and  $\mathbf{x}_j$ , respectively. Also, we denote the Frobenius norm and the  $\ell_{2,1}$ -norm of a matrix  $\mathbf{X}$  as  $\|\mathbf{X}\|_F = \sqrt{\sum_i \|\mathbf{x}^i\|_2^2} = \sqrt{\sum_j \|\mathbf{x}_j\|_2^2}$  and  $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}^i\|_2 = \sum_i \sqrt{\sum_j x_{ij}^2}$ , respectively. We further denote the transpose operator, the trace operator, the rank, and the inverse, of a matrix  $\mathbf{X}$ , as  $\mathbf{X}^T$ ,  $tr(\mathbf{X})$ ,  $rank(\mathbf{X})$ , and  $\mathbf{X}^{-1}$ , respectively.

### 3.2 Low-rank regression

Let  $\mathbf{X} = [\mathbf{x}^1; \dots; \mathbf{x}^n] = [\mathbf{x}_1, \dots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$  and  $\mathbf{Y} = [\mathbf{y}^1; \dots; \mathbf{y}^n] \in \{0, 1\}^{n \times c}$  be, respectively, a feature matrix and the corresponding class label matrix (as known as a response matrix), where  $n$ ,  $d$ , and  $c$  denote the numbers of samples (or subjects), feature variables, and response variables, respectively. As for the class label of the  $i$ -th sample  $\mathbf{x}^i$ , we use an indicator vector  $\mathbf{y}^i = [y_{i1}, \dots, y_{ij}, \dots, y_{ic}] \in \{0, 1\}^c$  such that when  $\mathbf{x}^i$  belongs to the  $j$ -th class, the corresponding  $j$ -th element in  $\mathbf{y}^i$  is set to one, *i.e.*,  $y_{ij} = 1$ , and all the other elements are set to 0.

A linear regression [49] finding the relationship between the response variables in  $\mathbf{Y}$  and the feature variables in  $\mathbf{X}$  is formulated as follows:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{e}\mathbf{b} \quad (1)$$

where  $\mathbf{W} \in \mathbb{R}^{d \times c}$  is a regression coefficient matrix,  $\mathbf{b} \in \mathbb{R}^{1 \times c}$  is a bias term, and  $\mathbf{e} \in \mathbb{R}^{n \times 1}$  denotes a column vector with all ones. Then the solution of  $\mathbf{W}$  for a loss function defined by least square error can be obtained by Ordinary Least Square (OLS) estimation [8, 38] as follows:

$$\hat{\mathbf{W}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{e}\mathbf{b}). \quad (2)$$

Note that for multiple response variables, *i.e.*,  $c > 1$ , the  $k$ -th column coefficients of  $\mathbf{W}$  are just the least square estimation in the regression of  $\mathbf{y}_k$  on  $\mathbf{x}_1, \dots, \mathbf{x}_d$ , thus (2) is equivalent to conduct the OLS estimation for each response variable separately. In other words, it does not make use of the possible relations among the response variables, thus limiting its modeling power.

One potential approach to circumvent such limitation is to explicitly take into account the possible relations among response variables by imposing a constraint on the rank of  $\mathbf{W}$ , *i.e.*,  $rank(\mathbf{W}) \leq \min(d, c)$ , as described in [28]. The motivation using the low-rank constraint is that, in real applications, noises or outliers often increase the rank of a feature matrix, or the features of  $\mathbf{X}$  are not linearly independent (*i.e.*,  $\mathbf{X}$  is not of full rank) [16].

Certainly,  $rank(\mathbf{W}) = r$ , where  $r \leq \min(d, c)$ , implies  $rank(\mathbf{X}\mathbf{W}) \leq r$  and the inverse is not necessarily true. With the constraint  $rank(\mathbf{X}\mathbf{W}) \leq r$ , we know that the rank of the predicted matrix  $\hat{\mathbf{Y}}$  (*i.e.*,  $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}$ ) of the response variables  $\mathbf{Y}$  is also less than  $r$ . We can interpret this as follows: there are less than  $r$  columns of  $\mathbf{Y}$  (corresponding to less than  $r$  response variables) such that each of  $c$  columns of  $\mathbf{Y}$  is actually a linear combination of those  $r$  columns, *i.e.*, there are only  $r$  effectively independent response variables. Such a low-rank constraint obviously considers the relations among response variables.

The low-rank constraint on  $\mathbf{W}$  also implies that the coefficient matrix can be expressed as the product of two lower rank matrices [14, 37], *i.e.*,

$$\mathbf{W} = \mathbf{B}\mathbf{A}^T \tag{3}$$

where  $\mathbf{B} \in \mathbb{R}^{d \times r}$  and  $\mathbf{A} \in \mathbb{R}^{c \times r}$ . For a fixed  $r$ , by replacing  $\mathbf{W}$  with  $\mathbf{B}\mathbf{A}^T$  in (1), a low-rank regression can be formulated as:

$$\min_{\mathbf{A}, \mathbf{B}, \mathbf{b}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T - \mathbf{e}\mathbf{b}\|_F^2 \tag{4}$$

Note that  $\mathbf{X}\mathbf{B}$  transforms the feature matrix into an  $r$ -dimensional space, spanned by  $r$  latent factors, and when  $r$  is smaller than  $d$ , it has the effect of reducing dimensionality. That is,  $r$  latent factors estimated from  $d$  features of  $\mathbf{X}$  are sufficient to explain the response variables. In this regard, the low-rank constraint on the coefficient matrix can be considered as conducting subspace learning on  $\mathbf{X}$ . Furthermore, the low-rank constraint also seeks the interaction between  $\mathbf{Y}$  and  $\mathbf{X}$  because it enables to use  $r$  latent response variables to linearly explain the original  $c$  response variables while the value of  $r$  depends on the rank of  $\mathbf{X}\mathbf{W}$ . Hence, the low-rank constraint could improve prediction accuracy by conducting subspace learning on  $\mathbf{X}$ , considering the relations in the columns of  $\mathbf{Y}$ , and seeking the interaction between  $\mathbf{Y}$  and  $\mathbf{X}$ .

### 3.3 OLRDR on multi-modality data

Multi-modality data (*e.g.*, MRI and PET) have been shown to provide complementary information to each other, thus helping enhance the performance of AD diagnosis [43, 46, 47]. Denoting the feature matrices of MRI and PET as  $\mathbf{X}_1 \in \mathbb{R}^{n \times d}$  and  $\mathbf{X}_2 \in \mathbb{R}^{n \times d}$ , respectively, we define the objective function of low-rank multi-modality regression<sup>4</sup> as follows:

$$\begin{aligned} \min_{\mathbf{A}_1, \mathbf{A}_2, \mathbf{B}_1, \mathbf{B}_2, \mathbf{b}_1, \mathbf{b}_2} & \|\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1\mathbf{A}_1^T - \mathbf{e}\mathbf{b}_1\|_F^2 \\ & + \|\mathbf{Y} - \mathbf{X}_2\mathbf{B}_2\mathbf{A}_2^T - \mathbf{e}\mathbf{b}_2\|_F^2 \end{aligned} \tag{5}$$

where  $\mathbf{B}_i \in \mathbb{R}^{d \times r}$ ,  $\mathbf{A}_i \in \mathbb{R}^{c \times r}$ , and  $\mathbf{b}_i \in \mathbb{R}^{1 \times c}$ ,  $i = 1, 2$ . Although the low-rank regression allows  $r$  latent factors of  $\mathbf{X}_i\mathbf{B}_i$  ( $i = 1, 2$ ) to directly represent the response variables, such  $r$  latent factors were basically obtained from  $d$ -dimensional features. When there are a large number of features from neuroimaging data, some of them may not be useful in prediction. The un-useful features may affect the extraction of  $r$  latent factors of  $\mathbf{X}_i$  and also the interpretation of  $\mathbf{Y}$ . Thus, it is helpful to perform feature selection on the low-rank multi-modality regression to exclude the redundant features. In this way, conducting feature selection on  $\mathbf{X}_i$  can be regarded as conducting subspace learning and explaining response variables using ‘clean’ data. To this end, we employ two  $\ell_{2,1}$ -norm regularizers, one for each modality, and additional orthogonal constraints to encourage the latent factors unrelated as follows:

$$\begin{aligned} \min_{\mathbf{A}_1, \mathbf{A}_2, \mathbf{B}_1, \mathbf{B}_2, \mathbf{b}_1, \mathbf{b}_2} & \|\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1\mathbf{A}_1^T - \mathbf{e}\mathbf{b}_1\|_F^2 \\ & + \alpha \|\mathbf{B}_1\|_{2,1} + \|\mathbf{Y} - \mathbf{X}_2\mathbf{B}_2\mathbf{A}_2^T - \mathbf{e}\mathbf{b}_2\|_F^2 \\ & + \beta \|\mathbf{B}_2\|_{2,1}, \text{ s.t., } \mathbf{A}_1^T \mathbf{A}_1 = \mathbf{I}_r, \mathbf{A}_2^T \mathbf{A}_2 = \mathbf{I}_r. \end{aligned} \tag{6}$$

<sup>4</sup>In this work, we extract the same number of features from MRI and PET as described in Section 2.1 and thus their feature dimensions are the same. However, it should be noted that the proposed method can be easily extended to multiple modalities with different numbers of features. Moreover, in the multi-modality case of this work,  $r < \min\{\text{rank}(\mathbf{B}_i), \text{rank}(\mathbf{A}_i)\}$  or  $r < \min\{\text{rank}(\mathbf{B}_i), \text{rank}(\mathbf{A})\}$ ,  $i = 1, 2$ .

where  $\mathbf{I}_r \in \mathbb{R}^{r \times r}$  and  $\alpha$  and  $\beta$  are the tuning parameters. The  $\ell_{2,1}$ -norm regularizers on  $\mathbf{B}_i$  penalize coefficients of  $\mathbf{B}_i$  in a row-wise manner for joint selection or un-selection of the features in predicting the response variables.

It is worth noting that the column-wise low-rank constraints and the row-wise  $\ell_{2,1}$ -norm regularizers on  $\mathbf{B}_i$  have the effects of conducting subspace learning and feature selection on  $\mathbf{X}_i$ , respectively. That is, the low-rank constraints obtain  $r$  latent factors and the  $\ell_{2,1}$ -norm regularizers remove redundant features, both from  $d$  features of  $\mathbf{X}_i$ . In this work, we propose to consider both low-rankness and sparsity in regression to conduct a low-rank dimensionality reduction (LRDR), with the goal of selecting informative features by considering the relations between the neuroimaging features and the response variables via  $\ell_{2,1}$ -norm regularizers and utilizing the relations among response variables via low-rank constraints. Moreover, motivated from the previous studies that the presentation of informative brain regions in MRI for AD diagnosis are different from those in PET [15, 34], we consider the difference between modalities by applying  $\ell_{2,1}$ -norm regularizers on  $\mathbf{B}_i$  for each modality separately with the assumption that MRI and PET may select different brain regions for AD diagnosis.

We also take advantage of the relations among modalities by replacing two orthogonal variables (*i.e.*,  $\mathbf{A}_1$  and  $\mathbf{A}_2$ ) in (6) with an orthogonal variable  $\mathbf{A}$  as follows:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{B}_1, \mathbf{B}_2, \mathbf{b}_1, \mathbf{b}_2} & \quad \|\mathbf{Y} - \mathbf{X}_1 \mathbf{B}_1 \mathbf{A}^T - \mathbf{e}\mathbf{b}_1\|_F^2 \\ & + \alpha \|\mathbf{B}_1\|_{2,1} + \|\mathbf{Y} - \mathbf{X}_2 \mathbf{B}_2 \mathbf{A}^T - \mathbf{e}\mathbf{b}_2\|_F^2 \\ & + \beta \|\mathbf{B}_2\|_{2,1}, \text{ s.t., } \mathbf{A}^T \mathbf{A} = \mathbf{I}_r. \end{aligned} \tag{7}$$

where  $\mathbf{B}_1$  and  $\mathbf{B}_2$ , respectively, are the subspace matrices of MRI and PET, and  $\mathbf{A}$  is the shared regression parameter matrix of two modalities. The reason of such replacement is that  $\mathbf{A}$  connects two modalities. Specifically, MRI and PET have different low-level representations (*i.e.*,  $\mathbf{X}_1$  and  $\mathbf{X}_2$ ) but share the same high-level representation  $\mathbf{Y}$ , thus it should have relation between each of the feature matrices and the response matrix. Once conducting dimensionality reduction on  $\mathbf{X}_1$  and  $\mathbf{X}_2$ , the new low-dimensional representations (*i.e.*,  $\mathbf{X}_1 \mathbf{B}_1$  and  $\mathbf{X}_2 \mathbf{B}_2$ ) could change the distribution of original low-level representations, so a rotation (or transformation)  $\mathbf{A}^T$  should be used to transfer them into the original label space spanned by the high-level representation  $\mathbf{Y}$ . In other words,  $\mathbf{Y}$  should take a rotation (*i.e.*,  $\mathbf{A}$ ) for seamlessly connecting these two new spaces. Such an orthogonal rotation step naturally makes up for the disadvantage resulted by the LRDR step, and explores the relation among modalities.

### 3.4 OLRDR on multi-output multi-modality data

In neuroimaging study, one often conducts a binary classification, such as AD vs. NC, MCI vs. NC, and progressive MCI (pMCI) vs. stable MCI (sMCI), for AD diagnosis. In this way, the dimensions  $c$  of  $\mathbf{Y}$  is low, *e.g.*,  $c = 2$ , for binary classification with 0-1 encoding, so the value of  $r$  may be very small according to the constraint, *i.e.*,  $r \leq \min(d, c)$ . Thus (6) makes little improvement. We, therefore, extend the abstract low-dimensional representation of  $\mathbf{Y}$  in the conventional AD study to a concrete multi-output representation by exploiting the clustering methods, based on the inter-subject variability (or subject variability) of neuroimaging data [20, 39].

In imaging analysis, a low-level representation  $\mathbf{X}$  and a high-level representation  $\mathbf{Y}$  characterize, respectively, the concreteness and abstractness of imaging data. The high-level

representation can be characterized with more details since the inter-subject variability indicates that neuroimaging data may have multiple peaks in distribution [20]. In AD study, the high-level representation of MCI can be further categorized into sub-classes, such as pMCI and sMCI. In this way, the complicated distribution of a class with multiple peaks can be modeled with multiple simple distributions, one for each sub-class. By following the previous work in [39], we propose to divide each class, *i.e.*, each high-level representation (*e.g.*, either AD or NC in the classification of AD vs. NC) into sub-classes via a clustering method (*e.g.*, hierarchical clustering [9] in this paper) and each of the resulting clusters is regarded as a sub-class.

Specifically, given the response matrix  $\mathbf{Y} \in \mathbb{R}^{n \times c}$ , we separately cluster the subjects in each  $i$ -th original class to form its corresponding sub-classes by denoting the  $m_i$  extended sub-classes of the  $i$ -th original class as  $[\mathbf{Y}_{\sum_{j=1}^{i-1} m_j+1}, \dots, \mathbf{Y}_{\sum_{j=1}^{i-1} m_j+m_i}]$  and each row has only one ‘1’, which means that each subject in  $m_i$  sub-classes only belongs to one sub-class. The new response matrix  $\hat{\mathbf{Y}}$  (*i.e.*,  $\hat{\mathbf{Y}} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_c, \mathbf{Y}_{c+1}, \dots, \mathbf{Y}_{c+m_1}, \dots, \mathbf{Y}_{c+\sum_{j=1}^{c-1} m_j+1}, \dots, \mathbf{Y}_{c+\sum_{j=1}^{c-1} m_j+m_c}] \in \{0, 1\}^{n \times m}$ , where  $m = c + \sum_{i=1}^c m_i$  includes two parts, *i.e.*, original labels copied in the first  $c$  columns and the extended labels produced by a clustering method in the remaining  $(m - c)$  columns. By rearranging  $\hat{\mathbf{Y}}$  as  $\hat{\mathbf{Y}} = [\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_m]$ , our final objective function can be written as follows:

$$\begin{aligned} & \min_{\hat{\mathbf{A}}, \hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2, \hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2} \|\hat{\mathbf{Y}} - \mathbf{X}_1 \hat{\mathbf{B}}_1 \hat{\mathbf{A}}^T - \mathbf{e}\hat{\mathbf{b}}_1\|_F^2 \\ & + \alpha \|\hat{\mathbf{B}}_1\|_{2,1} + \|\hat{\mathbf{Y}} - \mathbf{X}_2 \hat{\mathbf{B}}_2 \hat{\mathbf{A}}^T - \mathbf{e}\hat{\mathbf{b}}_2\|_F^2 \\ & + \beta \|\hat{\mathbf{B}}_2\|_{2,1}, \text{ s.t., } \hat{\mathbf{A}}^T \hat{\mathbf{A}} = \mathbf{I}_r \end{aligned} \tag{8}$$

where  $\hat{\mathbf{Y}} \in \mathbb{R}^{n \times m}$ ,  $\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2 \in \mathbb{R}^{d \times r}$ ,  $\hat{\mathbf{A}} \in \mathbb{R}^{m \times r}$  and  $\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2 \in \mathbb{R}^{1 \times m}$ .

After optimizing (8) by Algorithm 1, we have different zero row vectors in both  $\hat{\mathbf{B}}_1$  and  $\hat{\mathbf{B}}_2$ , so we can discard the irrelevant or noisy components, *i.e.*, the features whose regression coefficient vectors are zero in the rows on  $\hat{\mathbf{B}}_1$  and  $\hat{\mathbf{B}}_2$ . Given representative features from MRI and PET, we concatenate the reduced features and then use them to build binary classifiers with SVM.

### 3.5 Optimization

This section describes the optimization process to find optimal parameters of  $\hat{\mathbf{A}}$ ,  $\hat{\mathbf{b}}_1$ ,  $\hat{\mathbf{b}}_2$ ,  $\hat{\mathbf{B}}_1$  and  $\hat{\mathbf{B}}_2$ . Specifically, we iteratively conduct the following two steps until satisfying predefined conditions: (i) Update  $\hat{\mathbf{A}}$  with fixed  $\hat{\mathbf{b}}_1$ ,  $\hat{\mathbf{b}}_2$ ,  $\hat{\mathbf{B}}_1$  and  $\hat{\mathbf{B}}_2$ , *i.e.*, the orthogonal rotation step. (ii) Update  $\hat{\mathbf{b}}_1$ ,  $\hat{\mathbf{b}}_2$ ,  $\hat{\mathbf{B}}_1$  and  $\hat{\mathbf{B}}_2$  with fixed  $\hat{\mathbf{A}}$ , *i.e.*, the LRDR step.

#### 3.5.1 Update $\hat{\mathbf{A}}$ with fixed $\hat{\mathbf{b}}_1$ , $\hat{\mathbf{b}}_2$ , $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$ .

For fixed  $\hat{\mathbf{b}}_1$ ,  $\hat{\mathbf{b}}_2$ ,  $\hat{\mathbf{B}}_1$  and  $\hat{\mathbf{B}}_2$ , the optimization problem in (8) reduces to

$$\begin{aligned} & \min_{\hat{\mathbf{A}}} \|\hat{\mathbf{Y}} - \mathbf{X}_1 \hat{\mathbf{B}}_1 \hat{\mathbf{A}}^T - \mathbf{e}\hat{\mathbf{b}}_1\|_F^2 \\ & + \|\hat{\mathbf{Y}} - \mathbf{X}_2 \hat{\mathbf{B}}_2 \hat{\mathbf{A}}^T - \mathbf{e}\hat{\mathbf{b}}_2\|_F^2, \text{ s.t., } \hat{\mathbf{A}}^T \hat{\mathbf{A}} = \mathbf{I}_r. \end{aligned} \tag{9}$$



After simple mathematical manipulation, (9) becomes:

$$\min_{\hat{\mathbf{A}}} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\mathbf{A}}^T\|_F^2, \text{ s.t., } \hat{\mathbf{A}}^T \hat{\mathbf{A}} = \mathbf{I}_r, \tag{10}$$

where  $\tilde{\mathbf{Y}} = [\hat{\mathbf{Y}} - \mathbf{e}\hat{\mathbf{b}}_1; \hat{\mathbf{Y}} - \mathbf{e}\hat{\mathbf{b}}_2] \in \mathbb{R}^{2n \times m}$  and  $\tilde{\mathbf{X}} = [\mathbf{X}_1\hat{\mathbf{B}}_1; \mathbf{X}_2\hat{\mathbf{B}}_2] \in \mathbb{R}^{2n \times r}$ . Equation (10) is actually an orthogonal Procrustes problem [6]. The optimal solution of  $\hat{\mathbf{A}}$  is  $\mathbf{U}\mathbf{V}^T$ , where  $\mathbf{U} \in \mathbb{R}^{m \times r}$  and  $\mathbf{V} \in \mathbb{R}^{r \times r}$  are obtained from the singular value decomposition of  $\tilde{\mathbf{Y}}^T \tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ , and  $\mathbf{D} \in \mathbb{R}^{r \times r}$  is a diagonal matrix.

### 3.5.2 Update $\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2, \hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$ with fixed $\hat{\mathbf{A}}$ .

For fixed  $\hat{\mathbf{A}}$ , the optimization problem in (8) reduces to

$$\begin{aligned} \min_{\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2, \hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2} & \|\hat{\mathbf{Y}} - \mathbf{X}_1\hat{\mathbf{B}}_1\hat{\mathbf{A}}^T - \mathbf{e}\hat{\mathbf{b}}_1\|_F^2 + \alpha\|\hat{\mathbf{B}}_1\|_{2,1} \\ & + \|\hat{\mathbf{Y}} - \mathbf{X}_2\hat{\mathbf{B}}_2\hat{\mathbf{A}}^T - \mathbf{e}\hat{\mathbf{b}}_2\|_F^2 + \beta\|\hat{\mathbf{B}}_2\|_{2,1}, \end{aligned} \tag{11}$$

By setting the derivative of (11) w.r.t.  $\hat{\mathbf{b}}_1$  to zero, we have:

$$2\mathbf{e}^T \mathbf{X}_1 \hat{\mathbf{B}}_1 \hat{\mathbf{A}}^T + 2\mathbf{e}^T \hat{\mathbf{e}}\hat{\mathbf{b}}_1 - 2\mathbf{e}^T \hat{\mathbf{Y}} = 0 \tag{12}$$

After simple mathematical manipulation, we have:

$$\hat{\mathbf{b}}_1 = \frac{1}{n}\mathbf{e}^T \hat{\mathbf{Y}} - \frac{1}{n}\mathbf{e}^T \mathbf{X}_1 \hat{\mathbf{B}}_1 \hat{\mathbf{A}}^T \tag{13}$$

Similarly, by setting the derivative of (11) w.r.t.  $\hat{\mathbf{b}}_2$  to zero, we obtain optimal  $\hat{\mathbf{b}}_2$  as:

$$\hat{\mathbf{b}}_2 = \frac{1}{n}\mathbf{e}^T \hat{\mathbf{Y}} - \frac{1}{n}\mathbf{e}^T \mathbf{X}_2 \hat{\mathbf{B}}_2 \hat{\mathbf{A}}^T \tag{14}$$

Substituting (13) and (14) into (11) and letting  $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}^T \in \mathbb{R}^{n \times n}$ , where  $\mathbf{I}_n \in \mathbb{R}^{n \times n}$  is an identity matrix, we have:

$$\begin{aligned} \min_{\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2} & \|\mathbf{H}\hat{\mathbf{Y}} - \mathbf{H}\mathbf{X}_1\hat{\mathbf{B}}_1\hat{\mathbf{A}}^T\|_F^2 + \alpha\|\hat{\mathbf{B}}_1\|_{2,1} \\ & + \|\mathbf{H}\hat{\mathbf{Y}} - \mathbf{H}\mathbf{X}_2\hat{\mathbf{B}}_2\hat{\mathbf{A}}^T\|_F^2 + \beta\|\hat{\mathbf{B}}_2\|_{2,1}, \end{aligned} \tag{15}$$

As  $\hat{\mathbf{A}}$  has orthogonal columns, there is a matrix  $\hat{\mathbf{A}}^\perp$  with orthogonal columns such that  $(\hat{\mathbf{A}}, \hat{\mathbf{A}}^\perp)$  is an orthogonal matrix. Thus, we have

$$\begin{aligned} & \|\mathbf{H}\hat{\mathbf{Y}} - \mathbf{H}\mathbf{X}_1\hat{\mathbf{B}}_1\hat{\mathbf{A}}^T\|_F^2 + \|\mathbf{H}\hat{\mathbf{Y}} - \mathbf{H}\mathbf{X}_2\hat{\mathbf{B}}_2\hat{\mathbf{A}}^T\|_F^2 \\ & = \|(\mathbf{H}\hat{\mathbf{Y}} - \mathbf{H}\mathbf{X}_1\hat{\mathbf{B}}_1\hat{\mathbf{A}}^T)(\hat{\mathbf{A}}, \hat{\mathbf{A}}^\perp)\|_F^2 \\ & \quad + \|(\mathbf{H}\hat{\mathbf{Y}} - \mathbf{H}\mathbf{X}_2\hat{\mathbf{B}}_2\hat{\mathbf{A}}^T)(\hat{\mathbf{A}}, \hat{\mathbf{A}}^\perp)\|_F^2 \\ & = \|\mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}} - \mathbf{H}\mathbf{X}_1\hat{\mathbf{B}}_1\|_F^2 + \|\mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}}^\perp\|_F^2 \\ & \quad + \|\mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}} - \mathbf{H}\mathbf{X}_2\hat{\mathbf{B}}_2\|_F^2 + \|\mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}}^\perp\|_F^2 \end{aligned} \tag{16}$$

Both the second term and the fourth term in (16) do not involve  $\hat{\mathbf{B}}_1$  and  $\hat{\mathbf{B}}_2$ . Therefore, for fixed  $\hat{\mathbf{A}}$ ,  $\hat{\mathbf{b}}_1$  and  $\hat{\mathbf{b}}_2$ , the objective function in (15) reduces to

$$\min_{\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2} \|\mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}} - \mathbf{H}\mathbf{X}_1\hat{\mathbf{B}}_1\|_F^2 + \alpha\|\hat{\mathbf{B}}_1\|_{2,1} + \|\mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}} - \mathbf{H}\mathbf{X}_2\hat{\mathbf{B}}_2\|_F^2 + \beta\|\hat{\mathbf{B}}_2\|_{2,1}. \tag{17}$$

In this work, we used the framework of iteratively reweighted least square [10] to optimize (17), which is thus equivalent to

$$\min_{\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2} \|\mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}} - \mathbf{H}\mathbf{X}_1\hat{\mathbf{B}}_1\|_F^2 + \alpha tr(\hat{\mathbf{B}}_1^T \mathbf{P} \hat{\mathbf{B}}_1) + \|\mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}} - \mathbf{H}\mathbf{X}_2\hat{\mathbf{B}}_2\|_F^2 + \beta tr(\hat{\mathbf{B}}_2^T \mathbf{Q} \hat{\mathbf{B}}_2), \tag{18}$$

where  $\mathbf{P} \in \mathbb{R}^{d \times d}$  and  $\mathbf{Q} \in \mathbb{R}^{d \times d}$ , respectively, are the diagonal matrices with  $p_{jj} = \frac{1}{2\|\hat{\mathbf{B}}_1^j\|_2^2}$  and  $q_{jj} = \frac{1}{2\|\hat{\mathbf{B}}_2^j\|_2^2}$ ,  $j = 1, \dots, d$ . By setting (18) w.r.t.  $\hat{\mathbf{B}}_1$  to zero, we obtain:

$$\hat{\mathbf{B}}_1 = (\mathbf{X}_1^T \mathbf{H}\mathbf{X}_1 + \alpha \mathbf{P})^{-1} \mathbf{X}_1^T \mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}}. \tag{19}$$

Similarly, by setting the derivative of (18) w.r.t.  $\hat{\mathbf{B}}_2$  to zero, we have:

$$\hat{\mathbf{B}}_2 = (\mathbf{X}_2^T \mathbf{H}\mathbf{X}_2 + \beta \mathbf{Q})^{-1} \mathbf{X}_2^T \mathbf{H}\hat{\mathbf{Y}}\hat{\mathbf{A}} \tag{20}$$

We summarize the pseudo code of solving (8) in Algorithm 1. It can be proved that the objective function value in (8) monotonically decreases in each iteration using Algorithm 1 according to literature [10]. In Algorithm 1, the LRDR step and the orthogonal rotation step adjust each other to help find optimal parameters  $\hat{\mathbf{b}}_1$ ,  $\hat{\mathbf{b}}_2$ ,  $\hat{\mathbf{B}}_1$ ,  $\hat{\mathbf{B}}_2$ , and  $\hat{\mathbf{A}}$ , thus ensuring the output of class-discriminative features.

---

**Algorithm 1** Pseudo code of solving (8)

---

**Input:**  $\mathbf{X}_1 \in \mathbb{R}^{n \times d}$ ,  $\mathbf{X}_2 \in \mathbb{R}^{n \times d}$ ,  $\mathbf{Y} \in \mathbb{R}^{n \times c}$ ,  $\alpha, \beta$ ;

**Output:**  $\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2$ ;

- 1 Do clustering on each class of  $\mathbf{Y}$  to output  $\hat{\mathbf{Y}}$ ;
  - 2 Initialize  $t = 1$ ;
  - 3 Initialize  $\hat{\mathbf{b}}_1(1)$  and  $\hat{\mathbf{b}}_2(1)$  as two random vectors /\*  $\hat{\mathbf{b}}_i(t)$ : the  $t$ -th iteration result of  $\hat{\mathbf{b}}_i$ ,  $i = 1, 2$ . \*/
  - 4 Initialize  $\hat{\mathbf{B}}_1(1)$  and  $\hat{\mathbf{B}}_2(1)$  as two random matrices;
  - 5 **repeat**
  - 6     Update  $\hat{\mathbf{A}}(t + 1)$  via (10);
  - 7     Update  $\hat{\mathbf{b}}_1(t + 1)$  via (13);
  - 8     Update  $\hat{\mathbf{b}}_2(t + 1)$  via (14);
  - 9     Update  $\mathbf{P}(t + 1)$  via  $p_{jj} = \frac{1}{2\|\hat{\mathbf{B}}_{(t+1)}^j\|_2^2}$ ,  $j = 1, \dots, d$ ;
  - 10    Update  $\mathbf{Q}(t + 1)$  via  $q_{jj} = \frac{1}{2\|\hat{\mathbf{B}}_{(t+1)}^j\|_2^2}$ ,  $j = 1, \dots, d$ ;
  - 11    Update  $\hat{\mathbf{B}}_1(t + 1)$  via (19);
  - 12    Update  $\hat{\mathbf{B}}_2(t + 1)$  via (20);
  - 13     $t = t + 1$ ;
  - 14 **until** The difference between the objective function values of (8) within two sequential iterations is less than  $10^{-5}$ ;
-

## 4 Experimental results

We conducted various experiments on the ADNI dataset to compare our method with the state-of-the-art methods.

### 4.1 Competing methods

In order to validate the effectiveness of the proposed method, we compared with the following methods: 1) Original features based method ('Original') conducts classification on the concatenation of MRI data and PET data with all features, *i.e.*, without feature selection.

2) The feature selection methods on single-modality data include MRI-based Feature Selection (MRIFS) and PET-based Feature Selection (PETFS), where MRIFS and PETFS, respectively, conduct feature selection via Inter-Modality-based Feature Selection (IDFS) [15] with only MRI data and PET data.

3) The state-of-the-art feature selection methods include Multi-Modal Multi-Task (M3T) [34], IDFS [15], and Sparse Joint Classification and Regression (SJCR) [29]. M3T includes two steps: (1) using multi-task feature selection to determine a common subset for multiple response variables (or multiple tasks) from each modality, and (2) using a multi-kernel decision fusion to integrate the selected features from all modalities for prediction. IDFS conducts feature selection by simultaneously imposing the preservation of the inter-modality relationship (*i.e.*, preserving the relative distance between the feature vectors extracted from different modalities of the same subject) on multi-modality data, and also enforcing the sparseness of selected features from each modality. SJCR uses a logistic loss function and a least square loss function simultaneously along with an  $\ell_{2,1}$ -norm regularizer for multi-task feature selection. The methods (*e.g.*, M3T, IDFS, and SJCR) embedded their feature selection models into a multi-task learning framework and selected the same features for both MRI and PET. Both M3T and SJCR do not consider the relations among tasks, while IDFS considers the preservation of relative distance between features.

(4) Baseline: (6) that considers the subclass issue is used to test the effectiveness of the assumption of (8), *i.e.*, different modalities share the same high-level representation.

### 4.2 ADNI study

#### 4.2.1 Experimental setup

We considered three binary classification tasks (including AD vs. NC, MCI vs. NC, and pMCI vs. sMCI) on two-modality data (*i.e.*, MRI and PET) to evaluate the performance of all competing methods, in terms of the metrics of classification accuracy, sensitivity, specificity, and Area Under a receiver operating characteristic Curve (AUC).

We used a 10-fold cross-validation method to compare all methods. Specifically, we first randomly partitioned the whole dataset into 10 subsets. We then selected one subset for testing and used the remaining 9 subsets for training. We repeated the whole process 10 times to avoid the possible bias during dataset partitioning for cross-validation. The final result reported in Tables 2, 3, 4 and 5 was computed by averaging results from all ten experiments.

For the model selection, we set  $\alpha \in \{10^{-5}, \dots, 10^5\}$ ,  $\beta \in \{10^{-3}, \dots, 10^3\}$ , the number of sub-classes for each response variable  $\in \{5, 8, 10, 12\}$ , and  $r \in \{10\%, 30\%, 50\%$ ,

**Table 2** Classification performance of all methods for AD vs. NC

	Accuracy	Sensitivity	Specificity	AUC
Original	0.828±0.12◊	0.880±0.10◊	0.770±0.22◊	0.903±0.12◊
MRIFS	0.846±0.08◊	0.886±0.09◊	0.805±0.14◊	0.910±0.07◊
PETFS	0.834±0.08◊	0.857±0.12◊	0.809±0.16◊	0.889±0.11◊
M3T	0.855±0.09◊	0.866±0.10◊	0.843±0.12◊	0.925±0.06*
IDFS	0.862±0.22◊	0.876±0.11◊	0.847±0.12◊	0.927±0.06*
SJCR	0.848±0.08◊	0.892±0.10*	0.802±0.13◊	0.920±0.05*
Baseline	0.869±0.11◊	0.897±0.16*	0.856±0.12◊	0.925±0.09*
Proposed	<b>0.917±0.07</b>	<b>0.918±0.08</b>	<b>0.916±0.08</b>	<b>0.943±0.05</b>

70%, 90%} of the number of full-rank<sup>5</sup> in (8) and  $C \in \{2^{-5}, \dots, 2^5\}$  in SVM with a linear kernel by a 5-fold inner cross-validation, where the training data are further partitioned into five parts to select the best combination of parameters with the highest classification accuracy, to be used in testing.. For fair comparison, we also conducted 5-fold inner cross-validation to conduct model selection for each competing method. Specifically, for MRIFS, PETFS, and IDFS, we followed the literature [15] to set the values in the ranges of  $\lambda_1 \in \{10^{-3}, 10^{-2}, \dots, 10^3\}$  and  $\lambda_2 \in \{10^{-4}, 10^{-2}, \dots, 10^1\}$ . For SJCR and M3T, we optimized their sparsity parameters by cross-validating the value in the ranges of  $\{10^{-5}, 10^{-3}, \dots, 10^5\}$  (as in [29]) and  $\{10^{-5}, \dots, 10^2\}$ , respectively, to obtain their best performance.

### 4.3 Classification results

We summarized the performances of all methods on four binary classification tasks in Tables 2–5.<sup>6</sup> The proposed method outperformed all the competing methods in all classification tasks. For example, for four binary classification tasks, our method improved, on average, the classification accuracy by 4.9% (vs. Baseline), 5.4% (vs. IDFS), 6.9% (vs. M3T), 7.2% (vs. PETFS), 8.0% (vs. MRIFS), and 9.9% (vs. Original), respectively. Meanwhile, compared to all the competing feature selection methods, our method achieved the maximal and minimal improvements by 8.3% (vs. PETFS) and 4.8% (vs. Baseline) on AD vs. NC, 8.8% (vs. SJCR) and 5.0% (vs. Baseline) on MCI vs. NC, 7.4% (vs. PETFS) and 4.9% (vs. Baseline) on pMCI vs. sMCI, and 9.6% (vs. MRIFS) and 4.0% (vs. Baseline) on AD vs. MCI, respectively. The reason may be that, the proposed method considered the following constraints for conducting feature selection: 1) all kinds of relations inherent in data; 2) iteratively conducting the procedures of low-rank dimensionality reduction and orthogonal rotation; and 3) selecting different features from different modalities.

It is noteworthy that all methods achieved the highest classification performance on AD vs. NC. For example, the average classification accuracy of all methods was 85.7% (AD vs.

<sup>5</sup>In our experiments, we used matlab function ‘floor’ to discretize the real values of  $r$ .

<sup>6</sup>In Tables 2–5, the boldface denotes the maximum performance in each column. (Symbols \* and ◊, respectively, represent statistically significant difference between the proposed method and the comparison methods under  $p < 0.05$  and  $p < 0.001$ , on the paired-sample t-tests at 95% significance level.)

**Table 3** Classification performance of all methods for MCI vs. NC

	Accuracy	Sensitivity	Specificity	AUC
Original	0.631±0.13◇	0.201±0.26◇	0.703±0.08◇	0.590±0.11◇
MRIFS	0.662±0.05◇	0.204±0.15◇	0.891±0.23◇	0.660±0.11◇
PETFS	0.661±0.07◇	0.316±0.19◇	0.834±0.11◇	0.648±0.12◇
M3T	0.672±0.06◇	0.330±0.18◇	0.844±0.09◇	0.651±0.08◇
IDFS	0.676±0.23◇	0.329±0.18◇	0.850±0.09◇	0.661±0.08◇
SJCR	0.657±0.05◇	0.335±0.17◇	0.899±0.11*	0.671±0.08*
Baseline	0.689±0.11◇	0.371±0.12◇	0.886±0.15◇	0.668±0.11*
Proposed	<b>0.745±0.05</b>	<b>0.428±0.24</b>	<b>0.903±0.07</b>	<b>0.701±0.09</b>

NC), 67.4% (MCI vs. NC), 66.4% (pMCI vs. sMCI), and 72.7% (AD vs. MCI), respectively. Another observation is the low sensitivity (or specificity) for the classification of MCI vs. NC (or pMCI vs. sMCI). A possible reason for these two observations is that the difference between AD and NC is relatively prominent while there is no substantial difference between MCI and NC (or between pMCI and sMCI). Furthermore, our proposed method outperformed Baseline on all the datasets, which supports our assumption “different modalities share the same high-level representation”.

#### 4.4 Most discriminative brain regions

We investigated the brain regions as potential biomarkers for AD diagnosis based on the selected frequency of the brain regions by the proposed method. We listed the frequency (defined as the probability of a feature appeared in the 100 experiments - ten times of ten-fold cross-validation) of each feature in our 10 repeated 10-fold cross-validation experiments in Figure 2 and also visualized the top selected brain regions in Figure 3. In the classification task of MCI vs. NC, our method selected regions of uncus right, hippocampal formation right, uncus left, middle temporal gyrus left, hippocampal formation left, amyg-

**Table 4** Classification performance of all methods for pMCI vs. sMCI

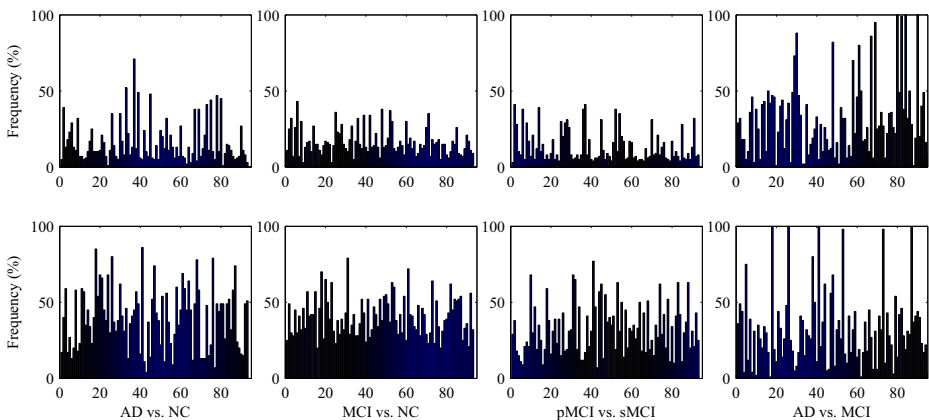
	Accuracy	Sensitivity	Specificity	AUC
Original	0.622±0.13◇	0.722±0.22◇	0.376±0.25◇	0.576±0.17◇
MRIFS	0.654±0.12◇	0.735±0.18◇	0.561±0.23*	0.681±0.18◇
PETFS	0.652±0.13◇	0.768±0.20◇	0.520±0.22◇	0.716±0.16*
M3T	0.659±0.13◇	0.801±0.17◇	0.496±0.23◇	0.712±0.15*
IDFS	0.674±0.05◇	0.834±0.18*	0.474±0.09◇	0.687±0.09◇
SJCR	0.655±0.11◇	0.741±0.19◇	0.556±0.23◇	0.699±0.17◇
Baseline	0.671±0.13◇	0.842±0.16◇	0.570±0.27*	0.715±0.19*
Proposed	<b>0.726±0.11</b>	<b>0.848±0.15</b>	<b>0.585±0.28</b>	<b>0.735±0.15</b>

**Table 5** Classification performance of all methods for AD vs. MCI

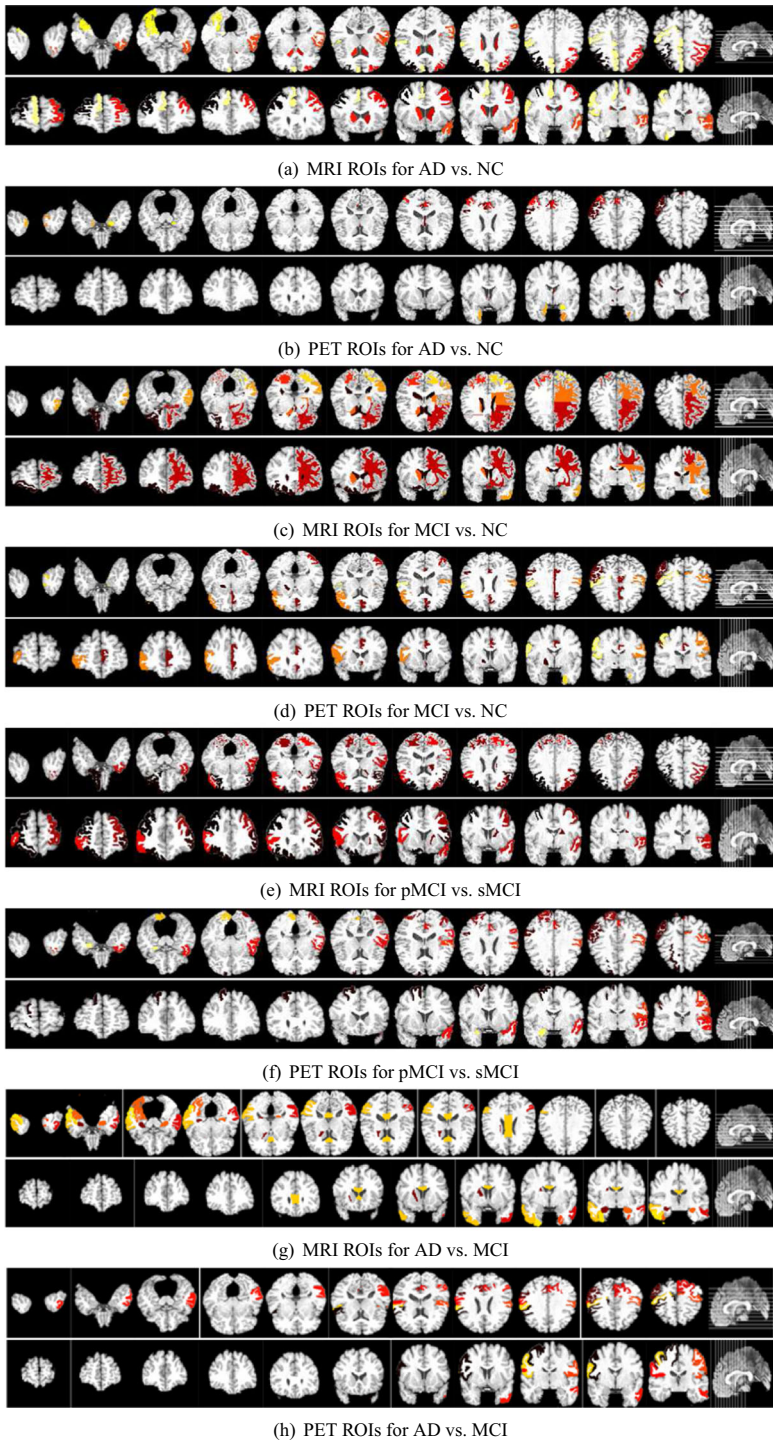
	Accuracy	Sensitivity	Specificity	AUC
Original	0.645±0.17◇	0.342±0.19◇	0.784±0.12◇	0.583±0.09◇
MRIFS	0.693±0.08◇	0.442±0.14*	0.884±0.07◇	0.623±0.10◇
PETFS	0.715±0.07◇	0.408±0.15◇	0.898±0.07◇	0.694±0.08◇
M3T	0.728±0.08◇	0.364±0.13◇	0.932±0.06*	0.705±0.09◇
IDFS	0.748±0.11*	0.388±0.12◇	0.930±0.09*	0.737±0.12◇
SJCR	0.745±0.17*	0.376±0.18◇	0.930±0.13*	0.744±0.12◇
Baseline	0.753±0.11*	0.428±0.18*	0.935±0.14*	0.752±0.15*
Proposed	<b>0.789±0.17</b>	<b>0.463±0.09</b>	<b>0.955±0.12</b>	<b>0.787±0.11</b>

dala left, middle temporal gyrus right, and amygdala right as top selected ones, which have also been pointed out in the previous work [34] and have been shown to be highly related to AD or related dementia (*e.g.*, MCI) in clinical diagnosis [1, 4, 18]. Hence, brain regions selected by our method could be further incorporated for future clinical analysis.

Besides, we had some interesting observations: 1) In the process of feature selection, our method selected less number of MRI features than PET features in our experiments. For example, our method selected, on average, 16.3/40.3, 16.6/40.8, 13.3/32.6, and 32.1/31.6, respectively, of MRI/PET features on the classification tasks of AD vs. NC, MCI vs. NC, pMCI vs. sMCI, and AD vs. MCI. This observation manifested that MRI and PET could provide complementary information to each other to enhance the classification performance of multi-modality data for AD classification. 2) Different classification tasks have selected different brain regions from each modality.

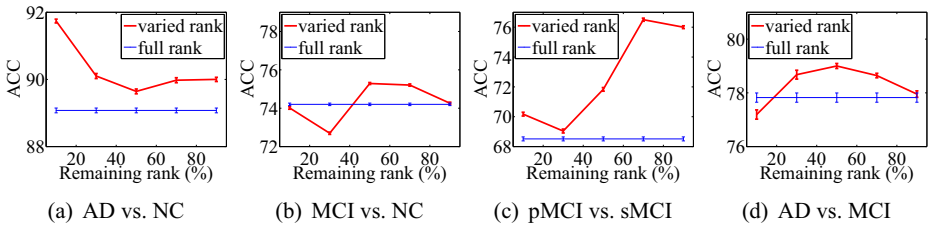


**Figure 2** Frequency of the selected ROIs of MRI (upper row) and PET (down row) by the proposed method on four binary classification tasks, where horizontal axis represents the number of ROIs and vertical axis represents the frequency (ranging from 0 to 100) in ten cross-validation experiments.  $Frequency_1 = 11$  in the first sub-figure (*i.e.*, the top left sub-figure) means that the first ROI was selected 11 times over 100 repeats by the proposed method



**Figure 3** Top selected regions of the proposed method on three binary classification tasks





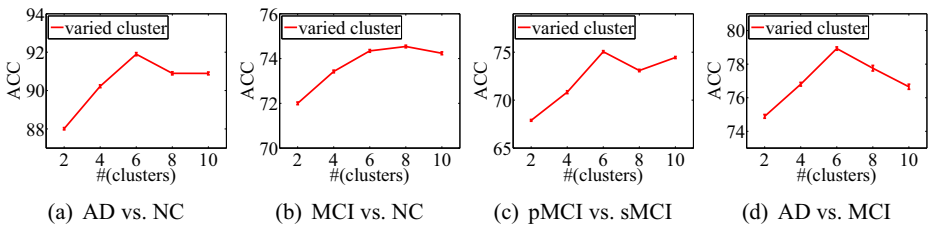
**Figure 4** Classification accuracy of our method with different number of ranks of coefficient matrix in the case of the fixed number of subclasses (*i.e.*, 6) for each response variable on four different classification tasks

### 4.5 Discussion

In this section, we investigate three aspects of the proposed method, *i.e.*, the effect of different number of ranks (*i.e.*,  $r$ ) in (8), and the effect of different number of sub-classes for each original class in Section 3.4.

Figures 4 and 5 visualize, respectively, the change of classification accuracies according to different values of a rank, *i.e.*,  $r \in \{10\%, 30\%, 50\%, 70\%, 90\%, 100\%$  of the number of the full rank and different number of clusters or sub-classes (*i.e.*,  $\{1, 2, 3, 4, 5\}$ ) in each class. From Figure 4, we observe that the performance with low-rank constraint in most of cases outperformed the performance of the cases with full-rank. This manifested that it is reasonable to analyze neuroimaging data with a low-rank constraint in AD study. The reason is that, the low-rank constraints, conducting subspace learning, help find the low-dimensional structure of high-dimensional neuroimaging data via considering the relations among response variables.

Figure 5 explains that different numbers of sub-classes for each original class outputted different results, and the case, in which the response variables are partitioned into 6 sub-classes, achieved the best classification performance on all three classification tasks in our experiments. Besides, we found that the results of using one cluster 1) were worse than the results of using more than one clusters, which indicates the variability of real datasets; and 2) were better than the results of all the competing methods, which shows the effectiveness of our proposed OLRDR on multi-modality data, *i.e.*, Section 3.3. By combining these two observations together, we conclude that the variability assumption (*i.e.*, Section 3.4) does work.



**Figure 5** Classification accuracy of the proposed method with different numbers of clusters in the case of the fixed number of ranks (*i.e.*,  $r = 5$ ) on four different classification tasks



## 5 Conclusion

In this paper, we focused on taking advantages of different aspects of data such as variability, sparsity, and low-rankness with multiple modalities for AD classification. Specifically, we first extended conventional label representation (*e.g.*, 0-1 encoding) to a multi-output representation, and then iteratively conducted a low-rank dimensionality reduction step and an orthogonal rotation step to select representative features, by taking advantages of all kinds of relations inherent in the neuroimaging data. The experimental results on the ADNI dataset with MRI and PET verified the efficacy of the proposed method over the state-of-the-art methods in terms of classification performance on three binary classification tasks.

According to Section 4.3, the classification performance on both MCI vs. NC and pMCI vs. sMCI are lower than the performance of AD vs. NC. This may result from the subtle changes between two clinical statuses. In our future work, we will focus on selecting representative features via transfer learning methods (*e.g.*, [21]) to further improve the classification performance of AD-to-MCI classification, such as MCI vs. NC and pMCI vs. sMCI. For example, we could use the information of the classification AD vs. NC to help enhance classification performance of the classification pMCI vs. sMCI according to the conclusion in [19], which demonstrated that the features selected by the classification AD vs. NC can be regarded as representative features of the classification pMCI vs. sMCI.

**Acknowledgments** This work was supported in part by NIH grants (EB008374, AG041721, AG049371, AG042599, EB022880). X. Zhu was also supported by the National Natural Science Foundation of China (Grants No: 61573270 and 61876046); the Project of Guangxi Science and Technology (GuiKeAD17195062); the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing; Strategic Research Excellence Fund at Massey University; and Marsden Fund of New Zealand (grant No: MAU1721). H.I. Suk was also supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

1. Chételat, G., Eustache, F., Viader, F., De La Sayette, V., Pélerin, A., Mézenge, F., Hannequin, D., Dupuy, B., Baron, J.-C., Desgranges, B.: FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. *Neurocase* **11**(1), 14–25 (2005)
2. Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C.: Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage* **60**(1), 59–70 (2012)
3. Deng, X., Li, Y., Weng, J., Zhang, J.: Feature selection for text classification: A review. *Multimedia Tools and Applications*, 1–20 (2018)
4. Fox, N.C., Schott, J.M.: Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *Lancet* **363**(9406), 392–394 (2004)
5. Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H.T.: Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. Multimed.* **19**(9), 2045–2055 (2017)
6. Gower, J.C., Dijksterhuis, G.B.: Procrustes problems, vol. 3 (2004)
7. Hu, R., Zhu, X., Cheng, D., He, W., Yan, Y., Song, J., Zhang, S.: Graph self-representation method for unsupervised feature selection. *Neurocomputing* **220**, 130–137 (2017)
8. Izenman, A.J.: Reduced-rank regression for the multivariate linear model. *J. Multivar. Anal.* **5**(2), 248–264 (1975)
9. Johnson, S.C.: Hierarchical clustering schemes. *Psychometrika* **32**(3), 241–254 (1967)

10. Jorgensen, M.: Iteratively reweighted least squares. *Encyclopedia of Environmetrics* (2006)
11. Kabani, N.J.: 3D anatomical atlas of the human brain. In: *Human Brain Mapping Conference* (1998)
12. Lahmiri, S.: Image characterization by fractal descriptors in variational mode decomposition domain: application to brain magnetic resonance. *Physica A: Stat. Mech. Appl.* **456**, 235–243 (2016)
13. Lahmiri, S., Boukadoum, M.: New approach for automatic classification of Alzheimer's disease, mild cognitive impairment and healthy brain magnetic resonance images. *Healthcare Technol. Lett.* **1**(1), 32–36 (2014)
14. Lei, C., Zhu, X.: Unsupervised feature selection via local structure learning and sparse learning. <https://doi.org/10.1007/s11042-017-5381-7>, p. 11 (2017)
15. Liu, F., Wee, C.-Y., Chen, H., Shen, D.: Inter-modality relationship constrained multi-modality multi-task feature selection for alzheimer's disease and mild cognitive impairment identification. *Neuroimage* **84**, 466–475 (2014)
16. Lu, C., Lin, Z., Yan, S.: Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Trans. Image Process.* **24**(2), 646–654 (2015)
17. Ma, Z., Sun, T.: Adaptive sparse reduced-rank regression. arXiv:1403.1922 (2014)
18. Misra, C., Fan, Y., Davatzikos, C.: Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *Neuroimage* **44**(4), 1415–1422 (2009)
19. Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Alzheimer's Disease Neuroimaging Initiative, et al.: Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* **104**, 398–412 (2015)
20. Noppeney, U., Penny, W.D., Price, C.J., Flandin, G., Friston, K.J.: Identification of degenerate neuronal systems based on intersubject variability. *Neuroimage* **30**(3), 885–890 (2006)
21. Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**(10), 1345–1359 (2010)
22. Rocchi, L., Chiari, L., Cappello, A., Horak, F.B.: Identification of distinct characteristics of postural sway in parkinson's disease: a feature selection procedure based on principal component analysis. *Neurosci. Lett.* **394**(2), 140–145 (2006)
23. Sato, J.R., Hoexter, M.Q., Fujita, A., Rohde, L.A.: Evaluation of pattern recognition and feature extraction methods in ADHD prediction. *Frontiers in Systems Neuroscience*, 6 (2012)
24. Shen, D., Davatzikos, C.: HAMMER: hierarchical attribute matching mechanism for elastic registration, vol. 21 (2002)
25. Song, J., Guo, Y., Gao, L., Li, X., Hanjalic, A., Shen, H.T.: From deterministic to generative: Multi-modal stochastic rnns for video captioning. *IEEE Transactions on Neural Networks and Learning Systems.* <https://doi.org/10.1109/TNNLS.2018.2851077> (2017)
26. Song, J., Zhang, H., Li, X., Gao, L., Wang, M., Hong, R.: Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Trans. Image Process.* **27**(7), 3210–3221 (2018)
27. Sui, J., Adali, T., Yu, Q., Chen, J., Calhoun, V.D.: A review of multivariate methods for multimodal fusion of brain imaging data. *J. Neurosci. Methods* **204**(1), 68–81 (2012)
28. Velu, R., Reinsel, G.C.: *Multivariate reduced-rank regression: theory and applications*, vol. 136 (2013)
29. Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A.J., Li, S.: Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In: *MICCAI*, pp. 115–123 (2011)
30. Wang, X., Gao, L., Wang, P., Sun, X., Liu, X.: Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. *IEEE Trans. Multimed.* **20**(3), 634–644 (2018)
31. Yang, M., Zheng, H., Wang, H., McClean, S.: Feature selection and construction for the discrimination of neurodegenerative diseases based on gait analysis. In: *PCTH*, pp. 1–7 (2009)
32. Yang, Y., Zhou, J., Ai, J., Yi, B., Hanjalic, A., Shen, H.T.: Video captioning by adversarial lstm. *IEEE Transactions on Image Processing.* <https://doi.org/10.1109/TIP.2018.2855422> (2018)
33. Yi, B., Yang, Y., Shen, F., Xie, N., Shen, H.T., Li, X.: Describing video with attention based bidirectional lstm. *IEEE Transactions on Cybernetics.* <https://doi.org/10.1109/TCYB.2018.2831447> (2018)
34. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* **59**(2), 895–907 (2012)
35. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**(1), 45–57 (2001)
36. Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R.: Efficient knn classification with different numbers of nearest neighbors. *IEEE Trans. Neural Netw. Learn. Syst.* **29**(5), 1774–1785 (2018)
37. Zheng, W., Zhu, X., Zhu, Y., Hu, R., Lei, C.: Dynamic graph learning for spectral feature selection. *Multimedia Tools and Applications.* <https://doi.org/10.1007/s11042-017-5272-y> (2017)

38. Zheng, W., Zhu, X., Wen, G., Zhu, Y., Yu, H., Gan, J.: Unsupervised feature selection by self-paced learning regularization. *Pattern Recognition Letters*, <https://doi.org/10.1016/j.patrec.2018.06.029> (2018)
39. Zhu, M., Martinez, A.M.: Subclass discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(8), 1274–1286 (2006)
40. Zhu, X., Suk, H.-I., Shen, D.: A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. *Neuroimage* **14**(0), 1–30 (2014)
41. Zhu, X., Zhang, L., Zi, H.: A sparse embedding and least variance encoding approach to hashing. *IEEE Trans. Image Process.* **23**(9), 3737–3750 (2014)
42. Zhu, X., Li, X., Zhang, S.: Block-row sparse multiview multilabel learning for image classification. *IEEE Trans. Cybern.* **46**(2), 450–461 (2016)
43. Zhu, X., Suk, H.-I., Lee, S.-W., Shen, D.: Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. *IEEE Trans. Biomed. Eng.* **63**(3), 607–618 (2016)
44. Zhu, X., Li, X., Zhang, S., Ju, C., Wu, X.: Robust joint graph sparse coding for unsupervised spectral feature selection. *IEEE Trans. Neural Netw. Learn. Syst.* **28**(6), 1263–1275 (2017)
45. Zhu, X., Li, X., Zhang, S., Xu, Z., Yu, L., Wang, C.: Graph pca hashing for similarity search. *IEEE Trans. Multimed.* **19**(9), 2033–2044 (2017)
46. Zhu, X., Suk, H.-I., Huang, H., Shen, D.: Low-rank graph-regularized structured sparse regression for identifying genetic biomarkers. *IEEE Trans. Big Data* **3**(4), 405–414 (2017)
47. Zhu, X., Suk, H.-I., Wang, L., Lee, S.-W., Shen, D.: A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Med. Image Anal.* **38**, 205–214 (2017)
48. Zhu, X., Zhang, S., He, W., Hu, R., Lei, C., Zhu, P.: One-step multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering*, <https://doi.org/10.1109/TKDE.2018.2873378> (2018)
49. Zhu, X., Zhang, S., Hu, R., Zhu, Y., et al.: Local and global structure preservation for robust unsupervised spectral feature selection. *IEEE Trans. Knowl. Data Eng.* **30**(3), 517–529 (2018)
50. Zhu, X., Zhang, S., Li, Y., Zhang, J., Yang, L., Fang, Y.: Low-rank sparse subspace for spectral clustering. *IEEE Transactions on Knowledge and Data Engineering*, <https://doi.org/10.1109/TKDE.2018.2858782> (2018)