NeuroImage 91 (2014) 386–400

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion

Kim-Han Thung [a], Chong-Yaw Wee [a], Pew-Thian Yap [a], Dinggang Shen [a,b,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Biomedical Research Imaging Center (BRIC) and Department of Radiology, University of North Carolina at Chapel Hill, USA
[b] Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

ARTICLE INFO

Article history:
Accepted 18 January 2014
Available online 27 January 2014

Keywords:
Matrix completion
Classification
Multi-task learning
Data imputation

ABSTRACT

In this work, we are interested in predicting the diagnostic statuses of potentially neurodegenerated patients using feature values derived from multi-modality neuroimaging data and biological data, which might be incomplete. Collecting the feature values into a matrix, with each row containing a feature vector of a sample, we propose a framework to predict the corresponding associated multiple target outputs (e.g., diagnosis label and clinical scores) from this feature matrix by performing matrix shrinkage following matrix completion. Specifically, we first *combine the feature and target output matrices into a large matrix and then partition this large incomplete matrix into smaller submatrices*, each consisting of samples with complete feature values (corresponding to a certain combination of modalities) and target outputs. Treating each target output as the outcome of a prediction task, we apply a *2-step multi-task learning algorithm to select the most discriminative features and samples* in each submatrix. Features and samples that are not selected in any of the submatrices are discarded, resulting in a *shrunk version of the original large matrix*. The missing feature values and unknown target outputs of the shrunk matrix is then completed simultaneously. Experimental results using the ADNI dataset indicate that our proposed framework achieves higher classification accuracy at a greater speed when compared with conventional imputation-based classification methods and also yields competitive performance when compared with the state-of-the-art methods.

Introduction

Alzheimer's Disease (AD) is the most prevalent form of dementia. It is ultimately fatal and is ranked as the sixth leading cause of death in the United States in year 2012 (Alzheimer's Association, 2013). Neurodegeneration associated with AD is progressive and the symptoms usually begin with gradual memory decline followed by a gradual loss of cognitive and motor abilities that will cause difficulties in the daily lives of the patients. Eventually, the patients will lose the ability to take care of themselves and will need to rely on the intensive care provided by others. This has posed significant medical and socioeconomic challenges to the community (Alzheimer's Association, 2013).

Owing to the criticality of this issue, it is vital to diagnose AD accurately, especially at its prodormal stage, i.e., amnestic mild-cognitive impairment (MCI), so that an early treatment can be provided to possibly stop or slow down the progression of the disease. MCI, which is defined as a condition where the patient has noticeable cognitive decline, but without difficulty in carrying out daily activities, has high probability to develop into AD. With the help of emerging neuroimaging technology, the progress and severity of the neurodegeneration associated with AD or MCI can now be diagnosed and monitored in different ways (modalities). Magnetic resonance imaging (MRI) scans, for instance, provide 3D structural information about the brain, where features such as region-of-interest (ROI)-based volumetric measure and the cortical thickness can be extracted from the MRI to quantify brain atrophy that is usually associated with the diseases (Cuingnet et al., 2011; Desikan et al., 2009; Du et al., 2007; Fan et al., 2007b; Gerardin et al., 2009; Klöppel et al., 2008; Oliveira et al., 2010). Flourodeoxyglucose positron emission tomography (FDG-PET), on the other hand, can be used to detect abnormality in term of glucose metabolic rate at brain regions preferentially affected by AD (Chételat et al., 2003, 2005; Foster et al., 2007; Herholz et al., 2002; Higdon et al., 2004). Besides neuroimaging techniques, another line of research uses biological and genetic data to develop potential biomarkers for AD diagnosis. The important measurements in biological and genetic data that are closely related to cognitive decline in AD patients include the increase of cerebrospinal fluid (CSF) total-tau (t-tau) and CSF tau hyperphosphorylated

* Corresponding author.
E-mail addresses: khthung@email.unc.edu (K.-H. Thung), dgshen@med.unc.edu (D. Shen).

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.neuroimage.2014.01.033
1053-8119/

at threonine 181 (p-tau), the decrease of CSF amyloid β ($\boldsymbol{A}\beta$), and the presence of gene apolipoprotein E (APOE) $\epsilon 4$ allele (Fagan et al., 2007; Fjell et al., 2010; Morris et al., 2009).

Although it is common to use information from only one modality such as structural MRI for diagnosis of AD/MCI, complementary information from multiple modalities (Fjell et al., 2010; Walhovd et al., 2010; Landau et al., 2010; Zhang et al., 2011; Liu et al., 2014; Verma et al., 2005; Fan et al., 2007a; Wee et al., 2011, 2012; Li et al., 2012; Zhou et al., 2011) can be combined for more accurate diagnosis. This is supported by the results reported in recent studies (De Leon et al., 2006; Fan et al., 2008; Ye et al., 2008; Hinrichs et al., 2009, 2011; Davatzikos et al., 2011; Zhang and Shen, 2012; Zhang et al., 2011; Liu et al., 2012; Zhang et al., 2012). To support AD research using multi-modality data, Alzheimer's Disease Neuroimaging Initiative (ADNI) has been actively collecting data from multiple modalities (e.g., MRI, FDG-PET and CSF data) from AD, MCI and normal control (NC) subjects yearly or half-yearly. Unfortunately, not all the samples in ADNI dataset are completed with the data from all different modalities. For example, while all the samples in the ADNI baseline dataset contain MRI data, only about half of the samples contain FDG-PET data (which is referred to as PET throughout the manuscript) and another different half of the samples contain CSF data. The "missing" data in the ADNI dataset is due to several reasons, such as, high measurement cost (i.e., PET scans), poor data quality and unwillingness of the patients to receive invasive tests (i.e., collection of CSF samples through lumbar puncture).

There are basically two approaches to deal with missing data in a dataset, i.e., we can either 1) discard the samples with missing data, or 2) impute the missing data. Most existing approaches discard samples with at least one missing modality and perform disease identification based on the remainder of the dataset. However, this approach discards a lot of information that is potentially useful. In fact, in following this approach for multi-modality analysis using MRI, PET and CSF data, about 2/3 of the total samples at ADNI baseline dataset will have to be removed.

The data imputation approach, on the other hand, is more preferable as it provides the possibility to use as many samples as possible in analysis. In fact, incomplete dataset is ubiquitous in many applications and thus various imputation methods have been developed to estimate the missing values based on the available data (Schneider, 2001; Troyanskaya et al., 2001; Zhu et al., 2011). However, these methods work well only when a small portion of the data is missing, but become less effective when a large portion of the data is missing (e.g., PET data in ADNI). Recently, low rank matrix completion (Candès and Recht, 2009) has been proposed to impute missing values in a large matrix through trace norm minimization. This algorithm can effectively recover a large portion of the missing data if the ground truth matrix is low rank and if the missing data are distributed randomly and uniformly (Candès and Recht, 2009). Unfortunately, the latter assumption does not hold in our case since, for each subject, the data from one or more of the modalities might be entirely missing, i.e., the data is missing in blocks.

In this paper, we attempt to identify AD and MCI from the NCs by using incomplete multi-modality dataset from the ADNI database. Denoting the incomplete dataset as a matrix with each row representing a feature vector derived from multi-modality data of a sample, conventional approach for solving this problem is to impute the missing data and build a classifier based on the completed matrix. However, it is too time consuming (as matrix size is large) (Jollois and Nadif, 2007; Xu and Jordan, 1996) and inaccurate (as there are too many missing values) to apply the current imputation methods directly. In addition, the errors introduced during the imputation process may affect the performance of the classifier. In this paper, we largely avert the problems of the conventional approach by proposing a framework (Thung et al., 2013) that 1) shrinks the large incomplete matrix through feature and sample selections, and 2) predicts the output labels directly through matrix completion on the shrunk matrix (i.e., without building another classifier on the completed matrix).

Specifically, we first partition the incomplete dataset into two portions – training set and testing set. Each set is represented by an incomplete feature matrix (each row contains feature vector of a sample), and a corresponding target output matrix (i.e., diagnostic status and clinical scores). Our first goal is to remove redundant/noisy features and samples from the feature matrix so that the imputation problem can be simplified. However, due to the missing values in the feature matrix, feature and sample selections cannot be performed directly. We thus partition the feature matrix, together with the target output matrix, into submatrices with only complete data (Ghannad-Rezaie et al., 2010), so that a 2-step multi-task learning algorithm (Obozinski et al., 2006; Zhang and Shen, 2012) can be applied to these submatrices to obtain a set of discriminative features and samples. The selected features and samples then form a shrunk, but still incomplete, matrix which is more "friendly" to imputation algorithms, as redundant/noisy features and samples have been removed and there are now a smaller number of missing values that need to be imputed. We propose to impute the missing feature data and target outputs simultaneously using a matrix completion approach. Two matrix completion algorithms are explored: low rank matrix completion and expectation maximization (EM). Experimental results demonstrate that our framework yields faster imputation and more accurate prediction of diagnostic labels than the conventional imputation-based classification approach.

In brief, we propose a framework for a solution for this problem – classification using incomplete multi-modality data with large block of missing data. The contributions of our framework are summarized below:

- Feature selection using incomplete matrix (i.e., matrix with missing values) through data grouping and multi-task learning.
- Sample selection using incomplete matrix through data grouping and multi-task learning.
- Improve imputation effectiveness by focusing only on the imputation of important data.
- Improve classification performance by label imputation.

Data

ADNI background

Data used in the preparation of this article were obtained from the ADNI database (`adni.loni.ucla.edu`). ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see `www.adni-info.org`.

Subjects

We only used the baseline data in this study, amounting to a total of 807 subjects (186 AD, 395 MCI and 226 NC). All 807 subjects have MRI scanned, while only 397 subjects have FDG-PET scanned and 406 subjects have CSF sampled. The general inclusion/exclusion criteria used by ADNI are summarized as follow: 1) Normal subjects: Mini-Mental State Examination (MMSE) scores between 24 and 30 (inclusive), a Clinical Dementia Rating (CDR) of 0, non-depressed, non-MCI, and nondemented; 2) MCI subjects: MMSE scores between 24 and 30 (inclusive), a memory complaint, have objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia; 3) mild AD: MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0, and meets NINCDS/ADRDA criteria for probable AD.

Since MMSE and CDR were used as parts of the criteria in categorizing subjects to different disease groups in ADNI dataset, they might provide complementary information in the data imputation process. Thus, in this study, three clinical scores were also included (CDR global, CDR-SB[2] and MMSE) as target outputs in addition to target label. The information of the subjects (i.e., gender, age and education) and clinical scores (i.e., MMSE and CDR global) used in this study are summarized in Table 1.

Data processing

The MRI and PET images were pre-processed to extract ROI-based features. For the processing of MRI images, anterior commissure (AC)–posterior commissure (PC) correction was first applied to all the images using MIPAV software.[3] We then resampled the images to $256 \times 256 \times 256$ resolution and used N3 algorithm (Sled et al., 1998) to correct the intensity inhomogeneity. Next, the skull was stripped using the method described in (Wang et al., 2011, 2014), followed by manual editing and cerebellum removal. We then used FAST (Zhang et al., 2001) in the FSL package[4] to segment the human brain into three different types of tissues: gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF). After registration using HAMMER (Shen and Davatzikos, 2002; Jia et al., 2010; Shen et al., 1999; Shen and Davatzikos, 2004; Tang et al., 2009; Wu et al., 2006; Xue et al., 2006b,a; Yang et al., 2008; Yap et al., 2009; Zacharaki et al., 2008), we obtained the subject-labeled image based on a template with 93 manually labeled region-of-interests (ROIs) (Kabani, 1998). For each subject, we used the volumes of GM tissue of the 93 ROIs, which were normalized by the total intracranial volume (which is estimated by the summation of GM, WM and CSF volumes from all ROIs), as features. For PET image, we first aligned it to its corresponding MRI image of the same subject through affine transformation, and then computed the average intensity of each ROI in the PET image as feature. In addition, five CSF biomarkers were also used in this study, namely amyloid β ($A\beta_{42}$), CSF total tau (t-tau) and tau hyperphosphorylated at threonine 181 (p-tau) and two tau ratios with respective to $A\beta_{42}$ (i.e., t-tau/$A\beta_{42}$ and p-tau/$A\beta_{42}$). As a result, there are a total of 93 features derived from the MRI images, 93 features derived from the PET images and 5 features derived from the CSF biomarkers used in this study. Table 2 summarizes the number of samples and the number of features used in this study for each modality. The numbers under the column "All" represent the number of samples with all the three modalities available.

Classification through matrix shrinkage and completion

Fig. 1 illustrates our framework, which consisted of three components: 1) feature selection, 2) sample selection, and 3) matrix completion.

[2] CDR-SB: CDR Sum of Box, summation of six CDR subscores.
[3] http://mipav.cit.nih.gov/index.php.
[4] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/.

Table 1
Information about ADNI dataset used in this study (Edu.: Education, std: standard deviation).

	Gender		Age (years)	Edu. (years)	MMSE	CDR
Subjects	Male	Female	Mean ± std.	Mean ± std.	Mean ± std.	Mean ± std.
AD	99	87	75.4 ± 7.6	14.7 ± 3.1	23.3 ± 2.0	0.75 ± 0.25
MCI	254	141	74.9 ± 7.3	15.7 ± 3.0	27.0 ± 1.8	0.50 ± 0.03
NC	118	108	76.0 ± 5.0	16.0 ± 2.9	29.1 ± 1.0	0.00 ± 0.00

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ (n samples, d features) and $\mathbf{Y} \in \mathbb{R}^{n \times t}$ (n samples, t targets) denote the feature matrix (that contains features derived from MRI, PET and CSF data) and target matrix (that contains label $[-1\ 1]$ and clinical scores), respectively. As shown in the leftmost diagram in Fig. 1, $\mathbf{X}$ is incomplete, and about half of the subjects do not have PET and CSF data. The dataset is divided into two parts, one for training and one for testing. The target outputs for all the training samples are known, but the target outputs for the testing samples are set to unknown for testing purposes. The input features $\mathbf{X}$ and clinical scores of $\mathbf{Y}$ are first z-normalized across all the samples, by using mean and scale obtained only from the training data. All the missing data are ignored during the normalization process. Then, two stages of multi-task sparse regression are used to remove noisy or redundant features and samples in the training set. The remaining matrix is a matrix with the most discriminative features and samples from the training set. The same set of features selected in the training set are also selected for the testing set. The shrunk training feature matrix together with the testing feature matrix forms a shrunk feature matrix $\mathbf{X_s}$. We then stack $\mathbf{X_s}$ with the corresponding target outputs $\mathbf{Y_s}$ (where the values is unknown for the testing set) to form an incomplete matrix $\mathbf{Z}$. Finally, a matrix completion algorithm (Goldberg et al., 2010; Ma et al., 2011; Schneider, 2001) is applied to $\mathbf{Z}$, so that missing features and the unknown testing target outputs can be predicted simultaneously. The signs of the imputed target labels are then used as the classification output for the testing samples. The following subsections describe the three main components of the framework in more details.

Feature selection

Not all the features are useful in classification. In fact, noisy features may decrease imputation and classification accuracy. In this step, the noisy or redundant features in the incomplete dataset are identified and removed through multi-task sparse regression (with details provided later). However, due to the missing values in the dataset, we cannot apply sparse regression directly to the dataset. We first group the incomplete training set into several overlapping submatrices that comprised samples with complete feature data for different modality combinations, to which sparse regression algorithm can be applied. Some parts of the submatrices are overlapping as we use a grouping strategy that uses the maximum possible numbers of samples and features for each submatrix, so that as much information as possible is used for sparse regression. For example, Table 3 shows the seven possible types of modality combination, denoted as "combination pattern" (CP), for a dataset of 3 modalities, possibly with incomplete data. As shown in Table 3, a sample with lower CP is a "subset" of some higher CPs, where these higher CPs contain modality data that can be grouped

Table 2
Number of subjects (ADNI database at baseline) and number of features used in this study.

	Modalities			
	MRI	PET	CSF	All
Number of features	93	93	5	191
AD subjects	186	93	102	51
MCI subjects	395	203	192	99
NC subjects	226	101	112	52
Total subjects	807	397	406	202

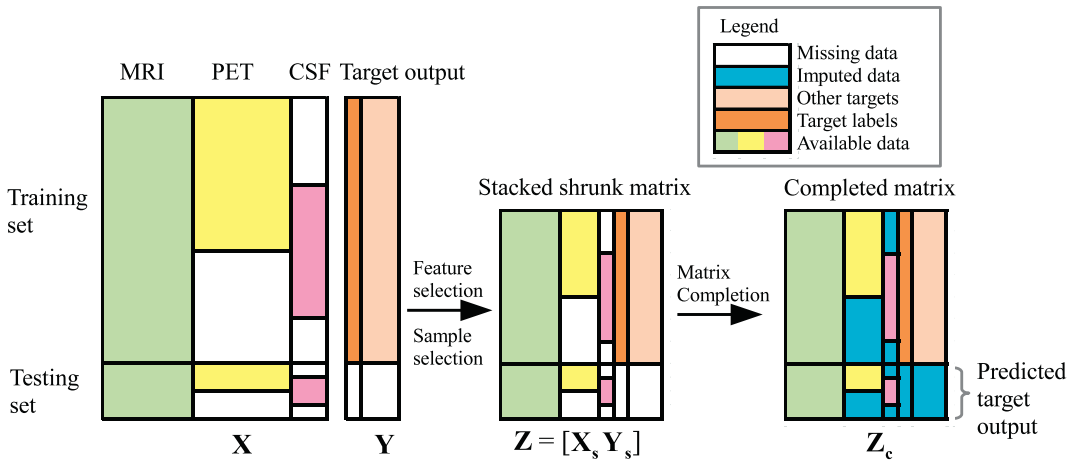

Fig. 1. Classification via matrix shrinkage and matrix completion. There are three main parts in this framework: feature selection, sample selection and matrix completion. Note that feature selection only involves training set. ($\mathbf{X_s}$, $\mathbf{Y_s}$: Shrunk version of $\mathbf{X}$ and $\mathbf{Y}$; $\mathbf{Z_c}$: Completed version of $\mathbf{Z}$.)

with the lower CP to form a submatrix. For instance, the first row of Table 3 indicates that CP1 is "subset" of CP3, CP5 and CP7, as CP1 contains only "Modality 1" data, which is also part of CP3, CP5 and CP7's data. Thus, we can combine "Modality 1" data from CP3, CP5 and CP7 with CP1 to form a submatrix that contains the maximum availability of samples with "Modality 1" data.

For the ADNI dataset used in this study, Modality 1, 2 and 3 are used to denote MRI, PET and CSF, respectively. At ADNI baseline, MRI data is complete while PET and CSF data is incomplete, resulting in four possible types of data combination, i.e., CP1, CP3, CP5 and CP7. Each CP can borrow data from the higher CPs as indicated in the last column of Table 3 to form a submatrix. The graphical illustration of the submatrices is shown in Fig. 2. The red blocks in Fig. 2 mark the four submatrices and their corresponding target outputs. Each submatrix has four interrelated target outputs (i.e. 1 label and 3 clinical scores), which can be learned together using a multi-task learning algorithm, by treating the prediction of each output target as a task. Let $\mathbf{X}_i \in \mathbb{R}^{n_i \times d_i}$ and $\mathbf{Y}_i \in \mathbb{R}^{n_i \times t_i}$ denote the input submatrix and its corresponding output matrix for the i-th multi-task learning in the training set, respectively. Then the multi-task sparse regression of each submatrix is given as

$$\min_{\alpha_i} \frac{1}{2} \|\mathbf{X}_i \alpha_i - \mathbf{Y}_i\|_2^2 + \lambda_f \|\alpha_i\|_{2,1}, \quad (1)$$

where n_i, d_i, t_i and $\alpha_i \in \mathbb{R}^{d_i \times t_i}$ denote the number of samples, the number of features, the number of target outputs and the weight matrix for the i-th multi-task learning, respectively. $\|.\|_{2,1}$ in Eq. (1) is the $l_{2,1}$-norm (group-lasso (Liu et al., 2009; Yuan et al., 2012)) operator which is defined as $\sum_{k=1}^{d_i} \left\| \alpha_i^{(k)} \right\|_2$, where $\boldsymbol{\alpha}_i^{(k)}$ denotes the k-th row of α_i. The use of l_2-norm for $\alpha_i^{(k)}$ forces the weights corresponding to the k-th feature (of $\mathbf{X}_i$) across multiple tasks to be grouped together, while the subsequent use of l_1-norm for $\alpha_i^{(k)}$ forces certain rows of α_i to be all zero. In other words, Eq. (1) tends to select only common features (corresponding to non-zero-valued rows of α_i) for all the prediction tasks. Thus, α_i is a sparse matrix with a significant number of zero-valued rows that correspond to redundant and irrelevant features in each submatrix. In Fig. 2, we arrange α_i according to the feature indices in $\mathbf{X}$, so that the shaded rows in α_i are corresponding to the columns in $\mathbf{X}_i$ (illustrated by red block in the Figure), while the empty rows in α_i are corresponding to the features not included in $\mathbf{X}_i$. In this way, each row of α_i is corresponding to the same feature index in $\mathbf{X}$. The features that are selected for at least one of the submatrices (i.e., rows with at least one non-zero value in $[\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2\ \boldsymbol{\alpha}_3\ \boldsymbol{\alpha}_4]$) are finally used for the training and the testing sets. In this study, we determined α_i for each multi-task learning by using 5-fold cross-validation test based on the accuracy of the label (i.e., first column of $\mathbf{Y}_i$) prediction of the training samples. The training and the testing sets with the selected features are then used in sample selection as described in the following subsection.

Table 3
Grouping of data according to maximum availability of samples for each combination pattern (CP) of modalities. The availability of modalities is represented by binary number at the center column of the Table ('0' denotes 'missing', '1'denotes 'available'), while its decimal equivalent is represented by the CP number on the leftmost column of the Table. Samples with lower CP number can be grouped with the samples with higher CP numbers at the last column of the Table to form a submatrix. In this study, the "Modality 1", "Modality 2" and "Modality 3" represent "MRI", "PET" and "CSF", respectively.

Combination pattern (CP)	Availability of data			Subset of CP
	Modality 1	Modality 2	Modality 3	
1	1	0	0	3, 5, 7
2	0	1	0	3, 6, 7
3	1	1	0	7
4	0	0	1	5, 6, 7
5	1	0	1	7
6	0	1	1	7
7	1	1	1	–

Sample selection

In this step, another multi-task learning is used to select representative samples from the training set that are closely related to the samples in the testing set. This is similar to sparse representation reported by other literatures (Huang and Aviyente, 2006; Wright et al., 2010), a subset of samples is selected to represent a test sample. The only difference here is that we perform sparse representation for a group of testing samples, instead of one testing sample, to 1) select common samples from the training set that well represent the samples in the testing set, and 2) remove unrelated or redundant samples from the training set. The procedure of sample selection is similar to feature selection described previously, with some modifications on the input and output matrices of the multi-task learning.

Let $\mathbf{X_{tr}}$ and $\mathbf{X_{te}}$ respectively denote the shrunk training and testing feature matrices from the previous step that contain only the selected features. $\mathbf{X_{tr}}$ and $\mathbf{X_{te}}$ are first transposed (or rotated by 90°) so that each column of $\mathbf{X_{tr}}^T$ and $\mathbf{X_{te}}^T$ contains features of a sample. Then $\mathbf{X_{tr}}^T$ and $\mathbf{X_{te}}^T$ are used as the input and output to the multi-task learning, where the task is now defined as the prediction of each testing sample from the training samples. If there are no missing values in $\mathbf{X_{tr}}^T$ and $\mathbf{X_{te}}^T$, this multi-task learning will select a set of common samples

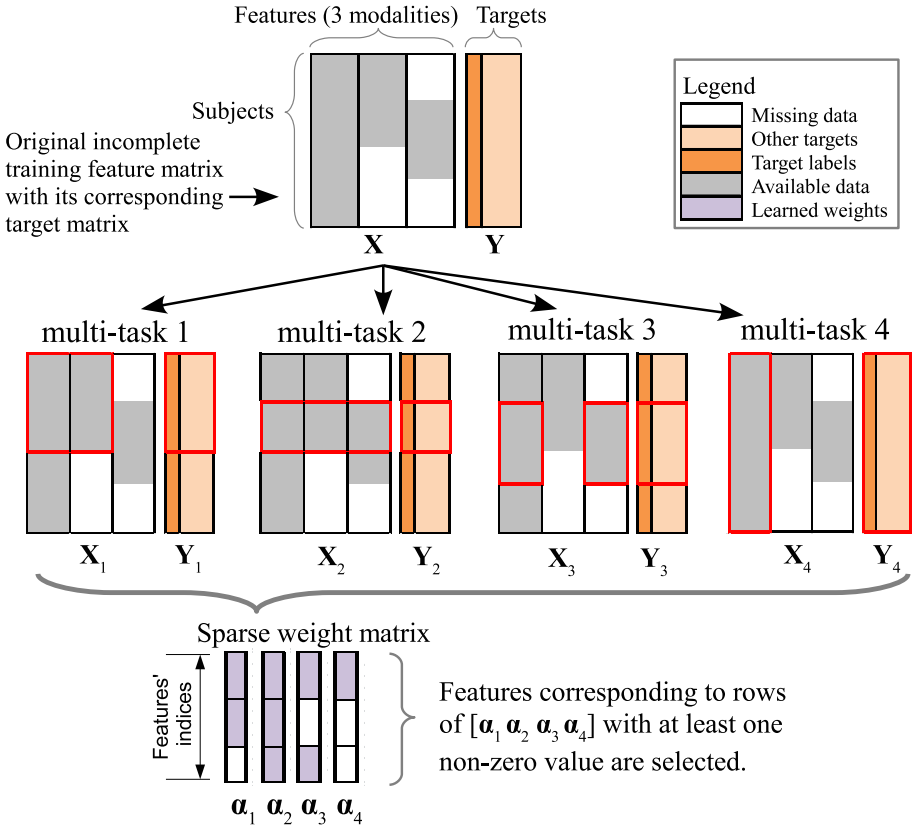

Fig. 2. Feature selection for incomplete multi-modal data matrix with multiple related target outputs by first grouping the data into submatrices and then using multi-task learning on each submatrix to extract common discriminative features. The red boxes come in pairs, which mark the submatrices that are comprised of largest possible number of samples for each pattern of modality combination and their corresponding target outputs.

(analogous to common features in feature selection) in the training set for all the prediction tasks. However, due to the missing values in $\mathbf{X_{tr}}^T$ and $\mathbf{X_{te}}^T$, we cannot perform sample selection directly. Instead, similar to feature selection, we group the input matrix $\left(\mathbf{X_{tr}}^T\right)$ into submatrices that contains complete data for the maximum possible number of samples and features. For each submatrix in $\mathbf{X_{tr}}^T$, all the samples in $\mathbf{X_{te}}^T$ that contain the same set of input features are identified. Each pair of input submatrix and output submatrix with the same features set forms a multi-task learning problem, with its optimization equation given as

$$\min_{\beta_i} \frac{1}{2}\left\|\mathbf{X_{tr_i}}^T\beta_i - \mathbf{X_{te_i}}^T\right\|_2^2 + \lambda_s\|\beta_i\|_{2,1}, \quad (2)$$

where $\mathbf{X_{tr_i}}^T \in \mathbb{R}^{d_i' \times n_{tri}}$, $\mathbf{X_{te_i}}^T \in \mathrm{R}^{d_i' \times n_{tei}}$, $\beta_i \in \mathbb{R}^{n_{tri} \times n_{tei}}$, d_i', n_{tri} and n_{tei} denote the input submatrix, output submatrix, weight matrix, length of the selected features, number of training samples, and number of testing samples of the i-th multi-task learning, respectively.

Fig. 3 summarizes the illustration of the sample selection. Note that the target matrix is incomplete like the input matrix. This causes different number of targets for each multi-task learning, which is reflected by different width of the weight matrix β_i. Due to the use of $\|.\|_{2,1}$ term in Eq. (2), β_i learned is a sparse matrix with some all-zero rows. Training subjects corresponding to all-zero rows of $[\beta_1\ ol\beta_2\ \beta_3\ \beta_4]$ are removed as noisy/irrelevant samples. We assume that removal of noisy or unrelated samples from the training set can consequently improve the accuracy of the missing values imputation, and thus the classification performance. To justify this assumption, we have included a simulation test on our proposed sample selection algorithm using synthetic data in Appendix A.

Matrix completion as classification

The original incomplete matrix is shrunk significantly after the feature and sample selection steps. Let $\mathbf{X_s}$ and $\mathbf{Y_s}$ denote the shrunk version of matrix $\mathbf{X}$ and $\mathbf{Y}$, respectively, while n_s and d_s denote the number of remaining samples and data features, respectively. The stacked matrix $\mathbf{Z} = [\mathbf{X_s}\mathbf{Y_s}] \in \mathbb{R}^{n_s \times (d_s + t)}$ still contains some missing values, including the target outputs of the test set which are to be estimated. The objective of this step is to impute the missing input features, missing target labels, and missing clinical scores simultaneously. Two imputation methods are tested for this step, namely the modified Fixed-point Continuation (FPC) algorithm (Goldberg et al., 2010; Ma et al., 2011) and the regularized expectation maximization (EM) algorithm (Schneider, 2001).

Modified FPC (mFPC)

The multi-task regressions used in the features and samples selection steps have selected the most discriminative input features for the (training) target outputs and the most representative training samples for the testing samples, respectively. As a consequence, the columns of target outputs ($\mathbf{Y_s}$) of the stacked matrix $\mathbf{Z}$ could be linearly represented by the columns of data features ($\mathbf{X_s}$); while the rows of the testing samples in $\mathbf{Z}$ could be linearlyrepresented by the rows of the training samples. The matrix $\mathbf{Z}$ is thus probably low rank (as some rows could be represented by other rows, etc.). However, in practice, measurements in $\mathbf{X_s}$ and $\mathbf{Y_s}$ could contain certain level of noises. Therefore, the incomplete $\mathbf{Z}$ can be completed using trace norm minimization (low trace norm is often used to approximate low rank assumption), together with two regularization terms (i.e., the second and third term in Eq. (3)) to penalize the noises in $\mathbf{X_s}$ and $\mathbf{Y}_s$ labels ($\mathbf{P}$), and 2) the rest of the data ($\mathbf{Q}$). The regularization terms are changed accordingly to have one logistic

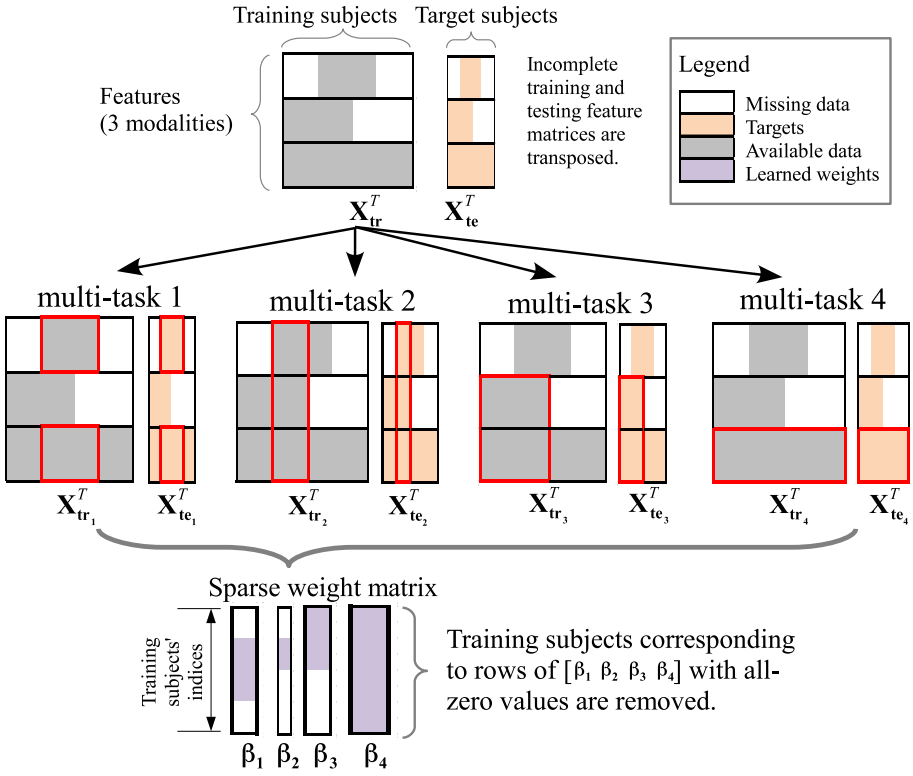

Fig. 3. Sample selection. In this study, sample selection is realized by modifying the input and output matrices in feature selection illustrated in Fig. 2. Specifically, we transpose the training and testing feature matrices, and use the transposed training and testing feature matrices as the input and target output of the multi-task learning, respectively.

loss function ($L_p(u,v) = log(1 + exp(-uv))$) for the output labels (as the output labels can only take value 1 or −1), and one square loss function ($L_q(u,v) = 1/2(u - v)^2$) for the rest of the data (as other data can take any value). The imputation optimization problem is thus given as:

$$\arg\min_{\mathbf{Z}} \mu\|\mathbf{Z}\|_* + \frac{\lambda_m}{|\Omega_\mathbf{P}|}\sum_{(i,j)\in\Omega_\mathbf{P}} L_p\left(z_{ij}, p_{ij}\right) + \frac{1}{\left|\Omega_\mathbf{Q}\right|}\sum_{(i,j)\in\Omega_\mathbf{Q}} L_q\left(z_{ij}, q_{ij}\right), \tag{3}$$

where $\Omega_\mathbf{P}$ and $\Omega_\mathbf{Q}$ denote the set of observed (i.e., non-missing) labels in $\mathbf{Y_s}$ and the set of observed values for the rest of the data, respectively; $|.|$ denotes an operator for the number of elements; $\|.\|_*$ denotes an operator for the trace norm; and z_{ij}, p_{ij} and q_{ij} are the predicted observed values, observed target labels and other observed data, respectively. λ_m and μ are the positive parameters used to control the focus of the minimization problem in Eq. (3). If λ_m is high, Eq. (3) will focus on minimizing the L_p term (second term); if μ is high, Eq. (3) will focus on minimizing the trace norm term (i.e., stronger low rank assumption), and vice versa.

This optimization problem is solved by using the modified FPC algorithm (Goldberg et al., 2010), which consists of two alternating steps for each iteration k:

1. Gradient step:

$$\mathbf{A}^k = \mathbf{Z}^k - \tau g\left(\mathbf{Z}^k\right) \tag{4}$$

where τ is the step size and $g(\mathbf{Z}^k)$ is the matrix gradient which is defined as:

$$g\left(z_{ij}\right) = \begin{cases} \dfrac{\lambda_m}{|\Omega_\mathbf{P}|}\dfrac{-p_{ij}}{1+exp\left(p_{ij}z_{ij}\right)}, & (i,j)\in\Omega_\mathbf{P} \\ \dfrac{1}{\left|\Omega_\mathbf{Q}\right|}\left(z_{ij}-q_{ij}\right), & (i,j)\in\Omega_\mathbf{Q} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

2. Shrinkage step:

$$\mathbf{Z}^{k+1} = S_{\tau\mu}\left(\mathbf{A}^k\right) \tag{6}$$

where $S(\cdot)$ is the matrix shrinkage operator. If SVD of $\mathbf{A}^k$ is given as $\mathbf{U\Lambda V}^T$, then the shrinkage operator is given as:

$$S_{\tau\mu}\left(\mathbf{A}^k\right) = \mathbf{U}\max(\Lambda-\tau\mu, 0)\mathbf{V}^T \tag{7}$$

where $\max(\cdot)$ is the elementwise maximum operator.

These two steps are iterated until convergence where the objective function in Eq. (3) at k-th iteration is stable.

Regularized EM (rEM)

We also use the regularized EM (rEM) algorithm developed in (Schneider, 2001) to impute missing values. Symbols defined in this subsection should not be confused with the symbols used in other sections. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be an incomplete matrix with n number of samples and d number of variables, its mean vector $\mu \in \mathbb{R}^{1\times d}$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ are to be estimated. For a given sample $\mathbf{x} \in \mathbb{R}^{1\times d}$ with missing values, let $\mathbf{x_m}\in\mathbb{R}^{1\times d_m}$ and $\mathbf{x_a}\in\mathbb{R}^{1\times d_a}$ denote the parts of vector $\mathbf{x}$ containing variables with missing values and available values, respectively. Then $\mathbf{x_m}$ can be estimated through linear regression model below

$$\mathbf{x_m} = \boldsymbol{\mu_m} + (\mathbf{x_a} - \boldsymbol{\mu_a})\mathbf{B} + \mathbf{e} \tag{8}$$

where $\boldsymbol{\mu_m}\in\mathbb{R}^{1\times d_m}$ and $\boldsymbol{\mu_a}\in\mathbb{R}^{1\times d_a}$ represent the portions of $\boldsymbol{\mu}$ that corresponding to $\mathbf{x_m}$ and $\mathbf{x_a}$, respectively, while $\mathbf{B}\in\mathbb{R}^{d_a\times d_m}$ and $\mathbf{e}\in\mathbb{R}^{1\times d_m}$ are the regression coefficient matrix and random residual vector (with zero mean and unknown covariance matrix $\mathbf{C}\in\mathbb{R}^{d_m\times d_m}$), respectively. We are now ready to describe the imputation using EM algorithm, which is an iterative process that consists of three steps, 1) expectation step: the mean μ and covariance matrix Σ is estimated, 2) maximization

step: the conditional maximum likelihood estimate (MLE) of the parameters of the regression model (i.e., **B** and **C**) is computed, based on the expected value of μ and Σ, and, 3) imputation step: the missing values is estimated using Eq. (8) based on the computed parameters. After missing values are imputed, it will iterate back to step 1, where a new set of μ and Σ is estimated based on the completed **x**, and the whole process is repeated until a convergence condition is met (i.e., the estimated μ and Σ become stable). Regularized EM algorithm consists of the same steps as EM algorithm, with a modification on the maximization step, where the regression coefficients in **B** are computed through ridge regression method (Hoerl and Kennard, 1970). For more detailed information about rEM algorithm, interested reader may refer to (Schneider, 2001). In our framework, rEM is used to estimate unknown target outputs and missing input features in the matrix completion step.

Results and discussions

The proposed framework was tested by using the ADNI multimodality dataset, which includes MRI, PET and CSF data. In this section, the proposed framework is first compared with the baseline frameworks which will be defined in the following subsection. Then, the proposed framework is compared with two state-of-the-art methods (i.e., incomplete Multi-Source Feature (iMSF) learning method and Ingalhalikar's algorithm for classification based on incomplete dataset) and also a unimodal classifier using only MRI features. In addition, we evaluate the effect of parameters selection (i.e., λ_s, λ_m and μ) of the proposed framework on the classification performance. Finally, we also identify the features that are always being selected in this study.

The classification performance of all the compared methods is evaluated by using a 10-fold cross-validation scheme. For each fold, another 5-fold cross-validation scheme is applied on the training dataset to select the best parameters for multi-task learning in feature selection and also for sparse regression based classifier in the baseline methods. The multi-task learning in feature selection and sample selection is realized by using matlab function *mcLeastR* from SLEP.[5] SLEP is a powerful sparse learning package where it achieves fast convergence in computation by using Nesterov's method (Liu et al., 2009; Nesterov, 1983) to solve smooth reformulation of the problem and accelerated gradient method (Liu and Ye, 2010; Nesterov, 2007) to solve regularized non-smooth optimization problem. There are infinite choices for λ_f (i.e., multi-task learning parameter in feature selection). Fortunately for the solver *mcLeastR* that we used, it automatically computes the maximum λ_{max} value for our problem. Thus, each λ_f value that we input to this solver is treated as a fraction to λ_{max}, e.g., the true regularization parameter used for $\lambda_f = 0.1$ is actually $0.1 \times \lambda_{max}$. Therefore, we choose parameter λ_f from these candidate values: {0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}, which roughly cover the whole range of possible λ_f values. The λ_f value used for each fold of experiment is determined based on the highest accuracy of regressed **Y** label of the training data through 5-fold cross-validation test on the training data. As a result, λ_f is different for each fold of experiment, i.e., different data sparsity for each fold of experiment is assumed. For sample selection, we fix a small value for λ_s, aiming to only remove unrelated samples from the training set. For mFPC matrix completion algorithm, we use grid search to select values of its parameters (i.e., μ and λ_m), i.e., fixed value of μ and λ_m are used for all the folds based on the best classification result in grid search.

Four classification performance measures are used in this study, namely 1) accuracy: the number of correctly classified samples divided by the total number of samples; 2) sensitivity: the number of correctly classified positive samples divided by the total number of positive samples; 3) specificity: the number of correctly classified negative samples divided by the total number of negative samples; and 4) area under receiver operating characteristic (ROC) curve (AUC). The positive samples are referred to AD in AD/NC classification and MCI in MCI/NC classification, respectively.

Comparison with baseline frameworks

Four imputation methods are included in the baseline framework for comparison in this study:

1. Zero imputation. In this method, the missing portion of the input data matrix is filled with zero. Since all the features were z-normalized (i.e., with zero mean and unit standard deviation) before the imputation process, "zero imputation" is equivalent to fill the missing feature values with the average observed feature values (i.e., all the missing values in a column of data matrix are filled with the mean of the observed values in the same column).
2. *k*-nearest neighbor (KNN) imputation (Speed, 2003; Troyanskaya et al., 2001). The missing values are filled with a weighted mean of the *k* nearest-neighbor rows. The weights are inversely proportional to the Euclidean distances from the neighboring rows. We set $k = 20$ after some empirical tests.
3. Regularized expectation maximization (rEM) (Schneider, 2001). Details are as described in the previous section. We used the default parameter values for the rEM code downloaded from `http://www.clidyn.ethz.ch/imputation/index.html`.
4. Fixed-point continuation (FPC) (Ma et al., 2011). FPC is one of the low rank matrix completion method that uses the fixed point and Bregman iterative algorithms. It is the original version of Eq. (3) with the regularization terms L_p and L_q replaced by a square loss function for all the observed data. The matlab code for FPC is included in the singular value thresholding (SVT) package.[6] The parameter value of FPC, i.e., μ, is determined empirically.

These imputation methods are used in two baseline frameworks for comparisons:

1. Baseline 1: Conventional method. Impute the incomplete data matrix and then train a classifier using the completed training set data.
2. Baseline 2: Use the proposed feature and sample selection method to shrink the incomplete dataset, impute the missing features in the shrunk incomplete feature matrix, and then train a classifier based on the completed shrunk training set data.

The only difference between the two baseline frameworks above, is that the first baseline framework imputes missing values on the original feature matrix, while the second baseline framework imputes missing values on the shrunk feature matrix. We use sparse regression classifier for the two baseline frameworks, its formulation is given as:

$$\min_{\alpha} \frac{1}{2}\|\mathbf{X}\alpha - \mathbf{Y}\|_2^2 + \lambda\|\alpha\|_{2,1}, \tag{9}$$

where **X**, **Y**, and α are defined as the input feature matrix, the output target matrix (including class labels and clinical scores), and the weight matrix, respectively. We obtain α based on the completed training set and multiply it with the completed feature matrix from the testing set to produce regressed outputs. The sign of the regressed output of a testing sample that corresponds to the class label is used as the predicted class label. There is one regularization parameter in Eq. (9), i.e., λ, which is always positive and is primarily used to control features sparsity in **X**. We determine the value of λ by performing a 5-fold cross-validation test based on the completed training dataset.

Table 4 summarizes the AD/NC classification performance of all the frameworks in comparison. Results reported are the average

[5] `http://www.public.asu.edu/jye02/Software/SLEP/index.htm`.

[6] `http://svt.stanford.edu/code.html`.

Table 4
AD/NC classification performance for all the frameworks in comparison. Total of 412 samples used. The best value for each performance measure is highlighted in bold (Acc: Accuracy, Sen: Sensitivity, Spe: Specificity, AUC: area under ROC curve, time: the average imputation time for each fold, rEM: Regularized EM, mFPC: modified FPC).

Framework	Imputation	Acc.	Sen.	Spe.	AUC	Time(s)
Baseline 1	Zero	0.802	0.802	0.804	0.887	**0.00**
	KNN	0.830	0.801	0.859	0.904	1.76
	rEM	0.816	0.792	0.840	0.892	84.36
	FPC	0.821	0.812	0.831	0.900	115.35
Baseline 2	Zero	0.853	0.843	0.864	0.922	**0.00**
	KNN	0.868	0.836	0.894	0.927	0.29
	rEM	0.857	0.833	0.879	0.922	23.76
	FPC	0.858	0.843	0.872	0.923	15.31
Proposed	KNN	0.850	0.745	**0.936**	0.914	0.32
	rEM	**0.885**	0.837	0.927	0.944	24.20
	mFPC	0.880	**0.852**	0.904	**0.947**	0.39

Table 5
MCI/NC classification performance for all the compared frameworks. Total of 621 samples used. The best value for each performance measure is highlighted in bold. (Please refer to Table 4 for the meaning of the abbreviations used in this table.)

Framework	Imputation	Acc.	Sen.	Spe.	AUC	Time(s)
Baseline 1	Zero	0.639	0.598	0.710	0.695	**0.00**
	KNN	0.635	0.623	0.657	0.687	2.72
	rEM	0.650	0.636	0.675	0.696	139.14
	FPC	0.643	0.602	0.714	0.686	130.20
Baseline 2	Zero	0.669	0.631	0.736	0.732	**0.00**
	KNN	0.666	0.628	0.733	0.724	0.61
	rEM	0.673	0.639	0.733	0.734	34.41
	FPC	0.670	0.632	**0.737**	0.736	19.06
Proposed	KNN	0.672	0.778	0.486	0.726	0.65
	rEM	0.701	**0.866**	0.414	**0.774**	36.56
	mFPC	**0.715**	0.753	0.649	0.773	1.21

measurements of 10 repetitions of 10-fold cross-validation test. As shown in Table 4, all performance of baseline 1 frameworks are improved in baseline 2 framework (i.e., from 0.80–0.83 to 0.85–0.87).

In fact, all the four performance measures (i.e., accuracy, sensitivity, specificity and AUC) increase after applying the proposed feature and sample selection steps before the imputation in baseline 2 framework. In addition, the average imputation time is significantly reduced, as shown in the last column of Table 4. For example, FPC and rEM respectively complete the imputation with 8 times and 4 times faster in the baseline 2 framework, if compared with the baseline 1 framework. We thus have shown the efficacy of the proposed feature and sample selection methods in removing the unrelated samples and noisy features, which is beneficial to the imputation process, both in terms of accuracy and speed. In addition, the classification performance is further improved to 0.88–0.89 if the target labels are imputed simultaneously with the incomplete data features using the modified FPC (mFPC) and rEM methods. Although the classification performances of both mFPC and rEM are similar, mFPC performs significantly better than rEM in terms of computation speed. Similar findings are observed for MCI/NC classification as shown in Table 5.

Comparison with non-imputation state-of-the-art methods

Recently, several algorithms have been proposed to deal with incomplete dataset where the data is missing in blocks. We compare our proposed framework with these methods, which are briefly described in the following:

1. Incomplete Multi-source Feature learning (iMSF)[7] (Xiang et al., 2013; Yuan et al., 2012). The iMSF predicts the target output labels of the incomplete multiple heterogeneous data without involving data imputation. This is a multi-task learning algorithm that is able to deal with missing feature values. The iMSF is available in two versions for multi-task learning part, i.e., the logistic version and the regression version, along with one regularization parameter. We test both versions of iMSF with a range of regularization parameters (i.e., {0.005, 0.01, 0.05, 0.1, 0.2, 0.3 and 0.4}) and finally choose the one with the highest classification accuracy for comparison.
2. Ingalhalikar's algorithm (Ingalhalikar et al., 2012). This algorithm uses an ensemble classification technique to fuse decision results from multiple classifiers constructed from subsets of data. The data subsets are obtained by applying a grouping strategy similar to ours. We implemented Ingalhalikar's algorithm and tested it on our dataset. Specifically, we group the data into subsets, select features using signal-to-noise ratio coefficient filter (Guyon and Elisseeff, 2003), use linear discriminant analysis (LDA) as classifier, and finally fuse all the classification results of the subsets into a single result for each sample. We used two fusion methods for this algorithm, i.e., 1) weighted average: each classifier is assigned a weight based on its training classification error, 2) average: all the classifiers are assigned with equal weight.

Tables 6 and 7 show the comparison of classification performance between the proposed framework (using rEM and mFPC imputation methods) and the iMSF and Ingalhalikar's algorithm. Both tables show that the proposed framework outperforms the Ingalhalikar's algorithm but performs competitively to iMSF.

iMSF-regression has the highest sensitivity for AD/NC classification and has the highest specificity for MCI/NC classification. iMSF-logistic performs well in MCI/NC classification, maybe because there is non-linear relationship between the features and MCI, which can be better captured by logistic loss function. However, iMSF-logistic does not perform as well in AD/NC classification, if compared with iMSF-regression and our proposed methods. In addition, both versions of iMSF have lower AUC for both categories of classification, if compared with our proposed methods.

Ingalhalikar's algorithm has the lowest performance in this study if compared with iMSF and our proposed method. The proposed framework, though not involving ensemble procedure, is competitive with state-of-the-art algorithm.

The proposed framework performs the best in term of classification accuracy and AUC values. In term of classification accuracy, the proposed framework using rEM performs the best in AD/NC classification while the proposed framework using mFPC performs the best in MCI/NC classification. Though the performance difference of the proposed framework and iMSF is small in term of classification accuracy (about 1%), there is a substantially significant difference in term of AUC, which is not sensitive to threshold. Both mFPC and rEM imputation algorithms achieve the highest AUC values for both AD/NC and MCI/NC classifications, which are the most important measure in classification.

We performed additional *t*-tests to examine the significance of our results. We picked AUC values for the *t*-test, as AUC values are not sensitive to threshold. All the AUC values obtained from the 10 repetitions of the 10-fold cross-validation are used for comparisons, i.e., 100 AUC values from the proposed methods, versus 100 AUC values from the methods of comparison. The null hypothesis is that both methods have no significant difference in term of AUC values, while the alternative hypothesis is there is significant difference in term of AUC values obtained by the two methods at 95% confidence level. We show the *p*-values of the *t*-test at the last two columns of Tables 6 and 7. The *p*-values that are marked with * indicates that the differences are significant at 95% confidence level.

Table 6 shows that our proposed framework using rEM and mFPC perform statistically significantly better than all the methods in comparison in both the AD/NC and MCI/NC classifications, in term of AUC values.

[7] http://www.public.asu.edu/~jye02/Software/MALSAR/.

Table 6
AD/NC classification comparison with iMSF and Ingalhalikar's algorithm. All the performance measures reported in this table are the means of the respective 10 repetitions of the 10-fold cross-validation test. The best value for each performance measure is highlighted in bold. The significance of the results is indicated by *p*-value of the *t*-test (using 100 pairs of AUC values of two methods in comparison) at the last two columns of the Table. *p*-values that are marked with * indicate that the proposed method is statistically better than the method in comparison at 95% confidence level (fusion1: weighted average of all the classification outputs from the data subsets, fusion2: average of all the classification outputs from the data subsets).

Framework	Methods	Performance measures				*t*-test on AUC (*p*-value)	
		Acc	Sen	Spe	AUC	with rEM	with mFPC
Proposed	rEM	0.885 ± 0.050	0.837	**0.927**	0.944		
	mFPC	0.880 ± 0.054	0.852	0.904	**0.947**		
iMSF	regression	0.873 ± 0.056	**0.861**	0.883	0.932	<0.0005*	<0.0005*
	logistic	0.866 ± 0.055	0.809	0.912	0.924	<0.0005*	<0.0005*
Ingalhalikar's	fusion1	0.843 ± 0.061	0.804	0.877	0.905	<0.0005*	<0.0005*
	fusion2	0.847 ± 0.058	0.809	0.880	0.913	<0.0005*	<0.0005*

Table 7
MCI/NC classification comparison with iMSF and Ingalhalikar's algorithm. (Please refer to Table 6 for detailed descriptions of this Table.)

Framework	Methods	Performance measures				*t*-test on AUC (*p*-value)	
		Acc	Sen	Spe	AUC	with rEM	with mFPC
Proposed	rEM	0.701 ± 0.047	**0.866**	0.414	**0.774**		
	mFPC	0.715 ± 0.056	0.753	0.649	0.773		
iMSF	Regression	0.692 ± 0.063	0.649	**0.768**	0.760	0.001	0.001
	Logistic	0.706 ± 0.051	0.818	0.509	0.733	<0.0005*	<0.0005*
Ingalhalikar's	fusion1	0.642 ± 0.062	0.644	0.639	0.664	<0.0005*	<0.0005*
	fusion2	0.649 ± 0.063	0.651	0.643	0.689	<0.0005*	<0.0005*

Comparison with unimodal classifier using MRI data

We also compare the performance of the proposed framework with a unimodal classifier using only MRI data, as shown in Table 8. Since all the samples have MRI data, the number of samples used is the same as the previous experiment. The same sparse regression classifier in Eq. (9) is used in this test. Superior performance of the proposed framework demonstrates the importance of including information from other modalities to improve disease diagnosis accuracy.

Effect of parameters selection of mFPC

It is important to select a set of robust parameters for matrix completion, so that the proposed framework works well for most of the situations. Fig. 4 shows the classification accuracies and AUC of the proposed mFPC-based framework for a range of λ_m and μ values. As shown in the figure, the classification accuracy is consistently high when small μ and large λ_m are used. With small μ and large λ_m, the objective function in Eq. (3) will focus on the minimization of logistic function (i.e., target label prediction) instead of the minimization of the tracenorm (i.e., low rank matrix completion). This implies that the incomplete matrix Z is completed using higher rank than expected. This is probably due to the measurement noise in the dataset, which causes an increase in the rank of **Z**. Based on the plot in Fig. 4, Eq. (3) that satisfy $\mu \leq 10^{-3}$ and $\lambda_m \geq 0.05$ yield reasonably good label prediction.

Table 8
Classification comparison with unimodal classifier using only MRI features for AD/NC and MCI/NC classifications. The best value for each performance measure is highlighted in bold.

Methods	AD/NC				MCI/NC			
	Acc	Sen	Spe	AUC	Acc	Sen	Spe	AUC
Proposed rEM	**0.885**	0.837	**0.927**	0.944	0.701	**0.866**	0.414	**0.774**
Proposed mFPC	0.880	**0.852**	0.904	**0.947**	**0.715**	0.753	0.649	0.773
MRI only	0.834	0.821	0.847	0.902	0.625	0.582	**0.700**	0.693

Effect of λ_s on sample selection

Fig. 5 shows the effect of λ_s on sample selection in Eq. (2) to the average number of samples selected (from the training dataset) and the average classification accuracy (i.e., accuracy of the label imputation) of the matrix completion. As shown in Fig. 5, the average number of samples selected reduces gradually when λ_s is decreasing, while relatively consistent in terms of classification accuracy for mFPC. This implies that there are a lot of redundant samples in the training set, which can be removed without significantly affecting the accuracy of the label imputation. To examine the performance of sample selection using synthetic data, please refer to Appendix A.

One of the possible limitations of the proposed sample selection is that the output space is not considered in the algorithm (as this information is not available for the testing samples), which might cause possible bias in the result if there is measurement noise in the output space. For example, the feature space for highly coherent samples is very similar, but due to measurement noise in the output space, they may have different outputs. In worst case scenario (e.g., using too large λ_s value), the l_1-regularized algorithms (i.e., the l_1-norm part of the $l_{2,1}$-norm) may select only one sample and discard the others, which cause bias in the result. This problem can be ameliorated by including the additional l_2-regularization, such as that done in Elastic Net (Zou and Hastie, 2005). This will help retain some coherent samples and allow some averaging effect. Another possible solution is to perform sample selection and output variable estimation iteratively, which we leave it as our future work.

Most discriminative features

Table 9 shows the statistics of the features selected by the proposed feature selection method for the incomplete ADNI data, during the AD/NC and MCI/NC classifications, respectively. On average, more than 60% of the features is removed for both cases. The number of features selected for each fold varied significantly, e.g., it can go as low as 45 or as high as 120 for AD/NC classification. This is probably because the regularization parameter λ_f in Eq. (1) is chosen from a wide range of values (i.e., {0.001, …, 0.9}).

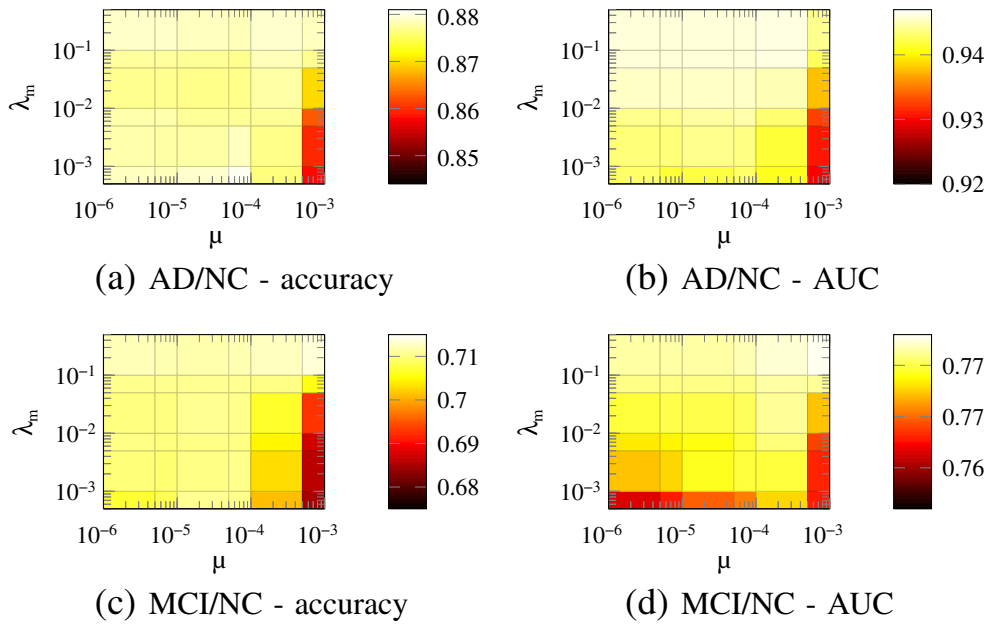

Fig. 4. Effect of the parameter changes in mFPC algorithm to AD/NC and MCI/NC classifications in terms of accuracy and AUC.

In addition, we also include the distribution of the most discriminative features according to modalities in Table 9. We define the most discriminative features (MDFs) as the features that were selected for more than 90% of the times, i.e., more than 90 times in the 10 repetitions of the 10-fold cross-validation run. Most of the MDFs are located in MRI modality, for both the AD/NC and MCI/NC classifications. We also observed that more features were selected for AD/NC than MCI/NC classification. This is probably because MCI, which is the early stage of AD, affects less brain regions (or ROIs) if compared with AD, where its abnormalities are widely spread across brain regions.

Table 10 shows the names of the MDFs for each modality. The common MDFs selected for AD/NC and MCI/NC classifications are also included in Table 10, if exist. The common MDFs for MRI modality include hippocampal formation, middle temporal gyrus, uncus, and amygdala. The atrophy at these ROIs has been reported to be associated with memory and cognitive impairments or closely related to the AD/MCI pathology (Convit et al., 2000; De Leon et al., 1997; Poulin et al., 2011; Yang et al., 2012). For AD/NC classification, since there are many MDFs from MRI, we only list the MDFs that were selected in all cross-validations and repetitions in Table 10.

On the other hand, the common MDFs for FDG-PET modality include middle frontal gyrus and precuneus, which are similar to the findings in (Mielke et al., 1998; Scarmeas et al., 2004). For CSF biomarkers, the selected MDFs were t-tau/$A\beta_{42}$ for AD/NC classification and $A\beta_{42}$ and t-tau for MCI/NC classification.

Figs. 6 and 7 graphically show the locations of the selected ROI-based features (for MRI and PET modalities) for both the AD/NC and MCI/NC classifications, respectively.

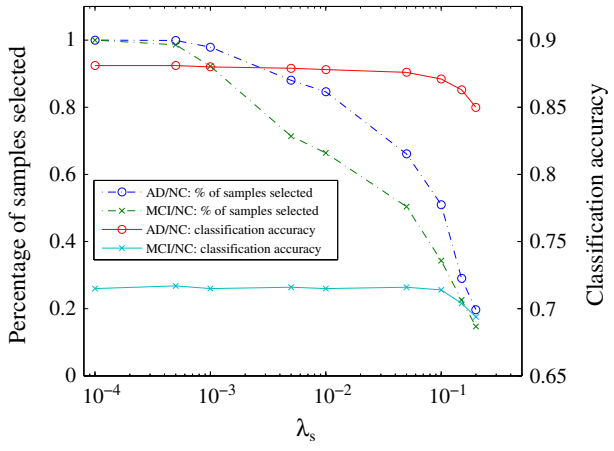

Fig. 5. Effect of the parameter λ_s on the number of samples selected and the corresponding classification accuracies for AD/NC and MCI/NC classification using mFPC of the proposed framework.

Conclusion

In this work, we propose a novel classification framework that is able to deal with datasets with significant amount of missing data (e.g., data missing in blocks). Conventional imputation-based classification approach is slow and inaccurate for this type of dataset. We accomplish accurate label prediction by applying matrix completion on a shrunk version of the data matrix. The matrix shrinkage operation simplifies the imputation task since redundant features and samples have been removed and less missing data needs to be imputed. The experimental results demonstrate the efficacy of feature selection and sample selection in improving the classification performance of the conventional imputation-based classification method, both in terms of speed and accuracy. The proposed framework also yields competitive performance, compared with the state-of-the-art methods such as iMSF and Ingalhalikar's algorithm. Based on the *t*-test of their AUC values, the proposed framework using rEM and mFPC are statistically significantly better than iMSF and Ingalhalikar's algorithm in AD/NC and MCI/NC classifications.

Acknowledgment

This work was supported in part by NIH grants AG041721, AG042599, EB006733, EB008374, and EB009634. This work was also partially supported by the National Research Foundation grant (No. 2012-005741) funded by the Korean government.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) National Institutes

Table 9
Numbers of features selected for AD/NC and MCI/NC classifications. (std.: standard deviation; MDFs: Most discriminative features, features that were selected more than 90% of the time for 10 repetitions of 10-fold cross-validation run.)

Classification	Original no. of features	No. of selected features			MDFs			
		Mean ± std.	Min	Max	MRI	PET	CSF	Total
AD/NC	191	69.9 ± 14.4	45	120	30	5	1	36
MCI/NC	191	63.0 ± 26.7	19	140	10	5	2	17

Table 10
Most discriminative features (MDFs) selected for each modality. (Please refer to Table 9 for definition of MDF. For MRI's MDFs in AD/NC classification, only those that are selected 100% of the time are listed here.)

Modality	AD/NC	MCI/NC
MRI	Common MDFs: Hippocampal formation right, hippocampal formation left, middle temporal gyrus left, uncus left, amygdala right.	
	Medial frontal gyrus, Angular gyrus right, precuneus right, superior parietal lobule left, precentral gyrus left, perirhinal cortex left, lateral occipitotemporal gyrus right, amygdala left, middle temporal gyrus right, corpus callosum, inferior temporal gyrus right, lateral occipitotemporal gyrus left	Entorhinal cortex left, cuneus left, lingual gyrus left, temporal pole left, middle occipital gyrus left
PET	Common MDFs: Middle frontal gyrus right, precuneus right, precuneus left, Medial front–orbital gyrus right	
	Insula right	Angular gyrus left
CSF	t-tau/$A\beta_{42}$	$A\beta_{42}$ and t-tau.

of Health grant U01 AG024904. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Amorfix, Bayer Schering Pharma AG, Bioclinica Inc., Biogen Idec, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, Innogenetics, IXICO, Janssen Alzheimer Immunotherapy, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Meso Scale Diagnostic, & LLC, Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Servier, Synarc, Inc., and Takeda Pharmaceuticals, as well as non-profit partners the Alzheimer's

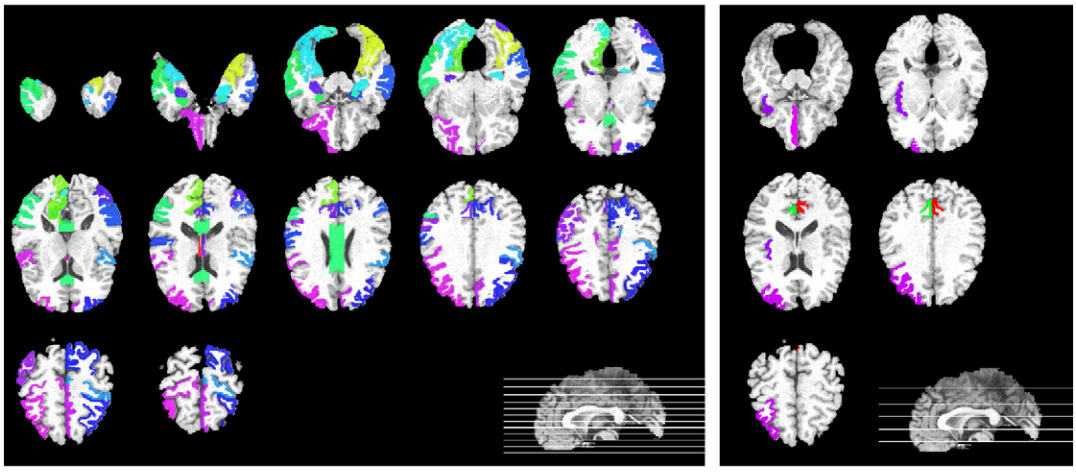

Fig. 6. MDFs in AD/NC classification (Left: MRI, right: PET).

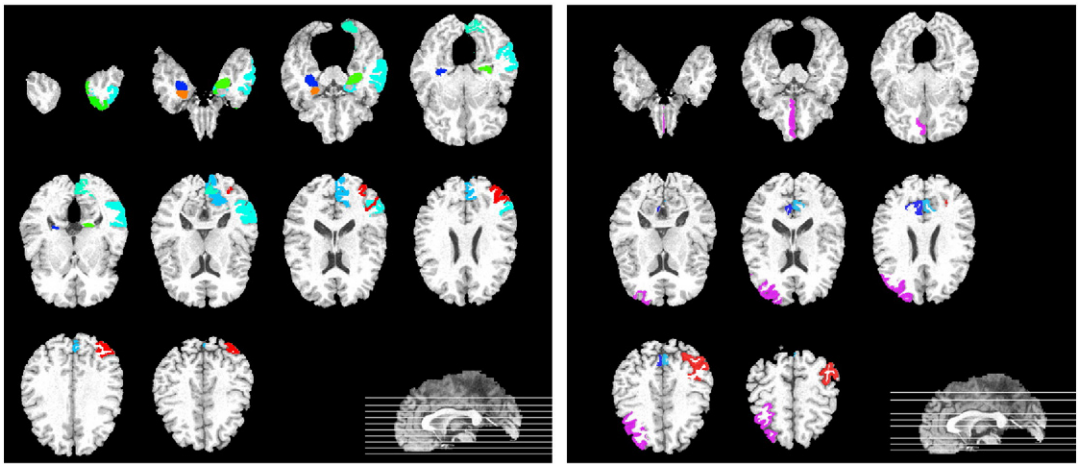

Fig. 7. MDFs in MCI/NC classification (Left: MRI, right: PET).

Table A.11
Synthetic data: One homogeneous data matrix and two inhomogeneous data matrices. $\mathbf{X}_m, m = \{1,2,3\}$ is a simulated two-modality data matrix, e.g., $[\mathbf{X}_{r1}\ \mathbf{X}_{r2}]$, with $n_{s\,m}$ number of samples and rank (r_1, r_2), where r_1 and r_2 are the ranks for $\mathbf{X}_{r1}$ and $\mathbf{X}_{r2}$, respectively. The inhomogeneous data matrix 1 is simulated by stacking $\mathbf{X}_1$ and $\mathbf{X}_2$, while the inhomogeneous data matrix 2 is simulated by stacking $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ data.

Data matrices	$\mathbf{X}_1$		$\mathbf{X}_2$		$\mathbf{X}_3$	
	n_{s1}	Rank	n_{s2}	Rank	n_{s3}	Rank
Homogeneous	100	(60,40)	0	–	0	–
Inhomogeneous 1	100	(60,40)	10	(20,10)	0	–
Inhomogeneous 2	100	(60,40)	10	(20,10)	10	(10,10)

Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

Appendix A. Test on sample selection algorithm using synthetic data

Sample selection was used in this work to select samples from the training set that are closely related to the testing samples before imputation of missing values and class labels. We assume that sample selection can remove outlier or unrelated samples from inhomogeneous dataset, and consequently improves the classification performance. To justify our assumption, we have tested the proposed sample selection algorithm by using several sets of synthetic data. The synthetic data with n_s number of samples, n_f number of variables and σ noise level, is generated as follows:

1. Generate a rank-r matrix $\mathbf{X_r} \in \mathbb{R}^{n_s \times n_f}$ by multiplying a randomly generated $n_s \times r$ matrix with another randomly generated $r \times n_f$ matrix, where elements of both matrices are drawn i.i.d. from a standard normal distribution.
2. Add Gaussian noise $\mathcal{N}(0, \sigma^2)$ to each element of matrix $\mathbf{X_r}$.
3. Generate a weight vector $\mathbf{w} \in \mathbb{R}^{n_f \times 1}$ where its elements are drawn i.i.d. from a standard normal distribution.
4. Generate output label $\mathbf{Y} \in \mathbb{R}^{n_f \times 1}$ from $\mathbf{Y} = \text{sign}(\mathbf{X_r} \times \mathbf{w} + \mathbf{N})$, where $\mathbf{N}$ is a noise vector with its elements are drawn i.i.d. from $\mathcal{N}(0, \sigma^2)$.

We simulated a multi-modal dataset by generating two different $\mathbf{X_r}$ with the same label $\mathbf{Y}$, and arranging them side by side, e.g., $\mathbf{X} = [\mathbf{X_{r1}}\mathbf{X_{r2}}]$. We simulated heterogeneous dataset by generating several $\mathbf{X}$ with different rank or $\mathbf{W}$, and stacking them together, e.g., $\mathbf{X_{het}} = [\mathbf{X}_1; \mathbf{X}_2]$, where $\mathbf{X}_1 \in \mathbb{R}^{n_{s1} \times 2n_f}$ and $\mathbf{X}_2 \in \mathbb{R}^{n_{s2} \times 2n_f}$. We simulated missing data by randomly removing half of the feature data (row by row) from the second modality of $\mathbf{X_{het}}$ (i.e., $\mathbf{X_{r2}}$ part for both $\mathbf{X}_1$ and $\mathbf{X}_2$).

Table A.11 shows the details of the generated data. One homogeneous data and two inhomogeneous (heterogeneous) data were generated. The homogenous data was created by using a single matrix $\mathbf{X}_1$, the inhomogeneous data 1 was created by stacking $\mathbf{X}_1$ and $\mathbf{X}_2$, while inhomogeneous data 2 was created by stacking $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$. Each $\mathbf{X}_m$, $m = \{1,2,3\}$ is a "two-modality" simulated data, with each modality containing 80 features (i.e., $n_f = 80$), respectively. The rank for each modality data is shown in Table A.11. Each synthesized data with four different levels of noise (i.e., $\sigma = \{2,1,0.5,0.1\}$) were used in experiment.

We then tested our framework (specifically the sample selection algorithm) on the synthetic data by using 10-fold cross-validation scheme, similar to the scheme used in this manuscript. The simulation results using homogeneous and 2 types of inhomogeneous data are shown in Figs. A.8, A.9 and A.10, respectively. The x-axis of these figures is the λ in sample selection, the higher the λ value, the more the removed training samples. The average number of samples selected from the training set for each fold is shown at the bottom right corner

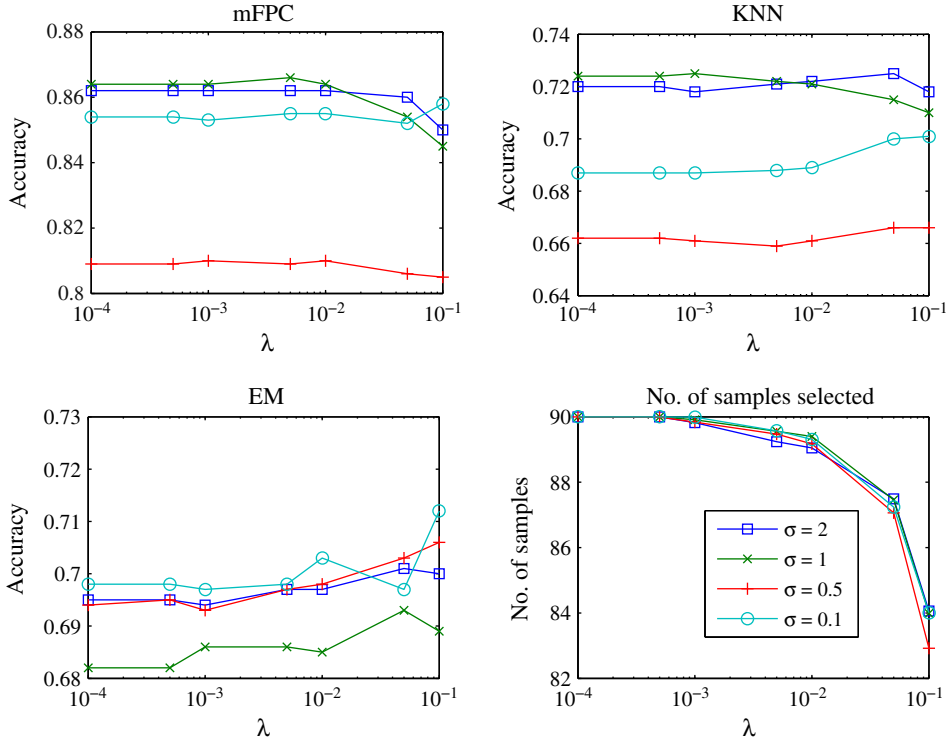

Fig. A.8. Classification result for homogenous data matrix.

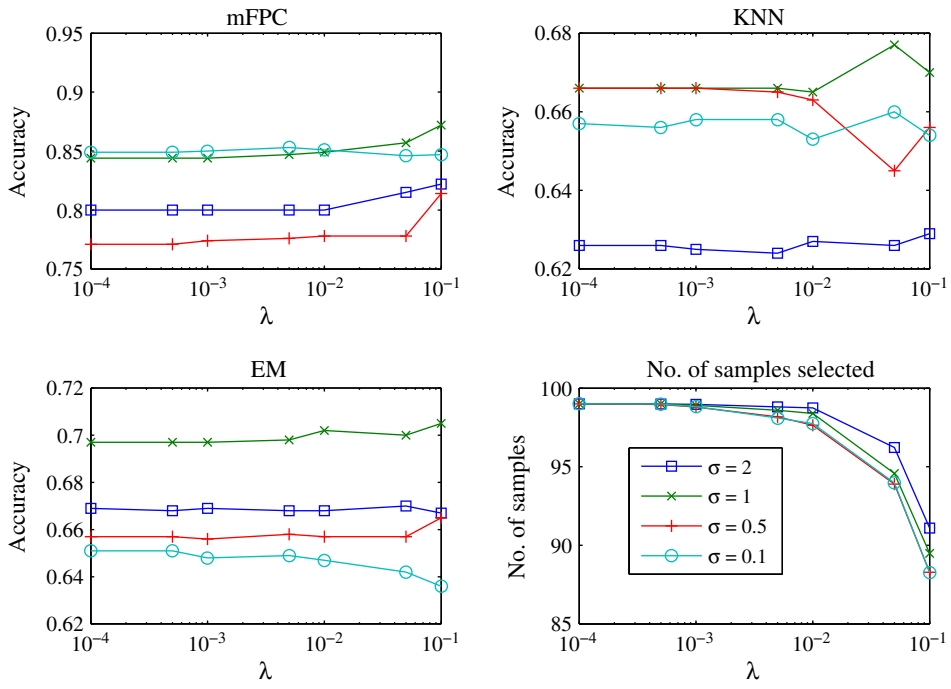

Fig. A.9. Classification result for inhomogeneous data matrix 1.

of all the three figures. The other three plots in these figures are the classification accuracies versus λ, using mFPC, KNN and EM imputations, respectively. From the Fig. A.8, the classification accuracies for mFPC and KNN are rather stable for all the λ values, as expected for homogeneous data. However, we surprisingly notice that the sample selection improves the classification accuracies for EM imputation using incomplete homogeneous data matrix. This is probably because sample selection removes some noisy samples from the training samples that improves the EM imputation. From the Fig. A.9, where the number of "outlier" samples is about 10% of the total samples, the sample selection algorithm slightly improves the classification accuracies for all the three imputation methods, especially when the noise level in the data is higher, i.e., $\sigma = \{2,1\}$. However, we also notice that there are some declines in classification accuracies for low noise curves ($\sigma = \{0.1,0.5\}$) using KNN and EM imputations, when higher λ values are used. When the number of "outlier" samples is increased to about 20% of the total samples, the classification accuracies of mFPC and KNN improve significantly, particularly for data with higher noise level, as shown in Fig. A.10. The effect of sample selection on EM imputation is not obvious for both the inhomogeneous data matrices 1 and 2.

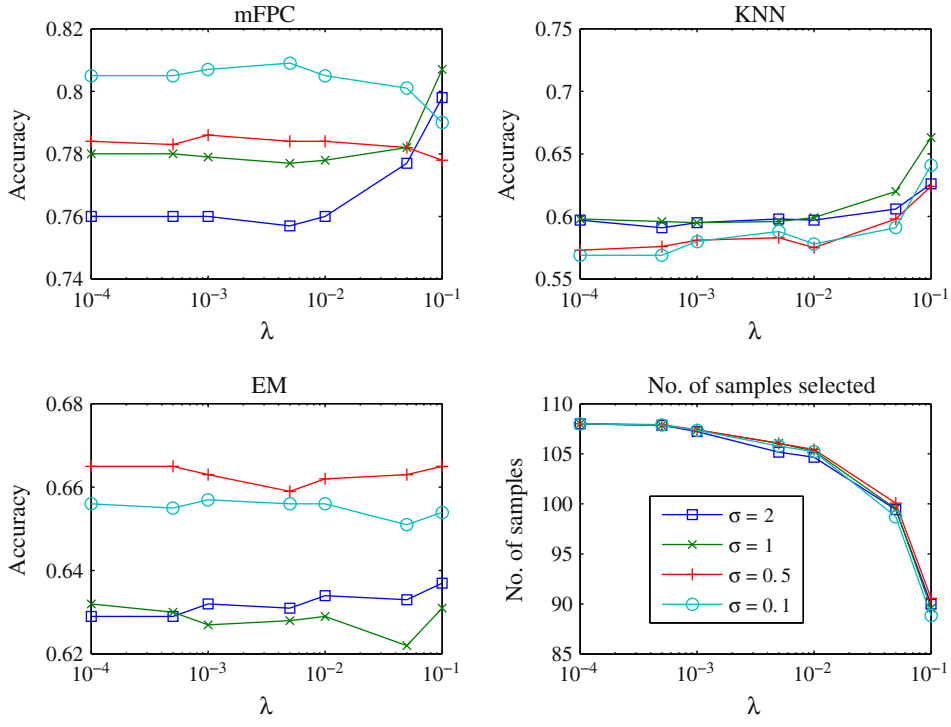

Fig. A.10. Classification result for inhomogeneous data matrix 2.

In summary, these simulation results support our assumption that removing noisy samples (due to Gaussian noise) or unrelated samples (due to inhomogeneous data) from the training dataset can improve classification performance. Sample selection improves mFPC and KNN imputation when the data is more noisy and inhomogeneous, while improves EM imputation when the data is homogeneous.

References

Alzheimer's Association, 2013. 2013 Alzheimer's disease facts and figures. Alzheimer's Dement. 9.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Found. Comput. Math. 9, 717–772.

Chételat, G., Desgranges, B., De La Sayette, V., Viader, F., Eustache, F., Baron, J.C., 2003. Mild cognitive impairment can FDG-PET predict who is to rapidly convert to Alzheimer's disease? Neurology 60, 1374–1377.

Chételat, G., Eustache, F., Viader, F., Sayette, V.D.L., Pélerin, A., Mézenge, F., Hannequin, D., Dupuy, B., Baron, J.C., Desgranges, B., 2005. FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. Neurocase 11, 14–25.

Convit, A., De Asis, J., De Leon, M., Tarshish, C., De Santi, S., Rusinek, H., 2000. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. Neurobiol. Aging 21, 19–26.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56, 766–781.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32, 2322.e19–2322.e27.

De Leon, M., George, A., Golomb, J., Tarshish, C., Convit, A., Kluger, A., De Santi, S., Mc Rae, T., Ferris, S., Reisberg, B., et al., 1997. Frequency of hippocampal formation atrophy in normal aging and Alzheimer's disease. Neurobiol. Aging 18, 1–11.

De Leon, M., DeSanti, S., Zinkowski, R., Mehta, P., Pratico, D., Segal, S., Rusinek, H., Li, J., Tsui, W., Louis, L.S., et al., 2006. Longitudinal CSF and MRI biomarkers improve the diagnosis of mild cognitive impairment. Neurobiol. Aging 27, 394–401.

Desikan, R.S., Cabral, H.J., Hess, C.P., Dillon, W.P., Glastonbury, C.M., Weiner, M.W., Schmansky, N.J., Greve, D.N., Salat, D.H., Buckner, R.L., et al., 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132, 2048–2057.

Du, A.T., Schuff, N., Kramer, J.H., Rosen, H.J., Gorno-Tempini, M.L., Rankin, K., Miller, B.L., Weiner, M.W., 2007. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 130, 1159–1166.

Fagan, A.M., Roe, C.M., Xiong, C., Mintun, M.A., Morris, J.C., Holtzman, D.M., 2007. Cerebrospinal fluid tau/beta-amyloid42 ratio as a prediction of cognitive decline in nondemented older adults. Arch. Neurol. 64, 343.

Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., Avants, B.B., Gee, J.C., Wang, J., Shen, D., 2007a. Multivariate examination of brain abnormality using both structural and functional MRI. Neuroimage 36, 1189–1199.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007b. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging 26, 93–105.

Fan, Y., Resnick, S.M., Wu, X., Davatzikos, C., 2008. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. Neuroimage 41, 277–285.

Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M., et al., 2010. CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. J. Neurosci. 30, 2088–2101.

Foster, N.L., Heidebrink, J.L., Clark, C.M., Jagust, W.J., Arnold, S.E., Barbas, N.R., DeCarli, C.S., Turner, R.S., Koeppe, R.A., Higdon, R., et al., 2007. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. Brain 130, 2616–2635.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., et al., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47, 1476.

Ghannad-Rezaie, M., Soltanian-Zadeh, H., Ying, H., Dong, M., 2010. Selection–fusion approach for classification of datasets with missing values. Pattern Recogn. 43, 2340–2350.

Goldberg, A., Zhu, X., Recht, B., Xu, J., Nowak, R., 2010. Transduction with matrix completion: three birds with one stone. Advances in Neural Information Processing Systems, 23, pp. 757–765.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Herholz, K., Salmon, E., Perani, D., Baron, J., Holthoff, V., Frölich, L., Schönknecht, P., Ito, K., Mielke, R., Kalbe, E., et al., 2002. Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET. Neuroimage 17, 302–316.

Higdon, R., Foster, N.L., Koeppe, R.A., DeCarli, C.S., Jagust, W.J., Clark, C.M., Barbas, N.R., Arnold, S.E., Turner, R.S., Heidebrink, J.L., et al., 2004. A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging. Stat. Med. 23, 315–326.

Hinrichs, C., Singh, V., Xu, G., Johnson, S., 2009. MKL for robust multi-modality AD classification. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2009. Springer, pp. 786–794.

Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., 2011. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55, 574–589.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Huang, K., Aviyente, S., 2006. Sparse representation for signal classification. Advances in Neural Information Processing Systems, pp. 609–616.

Ingalhalikar, M., Parker, W.A., Bloy, L., Roberts, T.P., Verma, R., 2012. Using multiparametric data with missing features for learning patterns of pathology. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012. Springer, pp. 468–475.

Jia, H., Wu, G., Wang, Q., Shen, D., 2010. ABSORB: Atlas building by self-organized registration and bundling. Neuroimage 51 (3), 1057–1070.

Jollois, F.X., Nadif, M., 2007. Speed-up for the expectation-maximization algorithm for clustering categorical data. J. Glob. Optim. 37, 513–525.

Kabani, N.J., 1998. A 3D atlas of the human brain. NeuroImage 7, S717.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689.

Landau, S., Harvey, D., Madison, C., Reiman, E., Foster, N., Aisen, P., Petersen, R., Shaw, L., Trojanowski, J., Jack, C., et al., 2010. Comparing predictors of conversion and decline in mild cognitive impairment. Neurology 75, 230–238.

Li, Y., Wang, Y., Wu, G., Shi, F., Zhou, L., Lin, W., Shen, D., 2012. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. Neurobiol. Aging 33 427–e15.

Liu, M., Zhang, D., Shen, D., 2012. Ensemble sparse classification of Alzheimer's disease. Neuroimage 60, 1106–1116.

Liu, J., Ye, J., 2010. Efficient l_1/l_q norm regularization. arXiv preprint arXiv:1009.4766.

Liu, J., Ji, S., Ye, J., 2009. Multi-task feature learning via efficient $l_{2,1}$-norm minimization. Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, pp. 339–348.

Liu, F., Wee, C.Y., Chen, H., Shen, D., 2014. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. Neuroimage 84, 466–475.

Ma, S., Goldfarb, D., Chen, L., 2011. Fixed point and Bregman iterative methods for matrix rank minimization. Math. Program. 128, 321–353.

Mielke, R., Kessler, J., Szelies, B., Herholz, K., Wienhard, K., Heiss, W.D., 1998. Normal and pathological aging—findings of positron-emission-tomography. J. Neural Transm. 105, 821–837.

Morris, J.C., Roe, C.M., Grant, E.A., Head, D., Storandt, M., Goate, A.M., Fagan, A.M., Holtzman, D.M., Mintun, M.A., 2009. Pittsburgh compound B imaging and prediction of progression from cognitive normality to symptomatic Alzheimer disease. Arch. Neurol. 66, 1469.

Nesterov, Y., 1983. A method of solving a convex programming problem with convergence rate O (1/k2). Soviet Mathematics Doklady, pp. 372–376.

Nesterov, Y., 2007. Gradient methods for minimizing composite objective function.

Obozinski, G., Taskar, B., Jordan, M.I., 2006. Multi-task feature selection. Tech. Rep. Statistics Department, UC Berkeley.

Oliveira Jr., P.P., Nitrini, R., Busatto, G., Buchpiguel, C., Sato, J.R., Amaro Jr., E., 2010. Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. J. Alzheimer's Dis. 19, 1263–1272.

Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., 2011. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res. Neuroimaging 194, 7–13.

Scarmeas, N., Anderson, K., Hilton, J., Park, A., Habeck, C., Flynn, J., Tycko, B., Stern, Y., 2004. APOE-dependent PET patterns of brain activation in Alzheimer disease. Neurology 63, 913–915.

Schneider, T., 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J. Clim. 14, 853–871.

Shen, D., Davatzikos, C., 2004. Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping. Neuroimage 21 (4), 1508–1517.

Shen, D., Davatzikos, C., 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21, 1421–1439.

Shen, D., Wong, W.H., Ip, H.H., 1999. Affine-invariant image retrieval by correspondence matching of shapes. Image and Vision Computing 17 (7), 489–499.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.

Speed, T., 2003. Statistical Analysis of Gene Expression Microarray Data. CRC Press.

Tang, S., Fan, Y., Wu, G., Kim, M., Shen, D., 2009. RABBIT: rapid alignment of brains by building intermediate templates. Neuroimage 47 (4), 1277–1287.

Thung, K.H., Wee, C.Y., Yap, P.T., Shen, D., 2013. Identification of alzheimers disease using incomplete multimodal dataset via matrix shrinkage and completion, in: Machine Learning in Medical Imaging, pp. 163–170. Springer.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525.

Verma, R., Mori, S., Shen, D., Yarowsky, P., Zhang, J., Davatzikos, C., 2005. Spatiotemporal maturation patterns of murine brain quantified by diffusion tensor MRI and deformation-based morphometry. Proceedings of the national academy of sciences of the United States of America 102, 6978–6983.

Walhovd, K., Fjell, A., Dale, A., McEvoy, L., Brewer, J., Karow, D., Salmon, D., Fennema-Notestine, C., 2010. Multi-modal imaging predicts memory performance in normal aging and cognitive decline. Neurobiol. Aging 31, 1107–1121.

Wang, Y., Nie, J., Yap, P.T., Li, G., Shi, F., Geng, X., Guo, L., Shen, D., Alzheimer's Disease Neuroimaging Initiative, 2014. Knowledge-Guided Robust MRI Brain Extraction for Diverse Large-Scale Neuroimaging Studies on Humans and Non-Human Primates. PLoS one 9 (1), e77810.

Wang, Y., Nie, J., Yap, P.T., Shi, F., Guo, L., Shen, D., 2011. Robust deformable-surface-based skull-stripping for large-scale studies. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer, pp. 635–642.

Wee, C.Y., Yap, P.T., Li, W., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2011. Enriched white matter connectivity networks for accurate identification of MCI patients. Neuroimage 54, 1812–1822.

Wee, C.Y., Yap, P.T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2012. Identification of MCI individuals using structural and functional connectivity networks. Neuroimage 59, 2045–2056.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T.S., Yan, S., 2010. Sparse representation for computer vision and pattern recognition. Proc. IEEE 98, 1031–1044.

Wu, G., Qi, F., Shen, D., 2006. Learning-based deformable registration of MR brain images. IEEE Trans. Med. Imaging 25 (9), 1145–1157.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., Ye, J., 2013. Bi-level multi-source learning for heterogeneous block-wise missing data. Neuroimage. http://dx.doi.org/10.1016/j.neuroimage.2013.08.015.

Xu, L., Jordan, M.I., 1996. On convergence properties of the em algorithm for Gaussian mixtures. Neural Comput. 8, 129–151.

Xue, Z., Shen, D., Karacali, B., Stern, J., Rottenberg, D., Davatzikos, C., 2006a. Simulating deformations of MR brain images for validation of atlas-based segmentation and registration algorithms. Neuroimage 33 (3), 855–866.

Xue, Z., Shen, D., Davatzikos, C., 2006b. Statistical representation of high-dimensional deformation fields with application to statistically constrained 3D warping. Medical Image Analysis 10 (5), 740–751.

Yang, J., Pan, P., Song, W., Huang, R., Li, J., Chen, K., Gong, Q., Zhong, J., Shi, H., Shang, H., 2012. Voxelwise meta-analysis of gray matter anomalies in Alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation. J. Neurol. Sci. 316, 21–29.

Yang, J., Shen, D., Davatzikos, C., Verma, R., 2008. Diffusion tensor image registration using tensor geometry and orientation features. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008. Springer, pp. 905–913.

Yap, P.T., Wu, G., Zhu, H., Lin, W., Shen, D., 2009. TIMER: tensor image morphing for elastic registration. Neuroimage 47 (2), 549–563.

Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., Janardan, R., Liu, H., Alexander, G., et al., 2008. Heterogeneous data fusion for Alzheimer's disease study. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1025–1033.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat Methodol. 68, 49–67.

Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J., 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. Neuroimage 61, 622–632.

Zacharaki, E.I., Shen, D., Lee, S.K., Davatzikos, C., 2008. ORBIT: a multiresolution framework for deformable registration of brain tumor images. IEEE Trans. Med. Imaging 27 (8), 1003–1017.

Zhang, D., Shen, D., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimaging 59, 895–907.

Zhang, D., Shen, D., et al., 2012. Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. PloS one 7 (3), e33182.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55, 856–867.

Zhou, L., Wang, Y., Li, Y., Yap, P.T., Shen, D., et al., 2011. Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. PloS one 6, e21935.

Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, Z., 2011. Missing value estimation for mixed-attribute data sets. IEEE Trans. Knowl. Data Eng. 23, 110–121.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat Methodol. 67, 301–320.