Contents lists available at ScienceDirect

Pattern Recognition

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

Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of Alzheimer's disease

Peng Cao[a,b,*], Xuanfeng Shan[a], Dazhe Zhao[b], Min Huang[c], Osmar Zaiane[d]

[a] *Computer Science and Engineering, Northeastern University, Shenyang, China*
[b] *Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang, China*
[c] *Information Science and Engineering, Northeastern University, Shenyang, China*
[d] *Computing Science, University of Alberta, Edmonton, Alberta, Canada*

ARTICLE INFO

Article history:
Received 4 November 2016
Revised 7 June 2017
Accepted 17 July 2017
Available online 18 July 2017

Keywords:
Alzheimer's disease
Multi-task learning
Proximal gradient
Regression
Biomarker discovery

ABSTRACT

Alzheimer's disease (AD), the most common form of dementia, not only causes progressive impairment of memory and other cognitive functions of patients, but also becomes the substantial financial burden to the health care system. There is thus an urgent need to (1) accurately predict the cognitive performance of the disease, and (2) identify potential MRI-related biomarkers most predictive of the estimation of cognitive outcomes. In this paper, we develop a novel multi-task learning formulation to explore the correlation existing in Magnetic Resonance Imaging (MRI) and cognitive measures by a mixed norm incorporating a hierarchical group sparsity and shared subspace uncovering regularization, to learn a shared structure from multiple related tasks with considering implicit shared subspace structure and explicit subset of features as well as Region-of-Interests (ROIs) simultaneously. An efficient alternating optimization algorithm is derived to solve the proposed non-convex and non-smooth objective formulation. We comprehensively evaluate the proposed algorithm for the cognitive outcome prediction including all subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The experimental results not only demonstrate the proposed method has superior performance over multiple state-of-the-art comparable approaches, but also identifies cognition-relevant MRI biomarkers that are consistent with prior knowledge.

1. Introduction

Dementia poses a serious challenge to the aging society. Alzheimer's disease (AD), the most common form of dementia, is a gradually progressive syndrome that mainly affects memory function, ultimately culminating in a dementia state where all cognitive functions are affected [1]. The worldwide prevalence of AD is predicted to quadruple from 46.8 million in 2016 to 131.5 million by 2050 according to ADI's World Alzheimer Report [2]. The huge price of caring for AD patients has made it one of the most costly diseases in the developed countries. The total estimated worldwide cost of ADntia is US $818 billion today, and it will increase to a trillion dollar by 2018 [3].

Diagnosed in an early stage, therapeutic interventions can be made to slow down the progression of AD. Previous AD diagnosis mainly relies on clinical observation and cognitive evaluation. Recent studies [4,5] show the image analysis of brain scans from neuroimaging is more reliable and accurate in detecting the presence of AD than traditional clinical evaluation. Thus more and more attention have been now shifting to finding effective biomarkers by image analysis of brain scans and applying machine learning methods to perform automatic early detection. To date, several biomarkers have been studied and proven to be sensitive to brain atrophy captured by MRI [6,11,12,25], PET [15,16], fMRI [13,14]. Recently, Jack Jr. et al. [17] reported that structural abnormalities can be observed in the human brain prior to any clinical symptom, indicating that structural abnormalities can be utilized for early detection of AD. Hence, magnetic resonance imaging (MRI) has been widely used in the diagnosis or cognitive performance prediction of AD [18,33].

Many classification approaches have been designed to the diagnosis of AD and its prodromal stage: mild cognitive impairment (MCI) based on the MRI scans [6,8,10,25,34,54]. However a definitive diagnosis of AD can only be made with histopathological confirmation of amyloid plaques and neurofibrillary tangles, usually at autopsy. Many clinical/cognitive measures have been designed to evaluate the cognitive status of the patients and used

[*] Corresponding author.
E-mail address: caopeng@cse.neu.edu.cn (P. Cao).

http://dx.doi.org/10.1016/j.patcog.2017.07.018
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

as important criteria for clinical diagnosis of probable AD. Compared with the discrete patient labels, cognitive performance evaluation provides additional valuable information for studying the underlying disease mechanisms. The most commonly used cognitive measures are Alzheimer's Disease Assessment Scale cognitive total score (ADAS), Mini Mental State Exam score (MMSE), Rey Auditory Verbal Learning Test (RAVLT). ADAS is the gold standard in AD drug trial for cognitive function assessment, extensively used to measure the severity of important symptoms of AD, including memory disturbances, language, praxis, attention, and other cognitive abilities. MMSE is used extensively in clinical and research settings to measure cognitive impairment, and examines orientation to time and place, attention and calculation, immediate and delayed recall of words, language and visuo-constructional functions. RAVLT is a test of episodic memory and sensitive to deficiencies of memory found in many groups of patients, and widely used for the diagnosis of memory disturbances. In the literature, regression models have been widely studied to reveal the relationship between neuroimaging markers and cognitive measures [47,53], and investigate the prediction performance of neuroimaging measures for inferring cognitive outcomes [47,53] or tracking disease progression [7,26,35].

It is known that there exist inherent correlations among multiple clinical cognitive variables of a subject. However, many works do not model dependence relation among multiple tasks and neglect the correlation between clinical tasks which is potentially useful. When the tasks are believed to be related, learning multiple related tasks jointly can improve performance relative to learning each task separately. Multi-Task Learning (MTL) is a statistical learning framework which seeks at learning several models in a joint manner. It has been commonly used to obtain better generalization performance than learning each task individually [19,31,46]. The critical issues in MTL is to identify how the tasks are related and build learning models to capture such task relatedness. Recently, the multi-task learning based feature learning methods (MTFL) with sparsity-inducing norm have been widely studied to select the discriminative feature subset from MRI features by incorporating inherent correlations among multiple clinical cognitive measures [26,20,21]. For example, the $\ell_{2,1}$-norm regularization penalizes each row of parameters matrix as a whole and enforces sparsity among the rows, it is able to select the most discriminative features. Wang et al. [20] and Zhang et al. [21] employed multi-task feature learning strategies for selecting biomarkers that could predict multiple clinical scores. Specially, Wang [20] further considers some important features are only correlated to a subset of tasks, and adds an ℓ_1-norm regularizer to impose the sparsity among all elements and propose to use the combined $\ell_{2,1}$-norm and ℓ_1-norm regularizations to select features; Zhang proposed a multi-task learning with $\ell_{2,1}$-norm to select the common subset of relevant features for multiple variables from each modality by assuming that the related tasks share a common relevant feature subset. The most limitation of the popular learning models assume linear relationship between the MRI features and the cognitive outcomes. To model these more complicated but more flexible relationship between them, Zhang develop a multi-modal support vector regression (SVR) to fuse the above-selected features from all modalities with the selected feature subset [21]. Kernel methods [9] have been studied to model the cognitive scores as nonlinear functions of neuroimaging measures. Recently, many kernel based classification or regression methods with faster optimization speed or stronger generalization performance have been proposed and investigated by theoretically analyzing and experimentally evaluating [22,23].

Despite of the above achievements, few regression models take into account the covariance structure among predictors. To achieve a certain function, brain imaging measures are often correlated with each other. For MRI data, the groups correspond to specific regions-of-interest (ROIs) in the brain, e.g., entorhinal and hippocampus. Individual features are specific properties of those regions, e.g., cortical volume and thickness. For each region (group), the multiple features are extracted to measure the atrophy information of each ROI involving cortical thickness, surface area and volume from gray matters and white matters in this study. The multiple shape measures from the same region provide a comprehensively quantitative evaluation of cortical atrophy, and tend to be selected together as joint predictors. However, the $\ell_{2,1}$-norm regularization only consider the shared representation from the features in the original space, neglecting the potentially grouping information among multiple neuroimaging measures and interrelated correlation in the latent shared low-dimensional subspace among multiple cognitive outcomes.

In order to adequately exploit the intrinsic relatedness among the tasks and make our model interpretable, we propose a sparse shared structure based multi-task learning formulation, named **S**parse **G**roup **L**asso with shared **S**ubspace based MTL(SGLS-MTL). The regularizer in SGLS-MTL consists of three components including an $\ell_{2,1}$-norm penalty on the regression weight vectors, which ensures that a small subset of features will be selected for the cognitive outcomes prediction models, a group $\ell_{2,1}$-norm penalty, which takes into account the interrelated correlation among multiple neuroimaging measures and identifies the cognition-relevant brain region, and a subspace structure uncovering regularization, which aims to search for a suitable low dimensional subspace of the given input feature space to uncover the shared structure across tasks [31,43]. Therefore, the proposed MTL model captures the task relationships from sparse representation with respect to feature and region (ROI) on the original feature space, and underlying subspace structure on the latent low-dimensional subspace simultaneously, which can be seen as two different underlying structure. Fig. 1 shows the schematic flowchart of the proposed SGLS-MTL framework. The proposed formulation is challenging to solve due to the use of non-convex and non-smooth penalties. To solve the objective function efficiently, we propose an efficient iterative algorithm utilizing accelerated proximal gradient (APG) [28] due to its fast convergence property.

The main contributions of this paper can be summarized as follows:

(1) We propose a mixed sparse shared structure based multi-task learning combining a two-level sparsity and subspace structure uncovering. We show that SGLS-MTL provides better predictive performance and more interpretable than the state-of-the-art in the cognitive performance prediction of AD.
(2) We design an efficient accelerated projected gradient optimization algorithm to solve the non-convex and non-smooth objective function.
(3) We extend the regularization of sparse group lasso to clustering involving the subspace based multi-task clustering and multi-subspace clustering, and empirically demonstrates the effectiveness of the embedded sparse learning framework in unsupervised scenarios.
(4) We extend the concept of "group" structure from ROI to modality, to evaluate the performance of our method on fusing multiple modalities.
(5) The proposed method was evaluated on a large database using the entire 788 baseline MRI scans in the ADNI study [40,41]. We carried out extensive experiments to test the performance of SGLS-MTL along various dimensions including predicting the baseline cognitive outcomes and future cognitive outcomes, identifying biomarkers and the predic-

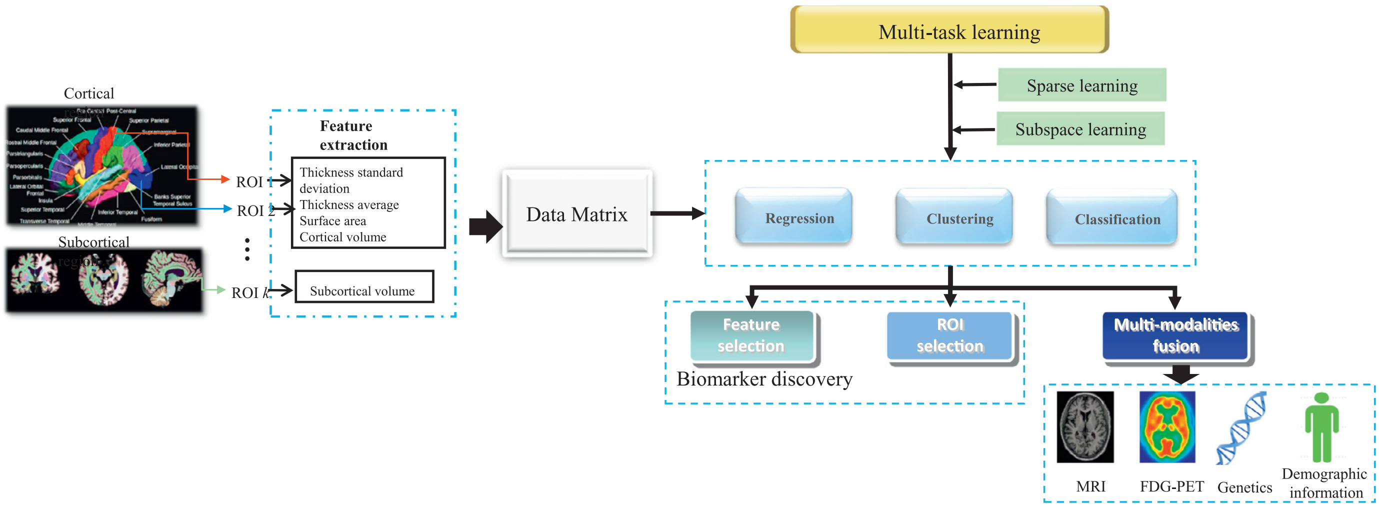

Fig. 1. Flow chart of the proposed SGLS-MTL method. The goal of our work is to predict subjects' cognitive scores in a number of neuropsychological assessments at baseline time using their MRI measures across the entire brain, respectively. From each ROI, multiple features are extracted to measure the atrophy information involving cortical thickness, surface area and volume from gray matters and white matters in this study. Sparse learning and subspace learning are incorporated into our multi-task learning framework to model the task relatedness and group structure of features. Our framework allows us not only to identify the MRI biomarkers from both the level of features and ROIs, but also to deal with fusion of multi-modalities data. Furthermore, the SGLS-MTL framework can be extended to classification model or clustering model to predict the stage of the disease or partition the instances without the label information.

tion of the heterogeneous tasks with incorporating the task of classification.

The rest of the paper is organized as follows. A description of the data used, the formulation and optimization procedure of the proposed SGLS-MTL are given in Section 2. Section 3 discusses experimental results using ADNI dataset. We conclude in Section 4.

2. Materials and methods

2.1. Data

MR images and data used in this work were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu) [40]. The 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. The current work focuses on MRI data. In ADNI, all participants received 1.5 Tesla (T) structural MRI. The MRI features used in our experiments are based on the imaging data from the ADNI database processed by a team from UCSF (University of California at San Francisco), who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) according to the atlas generated in [42]. The FreeSurfer software was employed to automatically label cortical and subcortical tissue classes for the structural MRI scan of each subject, and to extract thickness measures of cortical regions of interests (ROIs) and volume measures of cortical and subcortical. For each cortical region, the cortical thickness average (TA), standard deviation of thickness (TS), surface area (SA) and cortical volume (CV) were calculated as features. For each subcortical, subcortical volume was calculated as features. Left and right hemisphere SA and total intracranial volume (ICV) were also included. This yielded a total of $p = 319$ MRI features (including 275 cortical and 44 subcortical features, see Table S1). Details of the analysis procedure are available at http://adni.loni.ucla.edu/research/mri-post-processing/.

Table 1
Summary of ADNI dataset and subject information.

Category	CN	MCI	AD
Number	225	390	173
Gender (M/F)	119/109	257/142	99/90
Age (y, ag ± sd)	75.8 ± 5.0	74.7 ± 7.4	75.2 ± 7.5
Edu (y, ag ± sd)	16.1 ± 2.8	15.6 ± 3.0	14.7 ± 3.2

CN, Cognitively Normal; MCI, Mild Cognitive Impairment; AD, Alzheimer's Disease; M, male; F, female; Edu, Education; y, years; ag, average; sd, standard deviation.

Table 2
The amounts of patients of follow-up visits.

Time point	M06	M12	M24	M36	M48
Instance size	718	662	532	345	91

In this work, only ADNI subjects with no missing feature and cognitive outcome information baseline data are included. This yields a total of $n = 788$ subjects, who are categorized into 3 baseline diagnostic groups: Cognitively Normal (CN, $n_1 = 225$), Mild Cognitive Impairment (MCI, $n_2 = 390$), and Alzheimer's Disease (AD, $n_3 = 173$). Details of the demographics and clinical characteristics of the sample used in this paper are presented in Table 1.

ADNI is also a longitudinal project, in which the measurements are collected repeatedly over a 6-month or 1-year interval. The information of subjects used in this study at different time points is given in Table 2. The date when the patient performs the screening in the hospital for the first time is called baseline, and the time point for the follow-up visits is denoted by the duration starting from the baseline. For instance, we use the notation "M06" to denote the time point half year after the first visit. Currently ADNI has up to 48 months follow-up data for some patients. The amount of instances of each task is different in Table 2 since the data sets decrease in size due to the drop out of some patients for various reasons.

In this work, we remove features with more than 10% missing entries (for all patients and all time points), exclude patients without baseline MRI records and complete the missing entries using the average value.

2.2. *Sparse Group Lasso with shared Subspace based Multi-Task Learning (SGLS-MTL) Formulation*

Assume that we are given m supervised learning tasks. The majority of the proposed methods fall into the class of regularized multi-task learning, which has the form:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{p\times m}} J = \mathcal{L}(\boldsymbol{W};\boldsymbol{y},\boldsymbol{X}) + \lambda\mathcal{R}(\boldsymbol{W}) \tag{1}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^T \in \mathbb{R}^{n\times p}$ is the training set (n and p are the number of training instances and dimensionality of $\boldsymbol{x}_i$), $\boldsymbol{Y} = (\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)^T \in \mathbb{R}^{n\times m}$, where $\boldsymbol{y}_i$ is the target value vector (class label or cognitive score) for $\boldsymbol{x}_i$ and m is the number of tasks, $\boldsymbol{W} \in \mathbb{R}^{p\times m}$ is a coefficient matrix need to be learned. $\mathcal{L}(\cdot)$ denotes the loss function. In the context of regression, we assume the loss $\mathcal{L}_R(\boldsymbol{W};\boldsymbol{Y},\boldsymbol{X}) = \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{W}\|_F^2 = \sum_{i=1}^n \|\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{W}\|_2^2$. $\mathcal{R}$ is the regularizer encouraging the shared representation among the tasks and λ is a regularization parameter needed to be chosen by cross validation.

Shared structure learning [31,36,43] has been successfully used in multi-label image annotation [36], multi-view learning [37] and multimedia analysis [38]. Shared structure learning assumes that there is a certain common information shared among data samples and aims to search for a suitable low dimensional subspace of the given input feature space to uncover the shared structure across tasks. It has been claimed in [43] that there should be a shared subspace across multiple tasks and uncovering this shared subspace can improve classification performance. The concepts of an instance are predicted by its vector representation in the original feature space together with the embedding in the shared subspace, which can be generalized as the following demonstration: $f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{u} = \boldsymbol{x}\boldsymbol{w} + \boldsymbol{x}\boldsymbol{Q}^T\boldsymbol{v}$, where $\boldsymbol{u}\in\mathbb{R}^{p\times 1}$, $\boldsymbol{w}\in\mathbb{R}^{p\times 1}$ and $\boldsymbol{v}\in\mathbb{R}^{h\times 1}$ where h is the dimension of the shared subspace, are the weight vectors for the full feature space, the high-dimensional feature space and the shared low dimensional feature space (a linear form of feature map is considered for simplicity), respectively. $\boldsymbol{Q}\in\mathbb{R}^{h\times p}$ is a (to be learnt) linear low dimensional map common across the tasks and is constrained to be a matrix with orthonormal rows, $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$. Ando and Zhang [43] assumed that the tasks share a latent low-dimensional subspace and proposed an Alternating Structure Optimization (ASO) approach explicitly learn this subspace in the learning formulation. The formulation of ASO is non-convex and the alternating structure optimization procedure is not guaranteed to find a global optimum. In [31], Chen et al. presented an improved ASO formulation (called iASO) given by:

$$\underset{\boldsymbol{U},\boldsymbol{V},\boldsymbol{Q}^T\boldsymbol{Q}=I}{\arg\min} \ \|\boldsymbol{X}\boldsymbol{U}-\boldsymbol{Y}\|_F^2 + \lambda_1\left\|\boldsymbol{U}-\boldsymbol{Q}^T\boldsymbol{V}\right\|_F^2 + \lambda_2\|\boldsymbol{U}\|_2^2, \tag{2}$$

where $\boldsymbol{U} = \boldsymbol{W} + \boldsymbol{Q}^T\boldsymbol{V}$. The first regularization term $\left\|\boldsymbol{U}-\boldsymbol{Q}^T\boldsymbol{V}\right\|_F^2$ controls the task relatedness by sharing a low-dimensional feature map, while the second regularization term $\|\boldsymbol{U}\|_2^2$ controls the complexity of the models for each task. This formulation provides the foundation for our SGLS-MTL method. Then, Chen et al. proposed a convex relaxation of ASO (cASO) algorithm to solve the convex relaxation efficiently, and further showed that cASO converges to a global optimum. However, it is generally difficult for shared structure learning to interpret or investigate the results. Additionally, feature selection plays an important role in the diagnosis of AD since the extracted features or original neuroimaging data is extremely high dimensional. It is advantageous to exert the sparse feature selection models on the regularization term to discover a small set of imaging biomarkers that are easier to interpret and deal with the curse of dimensionality.

Sparse methods have attracted a great amount of research efforts in the past decade due to its sparsity-inducing property, leading to interpretable models by effective feature selecting [30]. Sparsity usually means that only a small portion of the solution components are non-zero. Inspired by the recent success of sparse methods, we develop a novel multi-task learning formulation to explore the correlation existing in imaging and cognitive measure or diagnosis status by a sparse shared structure regularization. The sparse shared structure regularization incorporates a hierarchical group sparsity and a subspace uncovering regularization in a unified framework, to learn a shared representation from multiple related tasks with considering the shared subset of features on the original feature subspace and shared subspace structure on the latent low-dimensional subspace jointly. Specifically, the regularization $\mathcal{R}(\cdot)$ in our formulation consists of two parts: the first part $\mathcal{R}_1$ is a hierarchical group sparsity and contributed from the representations in the original data space, and the second one $\mathcal{R}_2$ is contributed from the embedding in the latent subspace. $\mathcal{R}_1$ consists of a $\ell_{2,\,1}$-norm and a group $\ell_{2,\,1}$ norm $G_{2,\,1}$, to identify the discriminate features and regions relevant to infer cognitive outcomes and disease status from MRI. The $\ell_{2,\,1}$-norm regularization $\|\boldsymbol{U}\|_{2,1} = \sum_{i=1}^p \|\boldsymbol{u}_{i.}\|_2$, which is known to be an effective model for sparse feature selection for simultaneously enforcing sparsity over features for all tasks. The $\ell_{2,\,1}$-norm of $\boldsymbol{U}$ makes it sparse, meaning that some of its rows shrink to zero. Besides, in the context of AD, the groups correspond to specific ROIs in the brain, and the individual shape features are specific properties of those regions. The multiple shape measures from the same ROI tend to be selected together as joint predictors, and use this prior knowledge of interrelated structure to group relevant shape features together in the same ROI to guide the learning process. We assume the p features to be divided into k disjoint groups $G = \{G_1,\ldots,G_k\}$, with each group having p_j features respectively. For our data, the number of features in each group ranges is 4 for cortical region or 1 for subcortical region. included an extra group regularization to group the weights corresponding to the same brain region across multiple time points, which allowed the selection of brain regions based on multiple time points.

To enforce simultaneous group sparsity and allow the selection of brain regions, we introduce a group $\ell_{2,\,1}$ norm regularization which groups the weights corresponding to the same brain region and induces the desirable group-sparse structure in the matrix $\boldsymbol{U}$. The group $\ell_{2,\,1}$ norm is defined as: $\|\boldsymbol{U}\|_{G_{2,1}} = \sum_{j=1}^k w_j\sqrt{\sum_{i\in G_j}\|\boldsymbol{U}_i\|_2^2}$, where $w_j = \frac{1}{\sqrt{p_j}}$ is the weight for j-th group. Hence, the hierarchical sparsity regularization in $\mathcal{R}_1$ can exploit the correlation of tasks from the perspective of features and ROIs, leading to yield an anatomically meaningful biomarker discovery result. $\mathcal{R}_2$ in the regularization is a subspace structure penalty controlling the task relatedness with shared subspace, to encourage an underlying predictive latent subspace structure. Incorporating all of the above norms in the form of regularizers into the proposed multi-task learning model, the following minimization problem of SGLS-MTL is formulated:

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{Q}^T\boldsymbol{Q}=I} \ & \|\boldsymbol{X}\boldsymbol{U}-\boldsymbol{Y}\|_F^2 + \lambda_1\|\boldsymbol{U}\|_{2,1} + \lambda_2\|\boldsymbol{U}\|_{G_{2,1}} \\ & +\lambda_3\|\boldsymbol{U}-\boldsymbol{Q}^T\boldsymbol{V}\|_F^2\end{aligned} \tag{3}$$

Note that λ_1, λ_2 and λ_3 indicate the importance of the corresponding regularization component.

- Remark 1. Although we only consider the least squares loss function here, the above formulation can be easily generalized to other convex loss functions, such as hinge loss or logistic function.

- Remark 2. In the proposed general multi-task learning framework, both classification and regression (heterogeneous multi-task learning) can be performed simultaneously in a unified formulation (Experiment II).
- Remark 3. Several existing models can be viewed as special cases of the SGLS-MTL formulation.
  - When $\lambda_2 = \lambda_3 = 0$, SGLS-MTL simplifies to sparse multi task learning with $\ell_{2,1}$-norm norm. The $\ell_{2,1}$-norm regularized MKL has been used to select features that could predict all or most clinical scores in [21].
  - When $\lambda_2 = 0$ and $\|\boldsymbol{U}\|_{2,1}$ is replaced by $\|\boldsymbol{U}\|_2$, SGLS-MTL simplifies to shared structure learning [31].
  - When $\lambda_1 = \lambda_3 = 0$, SGLS-MTL simplifies to a simple group lasso learning, which is a single task learning method (where $m = 1$), with group lasso.
  - If $\lambda_3 = 0$, it reduce to sparse group lasso without shared subspace learning, named SGL-MTL and the formation is the same as Sparse Multimodal Learning (SMML) proposed in [32]. The difference is that our algorithm employ proximal gradient to solve the non-smooth norm while SMML focuses on a sub-gradient approach, which can be slow and inaccurate at times. The comparison is shown in the experiment section.

2.3. Optimization

The optimal $\boldsymbol{V}$ to the objective function in Eq. (3) can be expressed in the form of a function on $\boldsymbol{Q}$ and $\boldsymbol{U}$. It can be verified that $\mathcal{R}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{Q})$ in Eq. (3) is minimized with respect to $\boldsymbol{V}$ when $\boldsymbol{V} = \boldsymbol{QU}$. Therefore we can denote: $\mathcal{R}'(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{Q}) = \lambda_1\|\boldsymbol{U}\|_{2,1} + \lambda_2\|\boldsymbol{U}\|_{G_{2,1}} + \lambda_3 Tr(\boldsymbol{U}^T(I - Q^TQ)\boldsymbol{U})$. The orthonormality constraint on $\boldsymbol{Q}$ is non-convex, which makes the optimization problem non-convex. One method is to relax the feasible domain of it into a convex set firstly. Let $\boldsymbol{S} = \boldsymbol{Q}^T\boldsymbol{Q}$, the feasible domain of the optimization problem can be relaxed into a convex set according to [31], and a convex formulation of the problem in Eq. (3) can be defined as follows:

$$\begin{aligned} \arg\min_{\boldsymbol{U},\boldsymbol{S}} \; & \|\boldsymbol{XU} - \boldsymbol{Y}\|_F^2 + \lambda_1\|\boldsymbol{U}\|_{2,1} + \lambda_2\|\boldsymbol{U}\|_{G_{2,1}} \\ & + \lambda_3 \mathrm{Tr}(\boldsymbol{U}^T(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{U}) \\ \text{s.t.} \quad & \mathrm{Tr}(\boldsymbol{S}) = h, \boldsymbol{S} \preceq \boldsymbol{I}, \boldsymbol{S} \in \mathbb{R}_+^p \end{aligned} \tag{4}$$

For the optimization problem in Eq. (4), we symbolically denote the optimization variables $\boldsymbol{U}$ and $\boldsymbol{S}$ as: $\boldsymbol{M} = [\boldsymbol{U}\ \boldsymbol{S}], \boldsymbol{U} \in \mathbb{R}^{p\times m}, \boldsymbol{S} \in \mathbb{R}^{p\times p}$. The optimization can be done efficiently using Accelerated Proximal Gradient (APG) [28], which has been widely applied for solving mathematical formulations in the areas of machine learning due to its optimal convergence rate among all the first-order methods as well as its scalability for large-scale data analysis. It updates the intermediate solution point toward the globally optimal solution via computing the proximal operator and estimating the step size. APG maintains two sequences of variables: a feasible solution sequence $\{\boldsymbol{M}^{(t)}\}$ and a searching point sequence $\left\{\hat{\boldsymbol{M}}^{(t)}\right\}$. The general scheme of APG can be described: at the t-th iteration of APG, the solution point $\boldsymbol{M}^{(t+1)}$ can be computed via:

$$\begin{aligned} \boldsymbol{M}^{(t+1)} &= \mathbf{prox}_L^{\lambda_1,\lambda_2}(\hat{\boldsymbol{M}}^{(t)}) = \arg\min_{\boldsymbol{M}} \mathcal{R}_{\lambda_2}^{\lambda_1}(\boldsymbol{M}) \\ &+ \frac{L}{2}\|\boldsymbol{M} - \left(\hat{\boldsymbol{M}}^{(t)} - \frac{1}{L^{(t)}}\nabla f(\hat{\boldsymbol{M}}^{(t)})\right)\|^2, \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{M}}^{(t)}$ denotes a searching point constructed from a linear combination of $\boldsymbol{M}^{(t)}$ and $\boldsymbol{M}^{(t-1)}$ from previous two iterations, $f(\boldsymbol{M}) = \|\boldsymbol{XU} - \boldsymbol{Y}\|_F^2 + \lambda_3\mathrm{Tr}(\boldsymbol{U}^T(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{U})$ and $\mathcal{R}_{\lambda_2}^{\lambda_1}(\boldsymbol{M}) = \lambda_1\|\boldsymbol{U}\|_{2,1} + \lambda_2\|\boldsymbol{U}\|_{G_{2,1}}$ denote the smooth component and non-smooth component of Eq. (4), respectively. $L^{(t)}$ denotes a stepsize at the t-iteration, which is determined by a iteratively increasing its value until the inequality(backtracking line search) $f(\boldsymbol{M}^{(t+1)}) \le f(\hat{\boldsymbol{M}}^{(t)}) + \langle \boldsymbol{M}^{(t+1)} - \hat{\boldsymbol{M}}^{(t)}, \nabla f(\hat{\boldsymbol{M}}^{(t)})\rangle + \frac{L}{2}\|\boldsymbol{M}^{(t+1)} - \hat{\boldsymbol{M}}^{(t)}\|^2$ is satisfied. The procedure in Eq. (5) is commonly referred to as the proximal operator [39]. The efficient computation of the proximal operator is critical for the practical convergence of APG, as it is involved in each iteration of the APG algorithm. For the optimization problem in Eq. (4), its proximal operator can be expressed as an optimization problem of the general form:

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{S}} \; & \tfrac{L}{2}\left\|\boldsymbol{U} - \tilde{\boldsymbol{U}}\right\|_F + \tfrac{L}{2}\left\|\boldsymbol{S} - \tilde{\boldsymbol{S}}\right\|_F + \mathcal{R}_{\lambda_2}^{\lambda_1}(\boldsymbol{U}) \\ \text{s.t.} \quad & \mathrm{Tr}(\boldsymbol{S}) = h, \boldsymbol{S} \preceq \boldsymbol{I}, \boldsymbol{S} \in \mathbb{R}_+^p, \end{aligned} \tag{6}$$

where $\tilde{\boldsymbol{U}} = \hat{\boldsymbol{U}}^{(t)} - \frac{1}{L^{(t)}}\nabla_{\hat{\boldsymbol{U}}} f(\hat{\boldsymbol{M}}^{(t)})$ and $\tilde{\boldsymbol{S}} = \hat{\boldsymbol{S}}^{(t)} - \frac{1}{L^{(t)}}\nabla_{\hat{\boldsymbol{S}}} f(\hat{\boldsymbol{M}}^{(t)})$, $\nabla_{\hat{\boldsymbol{U}}} f(\hat{\boldsymbol{M}}^{(t)})$ and $\nabla_{\hat{\boldsymbol{S}}} f(\hat{\boldsymbol{M}}^{(t)})$ denote the derivatives of $f(\hat{\boldsymbol{M}}^{(t)})$ with respect to $\hat{\boldsymbol{U}}$ and $\hat{\boldsymbol{S}}$. Note that $\nabla_{\hat{\boldsymbol{U}}} f(\hat{\boldsymbol{M}}^{(t)}) = (\boldsymbol{XX}^T + \lambda_3(\boldsymbol{I} - \hat{\boldsymbol{S}}^{(t)}))\hat{\boldsymbol{U}}^{(t)} - \boldsymbol{XY}^T$, and $\nabla_{\hat{\boldsymbol{S}}} f(\hat{\boldsymbol{M}}^{(t)})) = -\mathrm{Tr}(\hat{\boldsymbol{U}}^{(t)T}\hat{\boldsymbol{U}}^{(t)})$.

It can be easily verified that the optimization of $\boldsymbol{U}$ and $\boldsymbol{S}$ to (6) are decoupled, and can be obtained by solving two subproblems with respect to $\boldsymbol{U}$ and $\boldsymbol{S}$ by fixing one of them and solving the corresponding convex optimization problem as below. Moreover, the optimal solution to Eq. (6) admits an analytic form as presented below.

- Computation of $\boldsymbol{U}$ for a given $\boldsymbol{S}$

When we keep $\boldsymbol{S}$ fixed and seek the optimal $\boldsymbol{U}$, Eq. (4) becomes an unconstrained regularization problem:

$$\min_{\boldsymbol{U}} \tfrac{L}{2}\left\|\boldsymbol{U} - \tilde{\boldsymbol{U}}\right\|_F + \mathcal{R}_{\lambda_2}^{\lambda_1}(\boldsymbol{U}) \tag{7}$$

A key building block in optimizing $\boldsymbol{U}$ is the computation of the proximal operator $R_{\lambda_2}^{\lambda_1}(\boldsymbol{U})$ in (7), which is challenging to solve due to the presence of two non-smooth terms. While proximal operators for individual simple regularizers (e.g. ℓ_1-norm, ℓ_2-norm and group lasso) are possible, proximal operators for a conic combination of such regularizers are more challenging. The proximal operator $\mathbf{prox}_L^{\lambda_1,\lambda_2}(\hat{\boldsymbol{U}}^{(t)}) := T_L^{\lambda_1,\lambda_2}(\tilde{\boldsymbol{U}}^{(t)})$ exhibits a certain decomposition property, based on which we can efficiently compute in two steps, as outlined below:

$$\tilde{\boldsymbol{B}}^{(t)} = T_L^{\lambda_1}(\tilde{\boldsymbol{U}}^{(t)}) \;, \tag{8}$$

$$\boldsymbol{U}^{(t+1)} = T_L^{\lambda_2}(\tilde{\boldsymbol{B}}^{(t)}) = T_L^{\lambda_1,\lambda_2}(\tilde{\boldsymbol{U}}^{(t)}) \;. \tag{9}$$

Next we show that both of these steps can be executed efficiently using suitable extensions of soft-thresholding. The update in (8) can be written as:

$$\begin{aligned} \tilde{\boldsymbol{B}}^{(t)} &= T_L^{\lambda_1}(\tilde{\boldsymbol{U}}^{(t)}) \\ &= \arg\min_{\tilde{\boldsymbol{B}}\in\mathbb{R}^{p\times m}} \left\{\lambda_1\|\tilde{\boldsymbol{B}}\|_{2,1} + \frac{L^{(t)}}{2}\|\tilde{\boldsymbol{B}} - \tilde{\boldsymbol{U}}^{(t)}\|_F^2\right\}. \end{aligned} \tag{10}$$

Following [44], the row-wise updates can be done by soft-thresholding as:

$$\tilde{\boldsymbol{b}}_i^{(t)} = \frac{\max\{\|\tilde{\boldsymbol{u}}_i^{(t)}\|_2 - \frac{\lambda_1}{L^{(t)}}, 0\}}{\|\tilde{\boldsymbol{u}}_i^{(t)}\|_2}\tilde{\boldsymbol{u}}_i^{(t)} \;, \tag{11}$$

where $\tilde{\boldsymbol{b}}_i^{(t)}, \tilde{\boldsymbol{u}}_i^{(t)}$ are the i-th rows of $\tilde{\boldsymbol{B}}^{(t)}, \tilde{\boldsymbol{U}}^{(t)}$ respectively.

Next we focus on the update (9), which can be written as:

$$\begin{aligned} \boldsymbol{U}^{(t+1)} &= T_L^{\lambda_2}(\tilde{\boldsymbol{B}}^{(t)}) \\ &= \arg\min_{\boldsymbol{U}\in\mathbb{R}^{p\times m}} \left\{\lambda_2\|\boldsymbol{U}\|_{G_{2,1}} + \frac{L^{(t)}}{2}\|\boldsymbol{U} - \tilde{\boldsymbol{B}}^{(t)}\|_F^2\right\} \;. \end{aligned} \tag{12}$$

Following [45], the group specific row-wise updates can be done by soft-thresholding as:

$$\boldsymbol{U}_{R_j}^{(t+1)} = \frac{\max\{\|\tilde{\boldsymbol{B}}_{R_j}^{(t)}\|_F - \frac{\lambda_2}{L^{(t)}}, 0\}}{\|\tilde{\boldsymbol{B}}_{R_j}^{(t)}\|_F} \tilde{\boldsymbol{B}}_{R_j}^{(t)} , \tag{13}$$

where $\boldsymbol{U}_{R_j}, \tilde{\boldsymbol{B}}_{R_j}$ are group specific $p_j \times m$ sub-matrices correspond to group R_j in $\boldsymbol{U}_{R_j}, \tilde{\boldsymbol{B}}_{R_j}$ (p_j denotes the number of features in group R_j) respectively. The proximal operators of these two non-smooth regularizations admit closed form solutions, and thus both the steps (8) and (9) can be efficiently computed.

- Computation of $\boldsymbol{S}$ for a given $\boldsymbol{U}$

The optimal M to (6) can be obtained by solving

$$\begin{aligned} &\min_{\boldsymbol{S}} \quad \left\|\boldsymbol{S} - \tilde{\boldsymbol{S}}\right\|_F \\ &\text{s.t.} \quad Tr(\boldsymbol{S}) = h, \boldsymbol{S} \preceq \boldsymbol{I}, \boldsymbol{S} \in \mathbb{R}_+^p \end{aligned} \tag{14}$$

Let $\tilde{\boldsymbol{S}}^{(t)} = \boldsymbol{P}^{(t)} \Sigma \boldsymbol{P}^{(t)T}$ be its singular value decomposition (SVD), where $\boldsymbol{P}^{(t)} \in \mathbb{R}^{p\times p}$ is column-wise orthogonal, and $\operatorname{rank}(\tilde{\boldsymbol{S}}^{(t)}) = q$. $\Sigma = \operatorname{diag}(\tilde{\sigma}_i, \ldots, \tilde{\sigma}_p) \in \mathbb{R}^{p\times p}$ is diagonal with the eigenvalues on its main diagonal, where $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \geq \tilde{\sigma}_q > 0 = \tilde{\sigma}_{q+1} = \ldots = \tilde{\sigma}_p$. According to the theorem in [46], let $\Sigma^* = \operatorname{diag}(\sigma_1^*, \ldots, \sigma_q^*, \boldsymbol{0}) \in \mathbb{R}^{p\times p}$ where $\left\{\sigma_i^*\right\}_{i=1}^q$ is the optimal solution to the following optimization problem:

$$\min_{\{\sigma_i\}_{i=1}} \sum_{i=1}^{q} (\sigma_i - \tilde{\sigma}_i)^2, \quad \text{s.t.} \quad \sum_{i=1}^{q} (\sigma_i) = h, 0 \leq \sigma_i \leq 1 \tag{15}$$

Then, the global solution of Eq. (14) is given by $\boldsymbol{S}^* = \boldsymbol{P}^{(t)} \Sigma^* \boldsymbol{P}^{(t)T}$. The optimization problem in Eq. (15) can be solved via a linear time algorithm. In summary, we present the algorithm of optimizing Eq. (4) in Algorithm 1, the change of objective values in two successive steps is smaller than a prespecified value (e.g., 10^{-5}). The optimization can guarantee the global convergence according to Theorem 4.2. in [31].

Algorithm 1 AGP Method for optimizing SGLS-MTL.

```
Input: Training Data $X$ and $Y$, $L_0 \geq 0$, $a^{(1)} = 1$, $\lambda_1$, $\lambda_2$, and $\lambda_3$
Output: $\boldsymbol{M}$
Initialize $\hat{\boldsymbol{U}}^{(0)}$ and $\hat{\boldsymbol{S}}^{(0)}$
$t = 0$
repeat
    Find the smallest nonmajority integers $i_t$ such that with $f(\boldsymbol{M}^{(t+1)}) \leq f(\hat{\boldsymbol{M}}^{(t)}) + \langle \boldsymbol{M}^{(t+1)} - \hat{\boldsymbol{M}}^{(t)}, \nabla f(\hat{\boldsymbol{M}}^{(t)}) \rangle + \frac{L}{2}\|\boldsymbol{M}^{(t+1)} - \hat{\boldsymbol{M}}^{(t)}\|^2$ , and set $L^{(t)} = 2^{i_t} L^{(t-1)}$
    Optimize $\boldsymbol{U}^{(t+1)}$ with fixed $\hat{\boldsymbol{S}}^{(t)}$ according to Eq. (11) and (13)
    Optimize $\boldsymbol{S}^{(t+1)}$ with fixed $\boldsymbol{U}^{(t+1)}$ according to Eq. (15)
    $\boldsymbol{M}^{(t+1)} = \left[\boldsymbol{U}^{(t+1)} \boldsymbol{S}^{(t+1)}\right]$, $a^{(t+1)} = \frac{(1+\sqrt{1+4(a^{(t)})^2})}{2}$
    $\hat{\boldsymbol{M}}^{(t+1)} = \boldsymbol{M}^{(t+1)} + \frac{a^{(t+1)} - 1}{a^{(t+1)}} (\boldsymbol{M}^{(t+1)} - \boldsymbol{M}^{(t)})$
    $t = t + 1$
until convergence criterion is satisfied
if convergence then
    $\boldsymbol{M} = \boldsymbol{M}^{(t+1)}$
end if
```

2.4. The sparse group lasso with shared subspace learning for cluster analysis

In terms of the label availability, the machine learning methods can be broadly categorized into supervised methods and unsupervised methods. The proposed sparse group lasso with shared subspace learning method can be extended to the clustering when both the clinical label and cognitive scores are unknown. Clustering is a well-established machine learning methodology aimed at grouping examples (patients) so that instances in the same cluster are as similar as possible. To identify the subpopulations of the ADNI data that are homogeneous with respect to the MRI features, we present two specific clustering algorithms based on the sparse group lasso with shared subspace learning: the first one is shared Subspace Multi-task Clustering with Sparse Group Lasso (named SGLS-MTC); the other is Multi-Subspace (Manifold) single-task Clustering with Sparse Group Lasso (named SGL-MSC).

2.4.1. Sparse Group Lasso with shared Subspace based Multi-Task Clustering, SGLS-MTC

Suppose that we want to cluster $\boldsymbol{X}$ into l clusters $(C_1, C_2, \ldots, C_l)$ under the matrix factorization framework as:

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{P}} \|\boldsymbol{X}^T - \boldsymbol{W}\boldsymbol{P}^T\|_F^2 \\ &\text{ s.t. } \boldsymbol{P} \in \{0, 1\}^{n_k \times l} \ . \end{aligned} \tag{16}$$

where $\boldsymbol{W} \in \mathbb{R}^{p\times l}$ is the latent feature matrix, and $\boldsymbol{P} \in \mathbb{R}^{n\times l}$ is the cluster indicator, which represents the clustering assignment, such that $\boldsymbol{P}_{ij} = 1$ if x_i belongs to cluster C_j and $\boldsymbol{P}_{ij} = 0$ otherwise.

The problem in Eq. (16) is difficult to solve due to the constraint on $\boldsymbol{P}$ [55,56]. Following the common relaxation for label indicator matrix [57], Eq. (16) can be rewritten as:

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{P}} \|\boldsymbol{X}^T - \boldsymbol{W}\boldsymbol{P}^T\|_F^2 \\ &\text{ s.t. } \boldsymbol{P}^T\boldsymbol{P} = I, \boldsymbol{P} \geq 0 \ . \end{aligned} \tag{17}$$

Furthermore, according to Theorem 1 in [57], the orthogonality constraint on $\boldsymbol{P}$ is to allow us to perform feature selection via $\boldsymbol{W}$. Therefore, by incorporating the proposed hierarchical sparsity regularization into the clustering model, the sparse group lasso based clustering model (SGL-C) is formulated:

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{P}} \|\boldsymbol{X}^T - \boldsymbol{W}\boldsymbol{P}^T\|_F^2 + \lambda_1\|\boldsymbol{W}\|_{2,1} + \lambda_2\|\boldsymbol{W}\|_{G_{2,1}} \\ &\text{ s.t. } \boldsymbol{P}^T\boldsymbol{P} = I, \boldsymbol{P} \geq 0 \ . \end{aligned} \tag{18}$$

The SGL-C method can only deal with a single clustering task. When there are multiple different but related clustering tasks, i.e. $\boldsymbol{X} = \left\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(m)}\right\}$, where $\boldsymbol{X}^{(t)} \in \mathbb{R}^{n_t\times p}$ is the data in the t-th task, the clustering performance of multiple tasks can be improved by appropriately capturing their intrinsic relationship among different tasks. In our work, we assume that the multiple tasks share a subspace $\boldsymbol{Q} \in \mathbb{R}^{h\times p}$, where all the clustering tasks have similar data distribution and can be performed together.

Here we give the objective function of our shared subspace based multi-task clustering with sparse group lasso regularization:

$$\begin{aligned} \min_{\boldsymbol{Q},\hat{\boldsymbol{W}},\boldsymbol{W}^{(t)},\boldsymbol{P}^{(t)}} \ & \sum_{t=1}^{m} \|\boldsymbol{X}^{(t)T} - \boldsymbol{W}^{(t)}\boldsymbol{P}^{(t)T}\|_F^2 + \lambda_1 \sum_{t=1}^{m} \|\boldsymbol{Q}\boldsymbol{X}^{(t)T} - \hat{\boldsymbol{W}}\boldsymbol{P}^{(t)T}\|_F^2 \\ & + \lambda_2(\|\boldsymbol{W}^{(t)}\|_{2,1} + \|\hat{\boldsymbol{W}}\|_{2,1}) + \lambda_3(\|\boldsymbol{W}^{(t)}\|_{G_{2,1}} + \|\hat{\boldsymbol{W}}\|_{G_{2,1}}) \\ & \text{s.t. } \boldsymbol{Q}^T\boldsymbol{Q} = I, \boldsymbol{P}^{(t)T}\boldsymbol{P}^{(t)} = I, \boldsymbol{P}^{(t)} \geq 0, t = 1, \ldots, m \ , \end{aligned} \tag{19}$$

where $\hat{\boldsymbol{W}} \in \mathbb{R}^{h\times l}$ is the latent feature matrix in the shared subspace, and each item $\hat{\boldsymbol{W}}_j$ in $\hat{\boldsymbol{W}}$ is the mean of cluster C_j of all the tasks in the shared subspace.

The first and second terms are the clustering of each task in its input space and the multi-task clustering in the shared subspace, respectively. The last two terms are the proposed sparse group lasso regularization to facilitate the selection of feature and ROI for the latent feature matrix of each task $\boldsymbol{W}^{(t)}$ and the latent feature matrix $\hat{\boldsymbol{W}}$ in the shared subspace. Note that the method of multi-task clustering without the regularization of sparse group lasso is called MTC.

Then, we derive an efficient alternating algorithm to solve this problem of Eq. (19).

(1) Update of $\hat{\boldsymbol{W}}$

Given $\boldsymbol{Q}$ and $\boldsymbol{W}^{(t)}$, optimizing Eq. (19) with respect to $\hat{\boldsymbol{W}}$ is equivalent to the following:

$$\min_{\hat{\boldsymbol{W}}} \|\boldsymbol{Q}\boldsymbol{X}^T - \hat{\boldsymbol{W}}\boldsymbol{P}^T\|_F^2 + \lambda_2\|\hat{\boldsymbol{W}}\|_{2,1} + \lambda_3\|\hat{\boldsymbol{W}}\|_{G_{2,1}} \tag{20}$$

where $\boldsymbol{X} = \left[\boldsymbol{X}^{(1)}; \boldsymbol{X}^{(2)}; \ldots; \boldsymbol{X}^{(m)}\right] \in \mathbb{R}^{n\times p}$, $\boldsymbol{P} = \left\{\boldsymbol{P}^{(1)}; \ldots; \boldsymbol{P}^{(m)}\right\} \in \mathbb{R}^{n\times l}$.

Using the fact that $\boldsymbol{P}^T\boldsymbol{P} = I$, we can reformulate Eq. (20) as:

$$\min_{\hat{\boldsymbol{W}}} \|\hat{\boldsymbol{W}} - \boldsymbol{Q}\boldsymbol{X}^T\boldsymbol{P}\|_F^2 + \lambda_2\|\hat{\boldsymbol{W}}\|_{2,1} + \lambda_3\|\hat{\boldsymbol{W}}\|_{G_{2,1}} \tag{21}$$

The above equation can be composited by the introduced proximal composition algorithm in Eq. (9) and optimized by both the row-wise and group-wise soft-thresholding according to Eqs. (11) and (13).

(2) Update of $\boldsymbol{W}^{(t)}$

Given $\boldsymbol{P}^{(t)}$, optimizing Eq. (19) with respect to $\boldsymbol{W}^{(t)}$ is equivalent to optimizing:

$$\min_{\boldsymbol{W}^{(t)}} \sum_{t=1}^{m} \|\boldsymbol{X}^{(t)T} - \boldsymbol{W}^{(t)}\boldsymbol{P}^{(t)T}\|_F^2 + \lambda_2\|\boldsymbol{W}^{(t)}\|_{2,1} + \lambda_3\|\boldsymbol{W}^{(t)}\|_{G_{2,1}} \tag{22}$$

Due to $\boldsymbol{P}^{(t)T}\boldsymbol{P}^{(t)} = I$, we can reformulate (22) as:

$$\min_{\boldsymbol{W}^{(t)}} \sum_{t=1}^{m} \|\boldsymbol{W}^{(t)} - \boldsymbol{X}^{(t)T}\boldsymbol{P}^{(t)}\|_F^2 + \lambda_2\|\boldsymbol{W}^{(t)}\|_{2,1} + \lambda_3\|\boldsymbol{W}^{(t)}\|_{G_{2,1}} \tag{23}$$

The proximal operator for the composite regularizer can also be computed efficiently in two steps.

(3) Update of $\boldsymbol{P}^{(t)}$

Given $\boldsymbol{Q}\boldsymbol{W}^{(t)}$ and $\hat{\boldsymbol{W}}$ are fixed, we have

$$\begin{aligned} &\min_{\boldsymbol{P}^{(t)}} \|\boldsymbol{X}^{(t)T} - \boldsymbol{W}^{(t)}\boldsymbol{P}^{(t)T}\|_F^2 + \lambda_1\|\boldsymbol{Q}\boldsymbol{X}^{(t)T} - \hat{\boldsymbol{W}}\boldsymbol{P}^{(t)T}\|_F^2 \\ &\text{s.t. } \boldsymbol{P}^{(t)T}\boldsymbol{P}^{(t)} = I, \boldsymbol{P}^{(t)} \geq 0 \ . \end{aligned} \tag{24}$$

The Lagrangian function of Eq. (24) is:

$$\begin{aligned} L(\boldsymbol{P}^{(t)}) = \ & \|\boldsymbol{X}^{(t)T} - \boldsymbol{W}^{(t)}\boldsymbol{P}^{(t)T}\|_F^2 + \lambda_1\|\boldsymbol{Q}\boldsymbol{X}^{(t)T} - \hat{\boldsymbol{W}}\boldsymbol{P}^{(t)T}\|_F^2 \\ & -\mathrm{tr}(\boldsymbol{\alpha}(\boldsymbol{P}^{(t)T}\boldsymbol{P}^{(t)} - I)) - \mathrm{tr}(\boldsymbol{\beta}\boldsymbol{P}^{(t)}) \\ = \ & \mathrm{tr}(\boldsymbol{X}^{(t)}\boldsymbol{X}^{(t)T} - 2\boldsymbol{X}^{(t)}\boldsymbol{W}^{(t)}\boldsymbol{P}^{(t)T} + \boldsymbol{P}^{(t)}\boldsymbol{W}^{(t)T}\boldsymbol{W}^{(t)}\boldsymbol{P}^{(t)T}) \\ & +\lambda_1\mathrm{tr}(\boldsymbol{X}^{(t)}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{X}^{(t)T} - 2\boldsymbol{X}^{(t)}\boldsymbol{Q}^T\hat{\boldsymbol{W}}\boldsymbol{P}^{(t)T} + \boldsymbol{P}^{(t)}\hat{\boldsymbol{W}}^T\hat{\boldsymbol{W}}\boldsymbol{P}^{(t)T}) \\ & -\mathrm{tr}(\boldsymbol{\alpha}(\boldsymbol{P}^{(t)T}\boldsymbol{P}^{(t)} - I)) - \mathrm{tr}(\boldsymbol{\beta}\boldsymbol{P}^{(t)}) \end{aligned} \tag{25}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{l\times l}$ and $\boldsymbol{\beta} \in \mathbb{R}^{l\times n}$ are Lagrangian multipliers.

Setting $\frac{\partial^2 L(\boldsymbol{P}^{(t)})}{\partial \boldsymbol{P}^{(t)}} = 0$, we obtain:

$$\begin{aligned} &-2\boldsymbol{X}^{(t)}\boldsymbol{W}^{(t)} + 2\boldsymbol{P}^{(t)}\boldsymbol{W}^{(t)T}\boldsymbol{W}^{(t)} - \lambda_1(2\boldsymbol{X}^{(t)}\boldsymbol{Q}^T\hat{\boldsymbol{W}} - 2\boldsymbol{P}^{(t)}\hat{\boldsymbol{W}}^T\hat{\boldsymbol{W}}) \\ &\quad -2\boldsymbol{P}^{(t)}\boldsymbol{\alpha}^T - \boldsymbol{\beta}^T = 0 \end{aligned} \tag{26}$$

The above function equivalently becomes:

$$\boldsymbol{\beta}^T = -2\boldsymbol{\Theta}_1 + 2\boldsymbol{P}^{(t)}\boldsymbol{\Theta}_2 \tag{27}$$

where $\boldsymbol{\Theta}_1 = \boldsymbol{X}^{(t)}\boldsymbol{W}^{(t)} + \lambda_1\boldsymbol{X}^{(t)}\boldsymbol{Q}^T\hat{\boldsymbol{W}}$ and $\boldsymbol{\Theta}_2 = \boldsymbol{W}^{(t)T}\boldsymbol{W}^{(t)} + \lambda_1\hat{\boldsymbol{W}}^T\hat{\boldsymbol{W}} - \boldsymbol{\alpha}$.

According to the Karush-Kuhn-Tucker condition, $\boldsymbol{\beta}_{ij}^T\boldsymbol{P}_{ij}^{(t)} = 0$, we get:

$$\left[-\boldsymbol{\Theta}_1 + \boldsymbol{P}^{(t)}\boldsymbol{\Theta}_2\right]_{ij}\boldsymbol{P}_{ij}^{(t)} = 0 \tag{28}$$

Introduce $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_1^+ - \boldsymbol{\Theta}_1^-$ and $\boldsymbol{\Theta}_2 = \boldsymbol{\Theta}_2^+ - \boldsymbol{\Theta}_2^-$ where $\boldsymbol{\Theta}_1^+(i,j) = \frac{|\boldsymbol{\Theta}_1(i,j)| + \boldsymbol{\Theta}_1(i,j)}{2}$ and $\boldsymbol{\Theta}_1^-(i,j) = \frac{|\boldsymbol{\Theta}_2(i,j)| - \boldsymbol{\Theta}_2(i,j)}{2}$ [58], we obtain:

$$\left[\boldsymbol{\Theta}_1^- + \boldsymbol{P}^{(t)}\boldsymbol{\Theta}_2^+ - \boldsymbol{\Theta}_1^+ - \boldsymbol{P}^{(t)}\boldsymbol{\Theta}_2^-\right]_{ij}\boldsymbol{P}_{ij}^{(t)} = 0 \tag{29}$$

Eq. (29) leads to the following updating formula:

$$\boldsymbol{P}_{ij}^{(t)} \leftarrow \boldsymbol{P}_{ij}^{(t)}\sqrt{\frac{\left[\boldsymbol{\Theta}_1^+ + \boldsymbol{P}^{(t)}\boldsymbol{\Theta}_2^-\right]}{\left[\boldsymbol{\Theta}_1^- + \boldsymbol{P}^{(t)}\boldsymbol{\Theta}_2^+\right]}} \tag{30}$$

(4) Update of $\boldsymbol{Q}$

Optimizing Eq. (19) with respect to $\boldsymbol{Q}$ yields the equation:

$$\begin{aligned} &\min_{\boldsymbol{Q}} \sum_{t=1}^{m} \|\boldsymbol{Q}\boldsymbol{X}^{(t)T} - \hat{\boldsymbol{W}}\boldsymbol{P}^{(t)T}\|_F^2 \ , \\ &\text{s.t. } \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I} \ . \end{aligned} \tag{31}$$

which results in the following problem:

$$\begin{aligned} &\min_{\boldsymbol{Q}}\|\boldsymbol{Q} - \hat{\boldsymbol{W}}\boldsymbol{P}^T(\boldsymbol{X}^T)^{-1}\|_F^2 \\ &\text{s.t. } \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}. \end{aligned} \tag{32}$$

We can further write the above equation into a more compact form as:

$$\min_{\boldsymbol{Q}^T\boldsymbol{Q}=\boldsymbol{I}} \|\boldsymbol{Q} - \boldsymbol{\Omega}\|_F^2 \tag{33}$$

where $\boldsymbol{\Omega} = \hat{\boldsymbol{W}}\boldsymbol{P}^T(\boldsymbol{X}^T)^{-1}$.

It can be solved using the lemma in [59]. Given the objective function with respect to $\boldsymbol{Q}$ in Eq. (33), the optimal $\boldsymbol{Q}^*$ is defined as: $\boldsymbol{Q}^* = \boldsymbol{\Omega}_l\boldsymbol{\Omega}_r^T$, where $\boldsymbol{\Omega}_l$ and $\boldsymbol{\Omega}_r$ are the left and right singular vectors of the singular value decomposition (SVD) of $\boldsymbol{\Omega}$.

The pseudo code of SGLS-MTC is summarized in Algorithm 2.

Algorithm 2 The optimization of SGLS-MTC.

```
Input: Training Data $\left\{\boldsymbol{X}^{(t)}\right\}_{k=1}^{m}$, regularization parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and the amount of cluster $l$
Output: The partitions of data $\boldsymbol{P}^{(t)}, t = 1, \ldots, m$
Initialize $\boldsymbol{P}^{(t)}$ using K-means, $t = 1, \ldots, m$;
Initialize $\boldsymbol{W}^{(0)}$;
repeat
    Update $\hat{\boldsymbol{W}}$ according to Eq. (11) and Eq. (13)
    for $t = 1$ to $m$ do
        Update $\boldsymbol{W}^{(t)}$ according to Eq. (11) and Eq. (13)
        Update $\boldsymbol{P}^{(t)}$ according to Eq. (30)
    end for
    Update $\boldsymbol{Q}$ according to Eq. (33)
until convergence criterion is satisfied
```

2.4.2. Multi-Subspace(Manifold) single-task Clustering with Sparse Group Lasso, SGL-MSC

Subspace clustering is an important unsupervised learning research topic. By modeling the distribution of data as a union of subspaces multiple subspace models improve on the single subspace assumption [63]. In many real-world problems, however, the data points lie in multiple subspaces (manifold) and the membership of the data points to the subspaces might be unknown [55]. Therefore, there is a need to simultaneously cluster the data into multiple subspaces and find a low-dimensional subspace fitting each group of points. In the case of linear manifolds, there are many existing subspace clustering methods including K-subspaces [61], and Generalized Principal Component Analysis (GPCA) [62], and Stable Subspace (SSS) [63]. However, all subspace clustering methods are formulated only for mixtures of linear manifolds and do not work in the presence of nonlinear manifolds.

Locally Linear Embedding (LLE) [64] is simple nonlinear dimensionality reduction method, and it assumes that the data lies on a smooth nonlinear manifold of dimensionality $h < p$. LLE exploits the fact that the local neighborhood of a point on the manifold can be well approximated by the affine subspace spanned by its k

nearest neighbors. It has been applied in transforming multivariate MRI data of regional brain volume and cortical thickness to a locally linear space with fewer dimensions [65]. The procedure of LLE can be summarized as follows:

1. Select k-nearest neighbors of each data points $\boldsymbol{x}_i$ using Euclidean distances.
2. Calculate the reconstructing weight matrix $\boldsymbol{C} = \left[c_{ij}\right]_{n\times n}$, which reconstructs each point x_i from its k-nearest neighbors. This $\boldsymbol{C}$ can be viewed as similarity between data points or the edge weights of a graph whose nodes are the data points.

$$\min J(\boldsymbol{C}) \;= \sum_{i=0}^{n} \left\| x_i - \sum_{j=0}^{n} c_{ij} x_i \right\|^2 \tag{34}$$

3. Reconstruct represented $\boldsymbol{z}_i = \boldsymbol{x}_i \boldsymbol{Q}^T$ by learning a projection matrix $\boldsymbol{Q} \in \mathbb{R}^{h\times p}$. To maintain the intrinsic geometrical feature of the data after the embedding process, the reconstruction error function must be minimized:

$$\begin{aligned} \min J(\boldsymbol{Z}) \;&= \sum_{i=0}^{n} \left\| z_i - \sum_{j=0}^{n} c_{ij} z_i \right\|^2 = \operatorname{tr}(\boldsymbol{X}\boldsymbol{Q}^T\boldsymbol{M}\boldsymbol{Q}\boldsymbol{X}^T) \\ &\text{s.t. } \boldsymbol{X}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{X}^T = I \end{aligned} \tag{35}$$

where $\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{C})^T(\boldsymbol{I} - \boldsymbol{C})$.

LLE is not designed to deal with the data that are disconnected, i.e. separated into groups. Polito proposed a variant of LLE for the purpose of clustering data living on different manifolds [60]. The algorithm can simultaneously group the data and calculate local embedding of each group. For a union of separated manifolds, the LLE algorithm computes a matrix whose null space contains vectors giving the segmentation of the data. However, according to the Proposition 2 in [60], there exist l vectors $\left\{\boldsymbol{v}^l_{j=1}\right\}$ in the $\boldsymbol{M}$ constructed from the local geometry of the manifold such that $\boldsymbol{v}_j$ corresponds to the jth group of points, i.e. $\boldsymbol{v}_{ij} = 1$ if the ith data point is in the j-th group, and $\boldsymbol{v}_{ij} = 0$ otherwise. However, these vectors are not the only vectors (the embedding vectors and membership vectors) in the null space of $\boldsymbol{M}$ and spectral clustering is not directly applicable. Goh et al. presented an algorithm to allows to distinguish the membership vectors from other vectors in the null space by analyzing of the variance of these vectors. Thanks to Proposition 5 in [66], the membership eigenvectors can be computed as $\boldsymbol{v} = \boldsymbol{B}\boldsymbol{\Gamma}^{-\frac{1}{2}}\gamma_i$, where $\boldsymbol{B}$ is a basis for the null space of $\boldsymbol{M}$, γ_i are the eigenvectors of $\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{B}^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{J})\boldsymbol{B}\boldsymbol{\Gamma}^{-\frac{1}{2}}$ associated with its smallest l eigenvalues and $\boldsymbol{\Gamma} = \boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B}$. Note that $\boldsymbol{J} = \frac{1}{n}\boldsymbol{1}_{n\times 1}\boldsymbol{1}^T_{n\times 1}$. By identifying the membership vectors in $\boldsymbol{M}$, the algorithm proposed in [66] allows simultaneous nonlinear dimensionality reduction and manifold clustering.

However, there are two significant problems in the joint nonlinear dimensionality reduction and manifold clustering: (1) the high dimensional MRI features with irrelevancy and redundancy may negatively influence the clustering performance and nonlinear dimensionality reduction; (2) the learned projection $\boldsymbol{Q}$ in Eq. (35) is a linear combination of all the original features, thus it is often difficult to interpret the results. In our work, we incorporate the sparse learning into the multi-subspace clustering by using the proposed regularization of sparse group lasso on the projection matrix $\boldsymbol{Q}$, which leads to both the row-sparsity and group-sparsity of the projection matrix.

By incorporating the sparse group lasso into the objective function of learning of the projection matrix $\boldsymbol{Q}$ in Eq. (35) function, we have:

$$\begin{aligned} &\min_{\boldsymbol{Q}} \operatorname{tr}(\boldsymbol{X}\boldsymbol{Q}^T\boldsymbol{M}\boldsymbol{Q}\boldsymbol{X}^T) + \lambda_1\|\boldsymbol{Q}\|_{2,1} + \lambda_2\|\boldsymbol{Q}\|_{G_{2,1}} \\ &\;\text{s.t. } \operatorname{tr}(\boldsymbol{X}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{X}^T) = I \end{aligned} \tag{36}$$

According to Theorem 1 in [67], the $\boldsymbol{Q}$ can be obtained through the following two steps:

1. Solve the eigen-problem in Eq. (37) to get $\boldsymbol{Z}$;

$$\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{\Lambda}\boldsymbol{Z} \tag{37}$$

2. Obtain a $\boldsymbol{Q}$ in Eq. (38);

$$\min\ (\boldsymbol{X}\boldsymbol{Q}^T - \boldsymbol{Z}) + \lambda_1\|\boldsymbol{Q}\|_{2,1} + \lambda_2\|\boldsymbol{Q}\|_{G_{2,1}}, \tag{38}$$

The above equation can be composited by the introduced proximal composition algorithm in Eq. (9) and optimized by both the row-wise and group-wise soft-thresholding according to Eqs. (11) and (13).

Based on the multi-subspace clustering and sparse group lasso regularization, we propose to iteratively repeat the two mains steps (clustering and representation learning) to progressively improve the clustering results. In the representation learning, the dimensionality reduction and feature selection can be jointly performed according to Eqs. (37) and (38). Here we use the selected features by sparse group lasso to reduce the irrelevant features, which in turn leads to a better clustering results. This is repeated until the process converges (details are given in Algorithm 3).

Algorithm 3 The optimization of SGL-MSC.

```
Input: Training Data $X$, regularization parameters $\lambda_1, \lambda_2$ and the amount of cluster $l$.
Output: The partitions of data $P$, and lower dimensional representation $Z$
Initialization: Feature selected indicator $\hat{I} = [1, 1, \ldots, 1]$
repeat
    Construct the dataset $X' = X \odot \hat{I}$ by feature selecting with selected feature indicator $\hat{I}$. (Note that $X' = X \odot \hat{I}$ denotes $x'_{i,j} = x_{i,j}\hat{I}_j$, for all $i, j$.)
    Apply the LLE algorithm on $X'$ to obtain the matrix $M$.
    Compute a basis $B$ for the null space of $M$.
    Compute the matrix $Q = B^T B$ and obtain the membership eigenvectors by solving a generalized eigenvalue problem according to the Proposition 5 in [71]
    Apply K-means to the rows of the matrix of membership vectors to cluster the data into $m$ different groups, and obtain the partitions of data $P$.
    for $j = 1$ to $l$ do
        Obtain the selected feature indicator $\hat{I}$ and the lower dimensional representation $Z$ according to solve the objective function of Eqs. (37) and (38) in $j$-th group
    end for
    Obtain a common selected feature indicator $\hat{I}$ by intersecting each selected feature indicator of the multiple groups
until Convergence.
```

3. Experiment

3.1. Experiment I: Regression for baseline cognitive measures

Five sets of baseline cognitive scores widely used clinical/cognitive assessment scores [47,51,53] were employed in this study, including Alzheimer's Disease Assessment Scale cognitive total score (ADAS), Mini Mental State Exam score (MMSE), Rey Auditory Verbal Learning Test (RAVLT) involving total score (TOTAL),

Table 3
Comparison of root mean squared error (RMSE) of baseline methods and SGLS-MTL across all tasks (Note that $\star$ stands for the case with $p \leq 0.05$).

	Methods	ADAS	MMSE	RAVLT		
				TOTAL	T30	RECOG
Single task learning	Linear	$1.0461 \pm 0.0436^\star$	$1.1682 \pm 0.0798^\star$	$0.8471 \pm 0.0441^\star$	$1.1615 \pm 0.0195^\star$	$0.9064 \pm 0.0569^\star$
	Ridge	$0.7839 \pm 0.0584^\star$	$0.8264 \pm 0.0290^\star$	$0.8556 \pm 0.0444^\star$	$0.8638 \pm 0.0477^\star$	$0.9167 \pm 0.0575^\star$
	Lasso	$0.7861 \pm 0.0496^\star$	$0.8391 \pm 0.0386^\star$	$0.8526 \pm 0.0466^\star$	$0.8627 \pm 0.0521^\star$	$0.9130 \pm 0.0509^\star$
	Group Lasso	$0.7872 \pm 0.0552^\star$	$0.8354 \pm 0.0327^\star$	$0.8529 \pm 0.0435^\star$	$0.8611 \pm 0.0415^\star$	$0.9189 \pm 0.0561^\star$
Multi-task learning	MTL	$0.7853 \pm 0.0410^\star$	$0.8301 \pm 0.0442^\star$	$0.8510 \pm 0.0568^\star$	$0.8576 \pm 0.0428^\star$	$0.9110 \pm 0.0404^\star$
	SGL-MTL	$0.7762 \pm 0.0458^\star$	$0.8270 \pm 0.0392^\star$	$0.8389 \pm 0.0536^\star$	0.8594 ± 0.0440	$\mathbf{0.9049 \pm 0.0395}$
	S-MTL	$0.7745 \pm 0.0573^\star$	$0.8290 \pm 0.0327^\star$	$0.8484 \pm 0.0454^\star$	$0.8528 \pm 0.0429^\star$	$0.9113 \pm 0.0582^\star$
	SGLS-MTL	$\mathbf{0.7596 \pm 0.0567}$	$\mathbf{0.8148 \pm 0.0317}$	$\mathbf{0.8117 \pm 0.0438}$	$\mathbf{0.8387 \pm 0.0430}$	0.9078 ± 0.0584
Multi-task learning	SMML	$0.7631 \pm 0.0452^\star$	$0.8277 \pm 0.0414^\star$	$0.8526 \pm 0.0505^\star$	0.8498 ± 0.0417	$0.9172 \pm 0.0411^\star$
	Robust-MTL	$0.7819 \pm 0.0549^\star$	$0.8257 \pm 0.0399^\star$	$0.8518 \pm 0.0622^\star$	$0.8679 \pm 0.0464^\star$	$0.9155 \pm 0.0437^\star$
	MTFL	$0.7794 \pm 0.0451^\star$	$0.8495 \pm 0.0308^\star$	$0.8508 \pm 0.0466^\star$	$0.8659 \pm 0.0440^\star$	$0.9173 \pm 0.0476^\star$
	Group-MTL	$0.7724 \pm 0.0507^\star$	$0.8208 \pm 0.0411^\star$	$0.8334 \pm 0.0417^\star$	$0.8606 \pm 0.0312^\star$	0.9109 ± 0.0512

RAVLT 30 minutes delay score (T30) and RAVLT recognition score (RECOG). We evaluated and compared all the methods with RMSE (root mean square error) and CC(Pearson correlation coefficient) between the actual and predicted scores of all the test subjects. The average and standard deviation of performance measures are calculated by 5 cross-validation. To show the superior performance of our algorithm, we selected several state-of-the-art or baseline methods for comparison:

(1) Single task learning: Ridge regression, Lasso, Group Lasso applied independently to each task;
(2) Basic multi-task learning: multi-task learning based on $\ell_{2,1}$-norm regularization(MTL), multi-task learning with shared subspace(S-MTL), multi-task group lasso based on $\ell_{2,1}$-norm regularization (SGL-MTL). For S-MTL, we use the implementation of cASO [31] provided in MALSAR package [52].
(3) The state-of-the-art multi-task learning:

SMML: Sparse Multimodal Learning [32], which takes into account coupled feature and group sparsity across tasks. The code was taken from the author's homepage: http://ranger.uta.edu/~heng/imaging-genetics/.

Group-MTL [48]: groups of related tasks are assumed and tasks belonging to the same group share a common feature representation. The code was taken from the author's homepage: http://www-scf.usc.edu/~zkang/GoupMTLCode.zip.

Sparse-LowRank MTL [49] (named Robust-MTL): it captures the task relationships using a low-rank structure, and simultaneously identifies the outlier tasks using a group-sparse structure.

MTFL [50]: It employs an $\ell_{2,1}$-norm regularization term to capture the task relationship from multiple related tasks constraining all models to share a common set of features. We used the code provided in MALSAR package [52] for Sparse-LowRank MTL and MTFL.

Regularization parameters for all the methods are using a nested cross-validation strategy on the training data. For each of the 5 trials, an internal 5-fold cross-validation is performed to optimize the parameters within the training data. The regularization parameters of λ and the lower-dimensionality parameter h are chosen by nested cross-validation strategy on the training data (trying values $10^{-2}, 10^{-1}, \ldots, 10^{2}, 10^{3}$ for γ and $20, 40, \ldots, 200$ for h) in this study. It is worth noting that we use the same training and testing data in each trial for all the comparable methods for fair comparison.

Experimental results are reported in Table 3 where the best results are boldfaced. A first glance at the results shows that SGLS-MTL generally outperforms all other compared methods on both metrics and across all the cognitive tasks. Additionally, a statistical analysis is performed on the results and reported in Table 3. As can be seen, our proposed method achieves statistically significant results compared to all other methods on most of the results. These results reveal several interesting points:

(1) Compared with 4 single task learning, all the multi-task learning methods with different assumptions improve the prediction performance by utilizing different intrinsic relationships among multiple related tasks.
(2) The results show that sparse learning methods (Lasso, Group Lasso, MTL, SGL-MTL, and SGLS-MTL) are more effective than ridge regression and Linear on most tasks, which demonstrates that sparsity can lead to better generalization. Moreover, the sparse learning methods of SGL-MTL and SGLS-MTL show superiority over the computing non-sparse multi-task learning involving Group-MTL, Robust-MTL.
(3) When comparing the performances summarized in Tables 3 and 4, we find that SGLS-MTL, on average, increases the regression performance by 1.79%/1.87% and 2%/3.47% compared to the intermediate methods: GLS-MTL and S-MTL in terms of RMSE/CC. The results demonstrate that the integration of two shared structure regularization can facilitate the prediction performance of the multi-task learning with only each shared structure regularization.
(4) Compared with MTL with only $\ell_{2,1}$-norm, SGL-MTL with both $\ell_{2,1}$ and $G_{2,1}$ obtains a better performance for all regression tasks. This observation verify the benefits of encouraging sparsity of group level.
(5) Compared with the computing multi-task learning with other task relatedness assumption, the performance enhancements by our method were 1.8%/3.5% (vs. SMML) and 2.6%/3.2% (vs. Robust-MTL), and 3%/4.2% (vs. MTFL) and 1.6%/3.1% (vs. Group-MTL) in terms of RMSE and CC. Moreover, the formulation of SGL-MTL and SMML is same, only difference of them is the optimization algorithm. Compared with the approximate gradient descent method used by SMML, the accelerated proximal gradient leads to a fast and correct algorithm for the optimization.

The good performance of SGLS-MTL for modeling Alzheimer's Disease can be attributed to the appealing property that it can select features a common set of biomarkers (features and ROIs) using the sparse group Lasso penalty across the whole feature space while simultaneously considering the correlation of multiple tasks by exploring the shared feature subspace.

Fig. 2 shows the histogram of regression weights of each method for ADAS score predicting. From these plots, we observe Ridge regression and S-MTL produced non-sparse results that are not appropriate for biomarker discovery while Lasso, MTL, SGL-MTL and SGLS-MTL presented a much better sparsity across all the cortical measures, where only a small portion of the ROI-based morphological features was identified to be relevant to the out-

Table 4
Comparison of correlation coefficient (CC) of baseline methods and SGLS-MTL across all tasks (Note that * stands for the case with $p \leq 0.05$).

	Methods	ADAS	MMSE	RAVLT		
				TOTAL	T30	RECOG
Single task learning	Linear	$0.4388 \pm 0.0274^{\star}$	$0.3291 \pm 0.0435^{\star}$	$0.3322 \pm 0.0611^{\star}$	$0.3164 \pm 0.0723^{\star}$	$0.1930 \pm 0.0560^{\star}$
	Ridge	$0.6174 \pm 0.0349^{\star}$	$0.5564 \pm 0.0478^{\star}$	$0.5111 \pm 0.0747^{\star}$	$0.4984 \pm 0.0935^{\star}$	$0.3914 \pm 0.0633^{\star}$
	Lasso	$0.6222 \pm 0.0347^{\star}$	$0.5412 \pm 0.0473^{\star}$	$0.5277 \pm 0.0784^{\star}$	$0.5137 \pm 0.1059^{\star}$	$0.4072 \pm 0.0552^{\star}$
	Group Lasso	$0.6184 \pm 0.0382^{\star}$	$0.5479 \pm 0.0532^{\star}$	$0.5222 \pm 0.070^{\star}$	$0.5152 \pm 0.0987^{\star}$	$0.3938 \pm 0.0694^{\star}$
Multi-task learning	MTL	$0.6233 \pm 0.0323^{\star}$	$0.5598 \pm 0.0398^{\star}$	$0.5247 \pm 0.0365^{\star}$	$0.5173 \pm 0.0948^{\star}$	$0.4108 \pm 0.0165^{\star}$
	SGL-MTL	$0.6345 \pm 0.0392^{\star}$	$0.5778 \pm 0.0349^{\star}$	$0.5436 \pm 0.0364^{\star}$	$0.5310 \pm 0.0970^{\star}$	0.4237 ± 0.0248
	S-MTL	$0.6328 \pm 0.0301^{\star}$	$0.5605 \pm 0.0465^{\star}$	$0.5281 \pm 0.0723^{\star}$	$0.5267 \pm 0.1036^{\star}$	0.4204 ± 0.0696
	SGLS-MTL	$\mathbf{0.6472 \pm 0.0308}$	$\mathbf{0.5879 \pm 0.0417}$	$\mathbf{0.5468 \pm 0.0803}$	$\mathbf{0.5446 \pm 0.0899}$	$\mathbf{0.4275 \pm 0.0793}$
Multi-task learning	SMML	$0.6371 \pm 0.0381^{\star}$	$0.5551 \pm 0.0307^{\star}$	$0.5370 \pm 0.0467^{\star}$	$0.5224 \pm 0.0414^{\star}$	$0.4174 \pm 0.0473^{\star}$
	Robust-MTL	$0.6281 \pm 0.0311^{\star}$	$0.5671 \pm 0.0309^{\star}$	$0.5307 \pm 0.0479^{\star}$	$0.5293 \pm 0.0623^{\star}$	$0.4198 \pm 0.0403^{\star}$
	MTFL	$0.6277 \pm 0.0362^{\star}$	$0.5612 \pm 0.0347^{\star}$	$0.5273 \pm 0.0417^{\star}$	$0.5279 \pm 0.0911^{\star}$	$0.4071 \pm 0.0145^{\star}$
	Group-MTL	$0.6411 \pm 0.0326^{\star}$	$0.5530 \pm 0.0329^{\star}$	$0.5329 \pm 0.0481^{\star}$	$0.5290 \pm 0.0434^{\star}$	0.4219 ± 0.0425

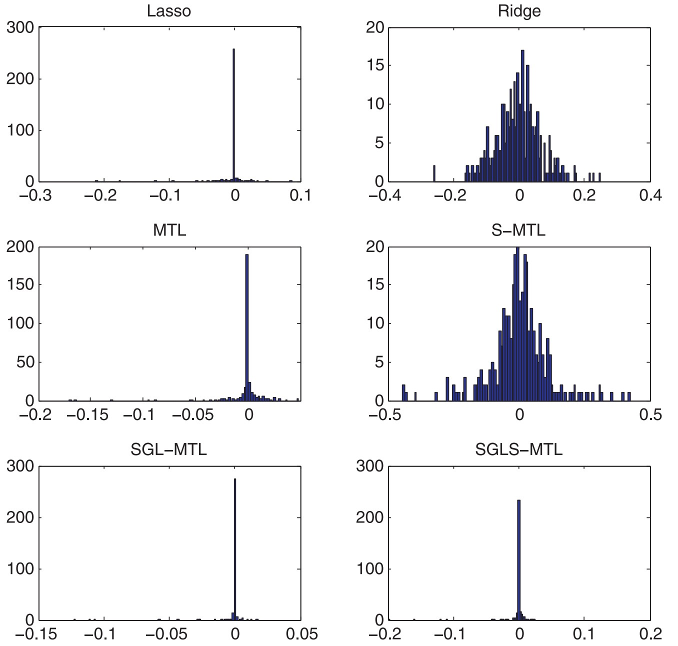

Fig. 2. Histogram of regression weights of all cortical measures for predicting the ADAS score.

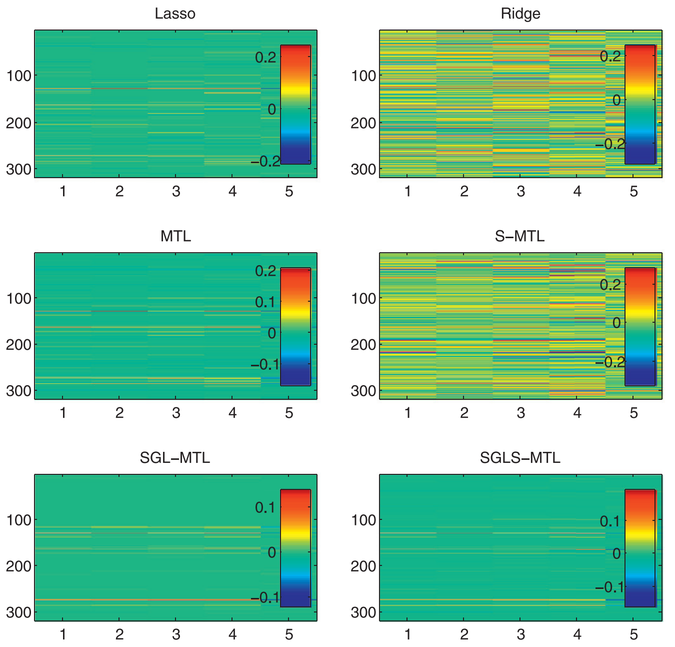

Fig. 3. The learned matrix of multiple methods on learning 5 tasks from feature level. Red indicates positive correlation while blue indicates negative correlation. The bigger the magnitude of an coefficient is, the more important its MRI measure is in predicting the corresponding cognitive score.

come. On the other hand, Figs. 3 and 4 illustrate the learned matrix of the comparable methods. From Fig. 3 we can observe the intrinsic sparse structural horizontal pattern of $\ell_{2,1}$-norm clearly, that is, selecting features for all the tasks. Additionally, Ridge regression and S-MTL yielded non-sparse results, making the result hard to interpret. Fig. 4 shows the group sparsity of the task parameters of multiple methods. The weight of each group j is calculated by $\sum_{j=1}^{k} w_j \sqrt{\sum_{i \in G_j} \|\boldsymbol{U}_i\|_2^2}$, and each row indicates a group (ROI). We only listed the 68 cortical regions, each of which involves 4 features as a group. Both of SGL-MTL and SGLS-MTL with group sparsity $G_{2,1}$ are able to identify a underlying compact set of relevant MRI biomarkers related to cognitive status, and we can find the both methods present a much better sparsity across all the cortical measures from group (ROI) level. Furthermore, from Figs. 3 and 4 we have the following observations: (1) SGL-MTL and SGLS-MTL have both sparsity at the feature level and group level since they take into account the sparsity of feature level and group level simultaneously; (2) SGL-MTL has more sparser than SGLS-MTL since SGLS-MTL also consider the latent subspace at the same time, which can influence the sparsity pattern on the original sparse.

3.2. Experiment II: jointly regression and classification

It is known that there exist inherent correlations between the disease diagnosis and clinical score prediction [21,24]. For better understanding of the underlying mechanism of AD, we tackle the heterogeneous tasks simultaneously in a unified framework to explore whether incorporating the classification tasks can improve the prediction performance of cognitive outcomes.

In this experiment, we perform the three binary classification tasks for the diagnostic groups respectively: AD, MCI(Mild Cognitive Impairment), and NC (Normal Control). We represented the class labels using a "1-of-m" encoding vector [29] $\boldsymbol{y}_i = [y^{(1)}, y^{(2)}, \ldots, y^{(m)}]^T$ such that $y^{(c)} = 1$ if $\boldsymbol{x}$ belongs to class c and $y^{(c)} = 0$ otherwise. The loss function of classification task is: $\mathcal{L}_C(\boldsymbol{W}; \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \left[\log \sum_{c=1}^{m} \exp(\boldsymbol{x}_i \boldsymbol{w}_c) - \sum_{c=1}^{m} y_i^{(c)} \boldsymbol{x}_i \boldsymbol{w}_c \right]$, and the overall loss function in Eq. (3) is $\mathcal{L}_R + \mathcal{L}_C$. The approach for jointly learning regression and classification problem is called SGLS-MTL-J. The performance of only regression task and heterogeneous tasks for our MTL model with respect to RMSE and CC is shown in Table 5.

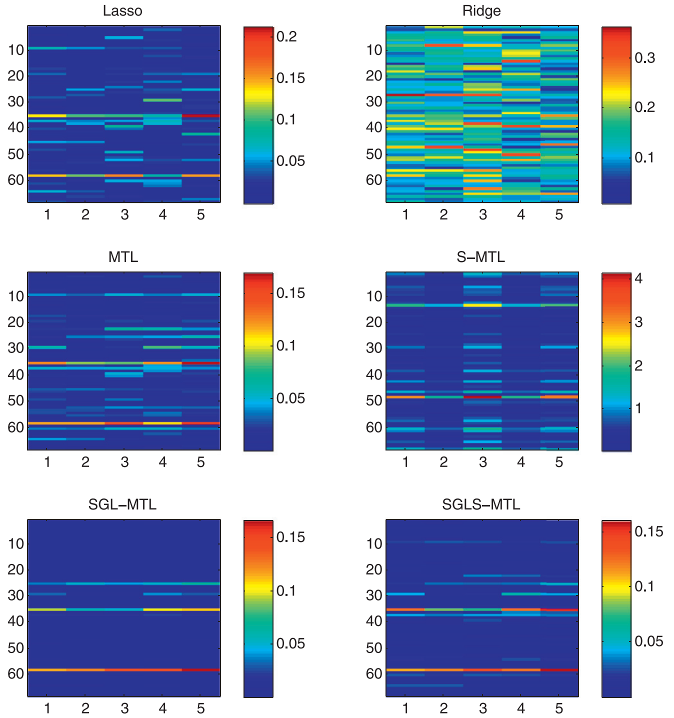

Fig. 4. The learned matrix of multiple methods on learning 5 tasks from group level.

Table 5
The results of only regression task and heterogeneous tasks for our MTL model with respect to RMSE and CC (Note that * stands for the case with $p \le 0.05$).

Metric	Score	SGLS-MTL	SGLS-MTL-J
RMSE	ADAS	0.7696 ± 0.0567*	**0.7447 ± 0.0601**
RMSE	MMSE	0.8148 ± 0.0317	**0.8025 ± 0.0395**
RMSE	TOTAL	0.8278 ± 0.0438	**0.8226 ± 0.0539**
RMSE	T30	0.8487 ± 0.0430*	**0.8219 ± 0.0714**
RMSE	RECOG	0.9078 ± 0.0584*	**0.8818 ± 0.0557**
CC	ADAS	0.6472 ± 0.0308	**0.6508 ± 0.0392**
CC	MMSE	0.5679 ± 0.0417	**0.5687 ± 0.0441**
CC	TOTAL	0.5468 ± 0.0803	**0.5476 ± 0.0589**
CC	T30	0.5346 ± 0.0199*	**0.5518 ± 0.0337**
CC	RECOG	0.4275 ± 0.0793*	**0.4418 ± 0.0326**

From the results of Table 5, SGLS-MTL-J consistently improved the performance of SGLS-MTL in all the test cases, which verifies the benefits of jointly learning from the heterogeneous tasks and implies that the features, ROIs and latent subspace used for these tasks were highly correlated. Moreover, the identified biomarkers will be correlated to memory scores and also be discriminative to disease categories as a result, the results will be shown in next subsection.

3.3. Experiment III: identify MRI Biomarker with stability selection

In Alzheimer's disease studies, researchers are not only interested in providing better cognitive scores prediction, but mainly to identify which are the brain areas more affected by the disease, which can help to diagnose early stages of the disease and how it spreads. One of the strengths of the SGLS-MTL formulation is that it facilitates the identification of biomarkers due to its sparse property.

In order to identify which areas of the brain region are closely related to cognitive measures, we conduct a further experiment to select the most discriminative features and ROIs. To find stable biomarkers (features and ROIs) with the SGLS-MTL model, a stability selection procedure is applied [54]. Stability selection [68], based on subsampling/bootstrapping, provides a general method to perform model selection using information from a set of regularization parameters. We propose to extend the idea of stability selection for feature and ROI based on multi-task learning. The stability score (between 0 and 1) of each feature or ROI is indicative of the importance of the specific feature or ROI for multiple cognitive outcomes prediction tasks. We calculated the stability score of each feature and ROI selected by SGLS-MTL for 5 tasks at the same time. Moreover, for obtaining the stable ROI, we computed the average of probabilities of the features belonging to each ROI G_j; $j = 1, \ldots, k$.

Top 20 selected ROIs (cortical regions) and features (from cortical and subcortical regions) by our SGLS-MTL with stability selection method on regression task are shown in Table 6. We found that the imaging biomarkers identified by SGLS-MTL yielded promising patterns that are expected from prior knowledge on neuroimaging and cognition. Some important features are selected, such as Hippocampus, Entorhinal, Middle Temporal Gyri and Fusiform, are highly relevant to the cognitive impairment. The brain regions that were selected for cognitive performance prediction, as well as the heterogeneous tasks for cognitive performance prediction and diagnosis are consistent with results reported in previous studies [47,53], which demonstrates the effectiveness of our proposed framework in identifying correct biomarkers closely related to cognitive measures. These included the Entorhinal [71,73], Hippocampus [69–71], Mid.Temporal [71], Parahippocampal [69,72], Bankssts [71], Fusiform [76], Paracentral [74], Amygdala [71,77]. The fact that our findings are consistent with results reported in previous studies demonstrates the correctness of the discovered biomarkers relevant to reveal the complex relationships between MRI measures and cognitive scores. Furthermore, we observe that the most of the top ranked MRI features in terms of prediction power are based on the average cortical thickness (TA) measurement, which implies the effectiveness of cortical thickness and is consistent with the previous studies [75]. On the other hand, the features based on volume and surface area (SA) estimation are less predictive. While most top markers are thickness measures from cortical regions, two markers are volume measures from subcortical structures including hippocampus and amygdala.

For better understanding of the underlying mechanism of AD, we can also find the identified neuroimaging biomarkers relevant to memory scores and disease categories at the same time by jointly learning the heterogeneous tasks. We found some regions identified by the heterogeneous tasks and only the regression task are common, such as Hippocampus, Entorhinal, Mid.Temporal and Fusiform, which implies that the tasks of cognitive outcomes prediction and diagnosis are highly correlated. Figs. 5 and 6 show the most relevant region areas for predicting all cognitive scores and all cognitive scores as well as clinical status, respectively.

3.4. Experiment IV: regression for future cognitive measures

In the previous experiments, the regression model is built for prediction baseline cognitive measures of on MRI data obtained at the baseline.

In this experiment, we use the longitudinal dataset from ADNI to further evaluate the performance of our proposed method. For AD, such longitudinal data usually consists of measurements at a starting time point ($t = 0$), after 6 months ($t = 6$), after 12 months ($t = 12$), after 24 months ($t = 24$), and so on usually up to 48 months ($t = 48$). We formulate the prediction of clinical scores at a sequence of time points as a multi-task regression problem, where the prediction of a clinical score for each time step is a task. In this experiment, we predict future ADAS-Cog scores of multiple times

Table 6
Top 20 selected features and ROIs by our proposed method.

Regression		Jointly regression and classification	
ROI	Feature	ROI	Feature
Entorhinal(R)	SV_Hippocampus(L)	Hippocampus(L)	SV_Hippocampus(L)
Hippocampus(L)	TA_Mid.Temporal(L)	Entorhinal(R)	CV_Entorhinal(R)
Mid.Temporal(L)	CV_Entorhinal(R)	Mid.Temporal(L)	TA_Inf.Temporal(L)
Parahippocampal(L)	TS_Mid.Temporal(R)	Inf.Temporal(L)	TA_Entorhinal(R)
Inf.Temporal(L)	TA_Entorhinal(R)	Parahippocampal(L)	TS_Entorhinal(R)
Entorhinal(L)	TA_Inf.Temporal(L)	Inf.Temporal(R)	CV_Inf.Temporal(L)
Fusiform(R)	CV_TransverseTemporal(R)	Entorhinal(L)	TS_TemporalPole(R)
Paracentral(L)	TA_Entorhinal(L)	TransverseTemporal(R)	TS_Paracentral(L)
Bankssts(L)	TA_Inf.Temporal(L)	ParsOpercularis(R)	TA_Entorhinal(L)
TemporalPole(L)	TS_Entorhinal(L)	TemporalPole(R)	SV_Inf.LateralVentricle(L)
Precentral(R)	TA_TransverseTemporal(R)	Fusiform(L)	TA_Fusiform(R)
CaudalAnt.Cingulate(L)	CV_Entorhinal(L)	CaudalMid.Frontal(R)	TA_Parahippocampal(L)
TransverseTemporal(L)	TA_Fusiform(R)	Paracentral(L)	TA_Mid.Temporal(R)
Paracentral(R)	TA_Bankssts	Mid.Temporal(R)	CV_TransverseTemporal(R)
Parahippocampal(R)	TA_Parahippocampal(L)	Amygdala(L)	CV_Cuneus(L)
Fusiform(L)	TA_Inf.Parietal(L)	Precuneus(L)	CV_Paracentral(L)
ParsOpercularis(L)	CV_Fusiform(R)	Paracentral(R)	TS_ParsOpercularis(R)
TemporalPole(R)	TS_TransverseTemporal(R)	FrontalPole(L)	TS_Sup.Frontal(L)
IsthmusCingulate(L)	SV_Amygdala(R)	Precentral(R)	CV_RostralAnt.Cingulate(R)
MedialOrbitofrontal(L)	TS_Paracentral(L)	Sup.Frontal(L)	SV_Amygdala(R)
Amygdala(L)	TA_Mid.Temporal(L)	TransverseTemporal(L)	SA_RostralAnt.Cingulate(R)

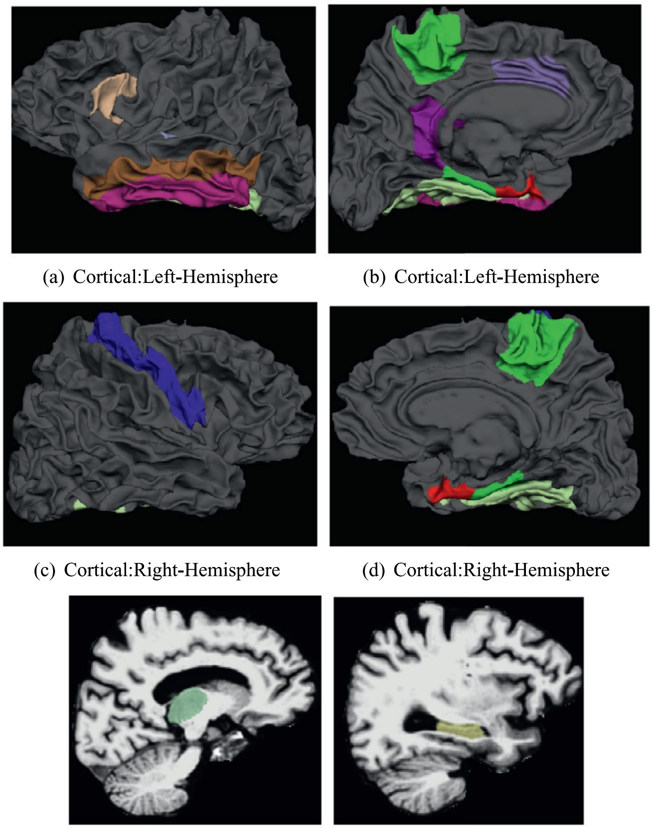
(a) Cortical:Left-Hemisphere
(b) Cortical:Left-Hemisphere
(c) Cortical:Right-Hemisphere
(d) Cortical:Right-Hemisphere
(e) Subcortical
(f) Subcortical

Fig. 5. The top 20 ROIs selected by SGLS-MTL.

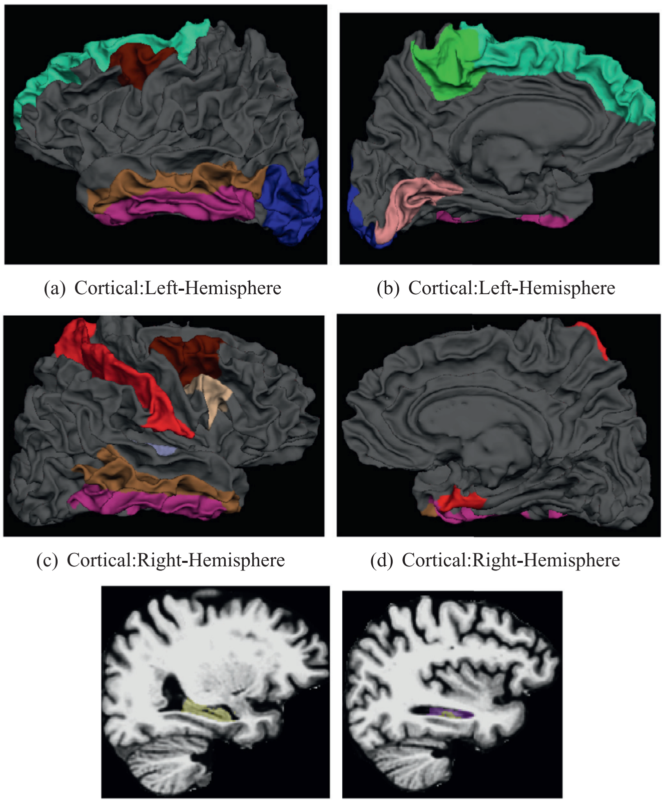
(a) Cortical:Left-Hemisphere
(b) Cortical:Left-Hemisphere
(c) Cortical:Right-Hemisphere
(d) Cortical:Right-Hemisphere
(e) Subcortical
(f) Subcortical

Fig. 6. The top 20 ROIs selected by SGLS-MTL-J.

on MRI data obtained at the baseline. For the disease progression considered in this paper, it is reasonable to assume that a small subset of features and ROI is predictive of the progression, and the multiple regression models from different time points share a common subspace structure among data samples. Since there exists many missing instances in the later time points, the formulation in Eq. (3) can be extended to the case with missing target values as:

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{Q}^T\boldsymbol{Q}=I} \|Z \odot (\boldsymbol{XU} - \boldsymbol{Y})\|_F^2 + \lambda_1\|\boldsymbol{U}\|_{2,1} + \lambda_2\|\boldsymbol{U}\|_{G_{2,1}}$$
$$+\lambda_3\|\boldsymbol{U} - \boldsymbol{Q}^T\boldsymbol{V}\|_F^2 \qquad (39)$$

We use a matrix $\boldsymbol{Z} \in R^{n\times m}$ to indicate missing target values, where $\boldsymbol{Z}_{i,j} = 0$ if the target value of sample i is missing at the jth time point, and $\boldsymbol{Z}_{i,j} = 1$ otherwise. We use the componentwise operator $\odot$ as follows: $Z = A \odot B$ denotes $z_{i,j} = a_{i,j}b_{i,j}$, for all i, j.

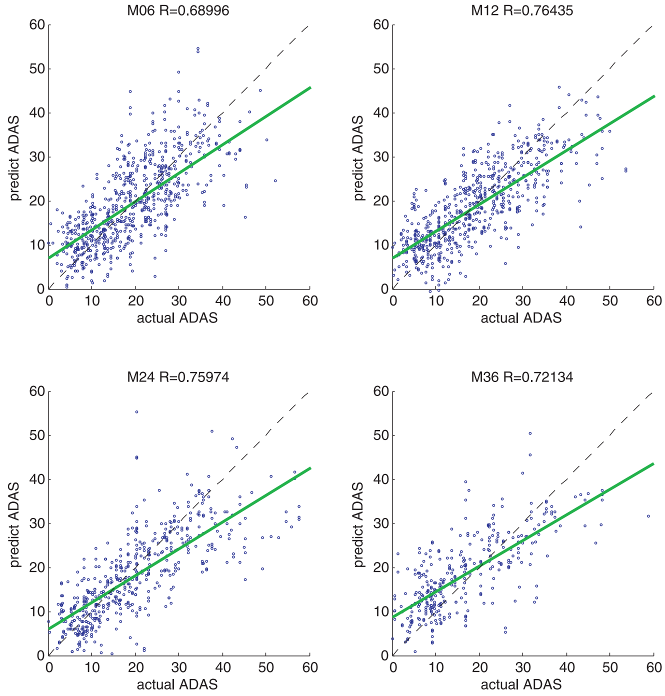

Fig. 7. Scatter plots of ADAS scores versus predicted values on testing data. The black dashed line in each figure is a reference of perfect correlation. We perform least squares regression on the points shown in the scatter plots and the green solid line is the regression line, which serves as a visual indicator of overall performance.

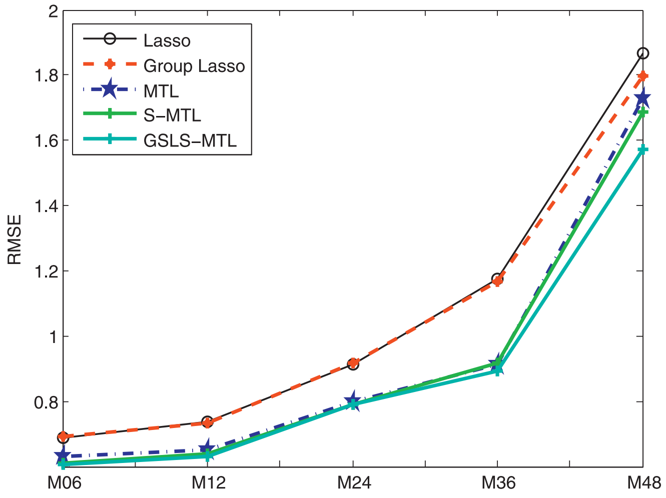

Fig. 8. Performances of different methods on a longitudinal dataset.

We show the scatter plots for the predicted values versus the actual values for ADAS-Cog and the correlation coefficient (R) on the testing data in Fig. 7. Since there are few samples available at the last time point (M48), we only show the scatter plots for the first four time points. Moreover, the experimental result of RSME is shown in Fig. 8. In the scatter plots, we see that the predicted values and actual clinical scores have a high correlation. Again, the proposed method achieved the best results, consistently performing better than the other methods.

3.5. Experiment V: subspace based clustering

The clustering result is evaluated by comparing the obtained label of each data point using clustering algorithms with that provided by the data set. The dataset is categorized into 4 classes: AD, NC, pMCI and sMCI, thus the number of clusters m is set to 4.

Depending on the space where the clustering is performed, we investigate the three different clustering methods: input space clustering methods (K-means and SGL-C in Eq. (18)), single subspace based multi-task clustering methods (MTC and SGLS-MTC in Eq. (19), and the multi-subspace based clustering (MSC [66] and SGL-MSC). SGL-C, SGLS-MTC and SGL-MSC integrate the sparse group lasso to conduct the feature selection while considering the group structure during the clustering. Consider the multi-task clustering methods, where each task corresponds to a time point $t = 1, \ldots, m$. For each time point t, we consider a clustering task based on data $(\boldsymbol{X}_t, \mathbf{y}_t)$, where $\boldsymbol{X}_t \in \mathbb{R}^{n \times p}$ denotes the matrix of covariates.

For each cluster algorithm, 10 tests were conducted on different randomly chosen clusters, and the average performance as well as the standard deviation was computed over these 10 tests. The experimental results of clustering for the baseline data are averaged over 10 repetitions and summarized in Table 7. Two widely used evaluation metrics, accuracy (ACC) and normalized mutual information (NMI), are employed to evaluate the quality of clusters.

From the result in Table 7, we make the following observations:

1. A first glance at the results shows that the SGLS-MTC algorithm achieved the best performance compared to the competing methods, which indicates that the tasks of clustering from multiple times are relevant, and exploiting the correlation among the tasks improves the clustering performance of the baseline data.
2. We can observe that SGL based clustering is better than the corresponding unsparse learning method (e.g., SGLS-MTC > MTC), which implies sparse learning is able to improve the clustering performance, and feature selection is necessary and effective for clustering analysis on the AD data with irrelevant and redundant features. Similar to regression, the feature selection with sparse group lasso can preserve the data similarity or manifold structure, so as to improve the performances for the clustering algorithms regardless of working in the input space or subspace. Since not all the brain regions are associated with AD, many of the features are irrelevant and redundant. The results indicate that the sparse based clustering methods capture essential characteristics of the high-dimensional MRI data, and are appropriate for the discovery of underlying concepts present in data.
3. When only the baseline data is used for the single task clustering, the multi-subspace based clustering achieved better performances than the clustering algorithms working in the input space, which demonstrates that the cluster structure is more clear in the underlying subspace than in the original input space.

3.6. Experiment VI: multi-modal fusion

Clinical and research studies commonly acquire complementary brain images for a more accurate and rigorous assessment of the disease status and likelihood of progression. MRI, which measures the structure of the cerebrum, has turned out to be an efficient tool for detecting the structural changes caused by AD or MCI. Fluorodeoxyglucose PET (FDG-PET), a technique for measuring glucose metabolism, is also a sensitive biomarker for the detection of AD or MCI. Each neuroimaging modality could offer valuable information for AD or MCI, and studies reported that biomarkers from different modalities could offer complementary information for different aspects of a given disease process [21,78].

To estimate the effectiveness of combining multi-modality image data with our SGLS-MTL method and provide a more compre-

Table 7
Clustering results of different clustering algorithms (Note that * and † indicate that SGLS-MTC and SGL-MSC, respectively, significantly outperformed that method on that score. with $p \leq 0.05$).

	Input space clustering		Subspace based multi-task clustering		Multi-subspace based clustering	
	K-means	SGL-C	MTC	SGLS-MTC	MSC	SGL-MSC
ACC	$0.4115 \pm 0.0779^{\star\dagger}$	$0.4626 \pm 0.0634^{\star\dagger}$	$0.5149 \pm 0.0475^{\star}$	**0.5567 ± 0.0410**	$0.4639 \pm 0.0524^{\star\dagger}$	$0.4862 \pm 0.0432^{\star}$
NMI	$0.3826 \pm 0.0582^{\star\dagger}$	$0.4057 \pm 0.0472^{\star\dagger}$	$0.4688 \pm 0.0372^{\star}$	**0.5152 ± 0.0340**	$0.4518 \pm 0.0433^{\star\dagger}$	$0.4771 \pm 0.0329^{\star}$

Table 8
Performance comparison with multi-modality data of MTL and SGLS-MTL across all tasks in terms of rMSE (Note that * stands for the case with $p \leq 0.05$).

	Methods	ADAS	MMSE	RAVLT		
				TOTAL	T30	RECOG
MTL	MRI	$0.7441 \pm 0.0356^{\star}$	$0.7981 \pm 0.0405^{\star}$	$0.7874 \pm 0.0515^{\star}$	$0.8070 \pm 0.0389^{\star}$	$0.8472 \pm 0.0369^{\star}$
	PET	$0.7258 \pm 0.0331^{\star}$	$0.7540 \pm 0.0227^{\star}$	$0.7316 \pm 0.0381^{\star}$	$0.7829 \pm 0.0355^{\star}$	$0.8033 \pm 0.0379^{\star}$
	MP	$0.7055 \pm 0.0302^{\star}$	$0.7414 \pm 0.0299^{\star}$	$0.7134 \pm 0.0288^{\star}$	$0.7535 \pm 0.0401^{\star}$	$0.7828 \pm 0.0330^{\star}$
	MPD	$0.6927 \pm 0.0288^{\star}$	$0.7226 \pm 0.0291^{\star}$	$0.7084 \pm 0.0332^{\star}$	$0.7189 \pm 0.0339^{\star}$	$0.7765 \pm 0.0416^{\star}$
SGLS-MTL	MP	$0.6861 \pm 0.0353^{\star}$	$0.7228 \pm 0.0340^{\star}$	$0.7059 \pm 0.0428^{\star}$	$0.7352 \pm 0.0377^{\star}$	$0.7772 \pm 0.0363^{\star}$
	MPD	**0.6746 ± 0.0318**	**0.7004 ± 0.0322**	**0.6975 ± 0.0355**	**0.7031 ± 0.0329**	**0.7604 ± 0.0397**

Table 9
Performance comparison with multi-modality data of MTL and SGLS-MTL in terms of correlation coefficient (CC) across all tasks (Note that * stands for the case with $p \leq 0.05$).

	Methods	ADAS	MMSE	RAVLT		
				TOTAL	T30	RECOG
MTL	MRI	$0.6592 \pm 0.0287^{\star}$	$0.5933 \pm 0.0325^{\star}$	$0.5316 \pm 0.0337^{\star}$	$0.5173 \pm 0.0595^{\star}$	$0.4428 \pm 0.0207^{\star}$
	PET	$0.6733 \pm 0.0308^{\star}$	$0.6154 \pm 0.0420^{\star}$	$0.5408 \pm 0.0456^{\star}$	$0.5363 \pm 0.0533^{\star}$	$0.4586 \pm 0.0360^{\star}$
	MP	$0.6881 \pm 0.0274^{\star}$	$0.6312 \pm 0.0408^{\star}$	$0.5519 \pm 0.0504^{\star}$	$0.5598 \pm 0.0466^{\star}$	$0.4611 \pm 0.0317^{\star}$
	MPD	$0.6979 \pm 0.0192^{\star}$	$0.6505 \pm 0.0442^{\star}$	**0.5717 ± 0.0388**	$0.5669 \pm 0.0428^{\star}$	$0.4812 \pm 0.0326^{\star}$
SGLS-MTL	MP	$0.6912 \pm 0.0279^{\star}$	$0.6497 \pm 0.0409^{\star}$	$0.5579 \pm 0.0463^{\star}$	$0.5682 \pm 0.0412^{\star}$	0.4833 ± 0.0332
	MPD	**0.7055 ± 0.0272**	**0.6616** ± 0.0387	0.5674 ± 0.0410	**0.5811 ± 0.0486**	**0.4886 ± 0.0313**

hensive comparison of the result from the proposed model, we further perform some experiments, that are (1) using only MRI modality, (2) using only PET modality, (3) combining two modalities: PET and MRI (MP), and (4) combining three modalities: PET, MRI and demographic information including age, years of education and ApoE genotyping (MPD). We compare the performance of the multi-task learning based on $\ell_{2,1}$-norm (MTL) and our SGLS-MTL on the fusing multi-modalities. For the MTL method, the features from multi-modalities are concatenated into a long vector features, while for our SGLS-MTL, the features from the same modality are considered as a group, and the $G_{2,1}$-norm are used to fuse the multi-modality data. There have been numerous reports on various ways of combining multi-modality data for efficient classification [21,79]. To our best knowledge, this is the first work that applies Multi-modal data to regression. Different from the above experiments, the samples from ADNI-2 are used instead of ADNI-1, since the number of the patients with PET is sufficient in ADNI-2. From the ADNI-2, we obtained all the patients with all MRI, PET and demographic information, totally 756 samples. The PET imaging data are from the ADNI-2 database processed by the UC Berkeley team, who use a native-space MRI scan for each subject that is segmented and parcellated with Freesurfer to generate a summary cortical and subcortical ROI, and coregister each florbetapir scan to the corresponding MRI and calculate the mean florbetapir uptake within the cortical and reference regions. The procedure of image processing is described in http://adni.loni.usc.edu/updated-florbetapir-av-45-pet-analysis-results/. The prediction performance results are shown in Tables 8 and 9.

We use the proposed sparse group lasso regularizer to explore both group-wise and individual importance of each feature for fusing multiple modalities data. We compared our SGLS-MTL with MTL by a straightforward concatenation of the different modality as baselines.

The results in these tables are summarized hereafter:

1. From the results in Tables 8 and 9, it is clear that the method with multi-modality outperforms the methods using one single modality of data. This validates our assumption that the complementary information among different modalities is helpful for cognitive function prediction. Particularly, when two modalities (MRI and PET) are used, MTL-MP and SGLS-MTL-MP improve the performances compared to MTL-MRI and MTL-PET using the unimodal information. Moreover when three modalities (MRI, PET and demographic information) are used, the regression performance is further improved.
2. Regardless of two or three modalities, the proposed multi-task learning SGLS-MTL achieves better performances than MTL. The results indicate that the group $\ell_{2,1}$-norm regularizer ($G_{2,1}$ norm) is able to capture the global relationships between modalities. This concatenated method provides a straightforward way for using multi-modality data. However, the simple concatenation method represents an equal confidence in each modality, which is often not enough to effectively fuse the heterogeneous feature sets. Furthermore, this also justifies the motivation of learning multiple tasks simultaneously with considering the group of variables regardless of the ROI structure or modality structure.

4. Conclusion

In this paper, we propose a framework for multi-task learning with hierarchical group sparsity and shared subspace to facilitate information sharing among different tasks and to better characterize Alzheimer's disease. The assumption of our SGLS-MTL is that there are common underlying structures shared by the multiple related tasks from three aspects: shared features, shared ROIs and shared subspace. Extensive experiments on ADNI data sets illustrate that proposed SGLS-MTL method not only yields superior performance on prediction performance of cognitive score prediction, but also is a powerful tool for discovering a small set of imaging biomarkers. Our current work is based on the summary statistics of ROI as input features. In the feature, we will extend our method on the higher dimensional voxel-based data [27]. Moreover, in future work, we are interested in investigating other underlying structure in features, such as graph structure, which can help gain additional insights to understand and interpret data.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (No.61502091), the Fundamental Research Funds for the Central Universities (No.N161604001 and N150408001) and the National Science Foundation for Distinguished Young Scholars of China under Grant (No.71325002 and No.61225012).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patcog.2017.07.018.

References

[1] Z.S. Khacha, Diagnosis of alzheimer's disease, Archives of Neurology 42 (11) (1985) 1097–1105.

[2] N.L. Batsch, M.S. Mittelman, World alzheimer report 2012, in: Overcoming the Stigma of Dementia. Alzheimer's Disease International (ADI), London; 2012. Accessed May, volume 5, 2015.

[3] Alzheimer's Association, et al., 2016 alzheimer's disease facts and figures, Alzheimer's & Dementia 12 (4) (2016) 459–509.

[4] R. Casanova, B. Wagner, C.T. Whitlow, J.D. Williamson, S.A. Shumaker, et al., High dimensional classification of structural MRI alzheimers disease data based on large scale regularization, Frontiers in neuroinformatics 5 (22) (2011).

[5] E.E. Tripoliti, D.T. Fotiadis, M. Argyropoulou, A supervised method to assist the diagnosis and monitor progression of alzheimer's disease using data from an fMRI experiment, Artificial intelligence in medicine 53 (1) (2011) 35–45.

[6] F. Liu, L. Zhou, C. Shen, J. Yin, Multiple kernel learning in the primal for multimodal alzheimer's disease classification, IEEE journal of biomedical and health informatics 18 (3) (2014) 984–990.

[7] R. Guerrero, C. Ledig, A. Schmidt-Richberg, D. Rueckert, ADNI, Group-constrained manifold learning: Application to AD risk assessment, Pattern recognition (2016).

[8] T. Tong, K. Graya, Q. Gao, L. Chen, D. Rueckert, ADNI, Multi-modal classification of alzheimer's disease using nonlinear graph fusion, Pattern recognition 63 (2017) 171–181.

[9] J. Peng, L. An, X. Zhu, Y. Jin, D. Shen, Structured sparse kernel learning for imaging genetics based alzheimers disease diagnosis, International Conference on Medical Image Computing and Computer-Assisted Intervention (2016) 70–78.

[10] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, et al., Automatic classification of patients with alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, NeuroImage 56 (2) (2011) 766–781.

[11] B. Shi, Y. Chen, P. Zhang, D.S. Charles, J. Liu, Nonlinear feature transformation and deep fusion for alzheimer's disease staging analysis, Pattern recognition (2016).

[12] S.P. Awate, R.M. Leahy, A.A. Joshi, Riemannian statistical analysis of cortical geometry with robustness to partial homology and misalignment, International Conference on Medical Image Computing and Computer-Assisted Intervention (2016) 237–246.

[13] E. Challis, P. Hurley, L. Serra, M. Bozzali, S. Oliver, M. Cerci, Gaussian process classification of alzheimer's disease and mild cognitive impairment from resting-state fMRI, NeuroImage 112 (2015) 232–243.

[14] S. Huang, J. Li, L. Sun, J. Liu, T. Wu, K. Chen, et al., Learning brain connectivity of alzheimer's disease from neuroimaging data, Advances in Neural Information Processing Systems (2009) 808–816.

[15] F. Rodrigues, M. Silveira, Longitudinal FDG-PET features for the classification of alzheimer's disease, in: The proceeding of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 1941–1944.

[16] C. Cabral, M. Silveira, Classification of alzheimer's disease from FDG-PET images using favourite class ensembles, in: The proceeding of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2013, pp. 2477–2480.

[17] C.R. Jack, D.S. Knopman, W.J. Jagust, L.M. Shaw, P.S. Aisen, et al., Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade, The Lancet Neurology 9 (1) (2010) 119–128.

[18] I. Beheshti, H. Demi, Alzheimers Disease Neuroimaging Initiative, et al., Feature-ranking-based alzheimers disease classification from structural MRI, Magnetic resonance imaging 34 (3) (2016) 252–263.

[19] A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, Machine Learning 73 (3) (2008) 243–272.

[20] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A.J. Saykin, S. Li, Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance, in: Proceedings of 2011 International Conference on Computer Vision, 2011, pp. 557–562.

[21] D. Zhang, D. Shen, et al., Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease, Neuroimage 59 (2) (2012) 895–907.

[22] B. Gu, V.S. Sheng, A robust regularization path algorithm for ν-support vector classification, IEEE Transactions on neural networks and learning systems (2016) 1–8.

[23] B. Gu, V.S. Sheng, K.Y. Tay, W. Romano, S. Li, Incremental support vector learning for ordinal regression, IEEE Transactions on Neural networks and learning systems 26 (7) (2015) 1403–1416.

[24] X. Zhu, H.I. Suk, D. Shen, A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis, Neuroimage 100 (2014) 91–105.

[25] X. Zhu, H. Suk, S.W. Lee, D. Shen, Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification, IEEE Transactions on Biomedical Engineering 63 (3) (2015) 607–618.

[26] J. Zhou, J. Liu, V.A. Narayan, J. Ye, Alzheimer's disease neuroimaging initiative, Modeling disease progression via multi-task learning, NeuroImage 78 (2013b) 233–248.

[27] R. Armananzas, M. Iglesias, D.A. Morales, L. Alonso-Nanclares, Voxel-based diagnosis of alzheimer's disease using classifier ensembles,IEEE journal of biomedical and health informatics.

[28] A. Beck, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM journal on imaging sciences (2009) 183–202.

[29] B. Krishnapuram, L. Carin, M.A. Figueiredo, A.J. Hartemink, Sparse multinomial logistic regression: Fast algorithms and generalization bounds, IEEE transactions on pattern analysis and machine intelligence 27 (6) (2005) 957–968.

[30] J. Ye, J. Liu, Sparse methods for biomedical data, ACM Sigkdd Explorations Newsletter 14 (1) (2012) 4–15.

[31] J. Chen, L. Tang, J. Liu, J. Ye, A convex formulation for learning shared structures from multiple tasks, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 137–144.

[32] H. Wang, F. Ni, H. Huang, C. Ding, A convex formulation for learning shared structures from multiple tasks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3097–3102.

[33] G. Prasad, S.H. Joshi, T.M. Nir, A.W. Toga, P.M. Thompson, ADNI, et al., Brain connectivity and novel network measures for alzheimer's disease classification, Neurobiology of aging 36 (2015) S121–S131.

[34] C. Davatzikos, P. Bhatt, L.M. Shaw, K.N. Batmanghelich, J.Q. Trojanowski, Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification, Neurobiology of aging 32 (12) (2011). 2322–e19

[35] L. Huang, Y. Jin, Y. Gao, K.H. Thung, D. Shen, ADNI, Longitudinal clinical score prediction in alzheimer's disease with soft-split sparse regression based random forest, Neurobiology of aging 46 (2016) 180–191.

[36] Z. Ma, F. Nie, Y. Yang, J.R.R. Uijlings, N. Sebe, Web image annotation via subspace-sparsity collaborated feature selection, IEEE Transactions on Multimedia 14 (4) (2012) 1021–1030.

[37] X. Jin, F. Zhuang, S. Wang, Q. He, Z. Shi, Shared structure learning for multiple tasks with multiple views, European Conference on Machine Learning and Knowledge Discovery in Databases (2013) 353–368.

[38] Y. Yang, Z. Ma, A.G. Hauptmann, N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks, IEEE Transactions on Multimedia 15 (3) (2013) 661–669.

[39] N. Parikh, S.P. Boyd, et al., Proximal algorithms, Foundations and Trends in optimization 1 (3) (2014) 127–239.

[40] M.W. Weiner, P.S. Aisen, C.R. Jack, W.J. Jagust, J.Q. Trojanowski, L. Shaw, A.J. Saykin, J.C. Morris, N. Cairns, L.A. Beckett, et al., The alzheimer's disease neuroimaging initiative: progress report and future plans, Alzheimers Dement 6 (2010) 202–211.

[41] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C.R. Jack, W. Jagust, J.Q. Trojanowski, et al., Ways toward an early diagnosis in alzheimers disease: the alzheimers disease neuroimaging initiative (ADNI), Alzheimer's & Dementia 1 (1) (2005) 55–66.

[42] R.S. Desikan, F. Ségonne, B. Fischl, B.T. Quinn, B.C. Dickerson, D. Blacker, R.L. Buckner, A.M. Dale, R.P. Maguire, B.T. Hyman, et al., An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, Neuroimage 31 (3) (2006) 968–980.

[43] R.K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, Journal of Machine Learning Research 6 (2005) 1817–1853.

[44] J. Liu, J. Chen, J. Ye, Multi-task feature learning via efficient $l_{2,1}$-norm minimization, UAI (2009) 339–348.

[45] J. Liu, J. Ye, Moreau-yosida regularization for grouped tree structure learning, Advances in Neural Information Processing Systems (2010) 1459–1467.

[46] J. Chen, L. Tang, J. Liu, J. Ye, A convex formulation for learning a shared predictive structure from multiple tasks, IEEE transactions on pattern analysis and machine intelligence 35 (5) (2013) 1025–1038.

[47] J. Wan, Z. Zhang, J. Yan, T. Li, B.D. Rao, S. Fang, et al., Sparse bayesian multi--task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease, 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 940–947.

[48] Z. Kang, K. Grauman, F. Sha, Learning with whom to share in multi-task feature learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 521–528.

[49] J. Chen, J. Zhou, J. Ye, Integrating low-rank and group-sparse structures for robust multi-task learning, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 42–50.

[50] A. Evgeniou, M. Pontil, Multi-task feature learning, Advances in neural information processing systems 19 (41) (2007).

[51] X. Liu, P. Cao, D. Zhao, A. Banerjee, Multi-task spare group lasso for characterizing alzheimers disease, in: The 5th Workshop on Data Mining for Medicine and Healthcare, 2016.

[52] J. Zhou, J. Chen, J. Ye, MALSAR: Multi-task learning via structural regularization, Arizona State University, 2011.

[53] J. Wan, Z. Zhang, B.D. Rao, S. Fang, J. Yan, A.J. Saykin, L. Shen, Identifying the neuroanatomical basis of cognitive impairment in alzheimer's disease by correlation-and nonlinearity-aware sparse bayesian learning, IEEE transactions on medical imaging 33 (7) (2014) 1475–1487.

[54] J. Ye, M. Farnum, E. Yang, R. Verbeeck, V. Lobanov, N. Raghavan, G. Novak, A. DiBernardo, V.A. Narayan, Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data, BMC neurology 12 (1) (2012) 1.

[55] U.V. Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (4) (2007) 395–416.

[56] J. Tang, H. Liu, Unsupervised feature selection for linked social media data, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 904–912.

[57] S. Wang, J. Tange, H. Liu, Embedded unsupervised feature selection, AAAI Conference on Artificial Intelligence (2015) 470–476.

[58] C.H. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE transactions on pattern analysis and machine intelligence 32 (1) (2010) 45–55.

[59] J. Huang, F. Nie, H. Huang, C. Ding, Robust manifold nonnegative matrix factorization, ACM Transactions on Knowledge Discovery from Data (TKDD) 8 (3) (2014) 11–19.

[60] D.C. Pham, O. Arandjelović, S. Venkatesh, Grouping and dimensionality reduction by locally linear embedding, in: 2016 International Joint Conference on Neural Networks (IJCNN)), 2001, pp. 1255–1262.

[61] J. Ho, M.H. Yang, J. Lim, K.C. Lee, D. Kriegman, Clustering appearances of objects under varying illumination conditions, in: IEEE computer society conference on Computer vision and pattern recognition, 2003, pp. 1–8.

[62] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (12) (2005) 1945–1959.

[63] M. Polito, P. Perona, Achieving stable subspace clustering by post-processing generic clustering results, Advances in neural information processing systems) (2016) 2390–2396.

[64] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, science 290 (5500) (2000) 2323–2326.

[65] X. Liu, D. Tosun, M.W. Weiner, N. Schuff, Alzheimer's Disease Neuroimaging Initiative, et al., Locally linear embedding (LLE) for MRI based alzheimer's disease classification, NeuroImage 83 (2013) 148–157.

[66] A. Goh, R. Vidal, Segmenting motions of different types by unsupervised manifold clustering, IEEE Conference on Computer Vision and Pattern Recognition) (2007) 1–6.

[67] D. Cai, X. He, J. Han, Spectral regression: A unified approach for sparse subspace learning, in: Proceedings of the 7th IEEE International Conference on Data Mining), 2007, pp. 73–82.

[68] N. Meinshausen, P. Bühlmann, Stability selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (4) (2010) 417–473.

[69] H. Wang, F. Nie, H. Huang, J. Yan, V. Lobanov, S. K., S. Risachervak, A. Saykin, L. Shen, High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction, Advances in Neural Information Processing Systems (2012) 1277–1285.

[70] T. Li, J. Wana, E. Yang, J. Zhang, J. Yan, S. Kim, S.L. Risacher, S. Fang, et al., Hippocampus as a predictor of cognitive performance: comparative evaluation of analytical methods and morphometric measures, MICCAI Workshop on Novel Imaging Biomarkers for Alzheimers Disease and Related Disorders (NIBAD12) (2012) 133–144.

[71] L. Shen, Y. Qi, S. Kim, K. Nhoang, J. Wan, S.L. Risacher, A.J. Saykin, et al., Sparse bayesian learning for identifying imaging biomarkers in AD prediction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (NIBAD12), 2010, pp. 611–618.

[72] K.A. Celone, V.D. Calhoun, B.C. Calhoun, A. Atri, E.F. Chua, S.L. Miller, K. DePeau, et al., Alterations in memory networks in mild cognitive impairment and alzheimer's disease: an independent component analysis, The Journal of neuroscience 26 (40) (2006) 10222–10231.

[73] R.S. Desikan, H.J. Cabral, C.P. Hess, W.P. Dillon, C.M. Glastonbury, M.W. Weiner, N.J. Schmansky, et al., Automated MRI measures identify individuals with mild cognitive impairment and alzheimer's disease, Brain (2009).

[74] L. Wang, F.C. Goldstein, E. Veledar, A.I. Levey, J.J. Lah, C.C. Meltzer, C.A. Holder, H. Mao, Alterations in cortical thickness and white matter integrity in mild cognitive impairment measured by whole-brain cortical thickness mapping and diffusion tensor imaging, American Journal of Neuroradiology 30 (50) (2009) 893–899.

[75] S. Zhe, Z. Xu, Y. Qi, P. Yu, Sparse bayesian multiview learning for simultaneous association discovery and diagnosis of alzheimer's disease, pp. 1966–1972, 2015.

[76] Y. Liu, T. Paajanen, Y. Zhang, E. Westman, L. Wahlund, A. Simmons, C. Tunnard, et al., Combination analysis of neuropsychological tests and structural MRI measures in differentiating AD, MCI and control groups the addneuromed study, Neurobiology of Aging 32 (7) (2011) 1198–1206.

[77] H. Matsuda, Voxel-based morphometry of brain MRI in normal aging and alzheimers disease, Aging and disease 4 (1) (2013) 29–38.

[78] L. Yuan, Y. Wang, P.M. Thompson, V.Q. Narayan, J. Ye, Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, 2012, pp. 1149–1157.

[79] T. Tong, K. Gray, A. Gao, L. Chen, D. Rueckert, ADNI, et al., Multi-modal classification of alzheimer's disease using nonlinear graph fusion, Pattern Recognition 63 (2017) 171–181.

Peng Cao is a postdoctoral fellow at Northeastern University, China. He earned his Ph.D. degree in computer application in 2014 at Northeastern University, China. His research interests include imbalanced data learning, multi-task learning and medical data mining.

Xuanfeng Shan is a graduate student in Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, China. His research interests include sparse learning, optimization and medical data mining.

Dazhe Zhao is professor in the Department of Computer Science and Engineering, Northeastern University, China. She is the chairman of Key Laboratory of Medical Image Computing of Ministry of Education, Northeast University, China. She is also a senior member of China Computer Association. She is interested in software engineering, and medical image computing.

Min Huang is a full professor in the Department of Systems Engineering at Northeastern University, China and was a senior visiting scholar in the Department of Industrial and Operations Engineering at University of Michigan (Ann Arbor) in 2011. Dr. Huang has also been recognized as the Distinguished Young Scholars by the National Science Foundation of China. Her research interests focuses on the modeling and optimization for manufacturing systems, logistics and supply chain systems, as well as healthcare systems, etc.

Osmar Zaiane is a Professor in Computing Science at the University of Alberta, Canada, and Scientific Director of the Alberta Innovates Centre for Machine Learning. He is Associate Editor of many International Journals on data mining and data analytics and served as program chair and general chair for scores of international conferences in the field of knowledge discovery and data mining. His current research interests include data mining, healthcare informatics.