Pattern Recognition 79 (2018) 195–215

Contents lists available at ScienceDirect

Pattern Recognition

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

$\ell_{2,1} - \ell_1$ regularized nonlinear multi-task representation learning based cognitive performance prediction of Alzheimer's disease

Peng Cao[a,b,*], Xiaoli Liu[a], Jinzhu Yang[a,b], Dazhe Zhao[a,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 12 August 2017
Revised 18 November 2017
Accepted 24 January 2018
Available online 2 February 2018

Keywords:
Alzheimer'S disease
Regression
Sparse learning
Multi-task learning
Kernel method

ABSTRACT

Alzheimer's disease (AD) has been not only a substantial financial burden to the health care system but also the emotional hardship to patients and their families. Predicting cognitive performance of subjects from their magnetic resonance imaging (MRI) measures and identifying relevant imaging biomarkers are important research topics in the study of Alzheimer's disease. Many previous works formulate the prediction task as a linear regression problem. The most critical limitation is that they assume a linear relationship between the MRI features and the cognitive outcomes. The linear models in original MRI feature spaces can be limited by their inability to exploit the nonlinear relation between the MRI features and cognitive measure prediction tasks. To better capture the complicated but more flexible relationship between the cognitive scores and the neuroimaging measures, we propose a $\ell_{2,1} - \ell_1$ norm regularized multi-kernel multi-task feature learning formulation with a joint sparsity inducing regularization. The formulation facilitates the shared kernel functions, as well as the high dimensional features in the kernel induced feature spaces simultaneously, to look for the common representation that are useful for all tasks by promoting use of few kernels and few learned features in each kernel. For optimization, we develop an alternating optimization method to effectively solve the proposed mixed norm regularized formulation. We evaluate the performance of the proposed method using the Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets and demonstrate that our proposed methods achieve not only clearly improved prediction performance for cognitive measurements with single MRI modality or multi-modalities data, but also a compact set of highly suggestive biomarkers relevant to AD.

1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia, which mainly affects memory function, ultimately culminating in a dementia state where all cognitive functions are affected. The disease poses a serious challenge to the aging society [1,2]. Predicting cognitive performance of subjects from neuroimage measures and identifying relevant imaging biomarkers are important research topics in the study of Alzheimer's disease. Many cognitive measures have been designed to clinically evaluate the cognitive status of the patients and used as important criteria for clinical diagnosis of probable AD, such as Alzheimer's Disease Assessment Scale cognitive total score (ADAS), Mini Mental State Exam score (MMSE), Rey Auditory Verbal Learning Test (RAVLT), Category Fluency (FLU), and Trail Making Test (TRAILS). Recently more attention has been given to the prediction of the cognitive outcomes and the identification of neuroimaging predictors for cognitive decline in AD. Magnetic resonance imaging (MRI) provides a chance to directly observe brain changes such as cerebral atrophy or ventricular expansion [3]. The relationship between structural changes in MRI and the cognitive measures has been previously studied by regression models [4–6].

Recently, instead of learning individual models, the multi-task learning (MTL) methods [7] have been widely studied to jointly build a better model for each task by incorporating inherent correlations among multiple clinical cognitive measures [6,8–13]. The primary motivation of using multi-task learning is due to its ability to learn a shared representation across related tasks and reduce the prediction error of each task. The most recent studies [14–16] employed multi-task based linear models with $\ell_{2,1}$ norm [17] regularization to identify the features that are important all or most clinical scores. The $\ell_{2,1}$-norm is chosen to be the regular-

* Corresponding author at: Computer Science and Engineering, Northeastern University, Shenyang, China.
E-mail address: caopeng@cse.neu.edu.cn (P. Cao).

https://doi.org/10.1016/j.patcog.2018.01.028
0031-3203/

ization because it encourages multiple predictors to share similar sparsity patterns. Thus, the $\ell_{2,1}$-norm regularized regression model is able to select some common features across all the tasks. However, the assumption of these existing linear models usually does not hold due to the inherently complex patterns between brain images and the corresponding cognitive outcomes. Modeling cognitive scores as nonlinear functions of MRI features may provide enhanced flexibility and the potential to better capture the complex relationship between the MRI features and cognitive outcomes.

Many kernel-based classification or regression methods with faster optimization speed or stronger generalization performance have been proposed and investigated by theoretical analysis and experimental evaluation [18,19]. The kernel methods also have been widely used for the predictive classification [20,21] or regression [14] in the current research on Alzheimer's disease. Since imaging markers relevant to a certain cognition task more or less affect the other cognitive scores, naturally these prediction tasks share some commonality. It is therefore important to extend the existing kernel-based learning methods to the multi-task learning paradigm, and how to incorporate the scheme of multi-task learning into the kernel methods is critical. Some work proposed methods incorporating task relations into regularization terms in kernel methods by assuming that the tasks share the same kernel. Hence, the problem of inferring task relationships boils down to the problem of learning a multi-task kernel [22,23]. The most related work to ours is CORrelation- and NonLINearity-aware SBL (CORNLIN) in [5], where a polynomial kernel function is adopted as a nonlinear mapping to introduce high-order features and the cognitive scores are modeled as nonlinear functions of neuroimaging variables. An $\ell_{2,1}$ norm is employed on the higher dimensional features to build the correlation among the tasks. It has shown the superiority of the nonlinearity-aware method compared with the competing linear methods. However, the limitations of CORNLIN [5] are: (1) The high-order features are explicitly represented by the polynomial kernel. It cannot be mapped to other high-order feature forms by more complicated kernels, e.g. RBF, leading to limited flexibility and predictive performance; (2) the choice of the types and parameters of the kernels is critical for a particular task [24], which determines the mapping between the input space and the feature space. The inappropriate kernels may not accurately capture the correlation structure of the data. It is necessary to emplmultiple kernels for learning multiple tasks. A possible way to address this problem is to learn an optimal kernel function by a weighted, linear combination of predefined candidate kernels within the framework of multiple kernel learning (MKL) [25]. The idea of exploiting multiple kernels to improve MTL has also been addressed in [26,27]. They model relationships between the function parameters by employing multiple kernels for multiple tasks via kernel regularizations. Although these works capture the nonlinear predictors-to-response relationship and encourage the parameters of kernel functions to be shared across the multiple tasks, they donnot consider the intrinsic relationships among multiple related tasks over the higher dimensional features in the RKHS (Reproducing Kernel Hilbert Spaces). The limitation lead to ignore the shared non-linear predictive information beneficial to tasks in the kernel space. In addition, the combination of MKL and MTL has also been applied in the diagnosis of AD. Two most recent studies employed MTL to select features from multi-modality data (MRI and PET) [28,29] or from multi-task (e.g. ADAS, MMSE) [30] by $\ell_{2,1}$ norm linearized MTL, then a MKL method is adopted to combine multi-modality data. However, in all the three methods, the MTL and MKL are conducted individually rather than collectively as in an unified framework, The two sub-problems influence each other, resulting in obtaining a suboptimal solution. Moreover, the combination of the linearized MTL for feature learning and the nonlinear MKL for fusion and classification tends to result in inconsistencies since they work in the different spaces.

Given these considerations, a question rises naturally: *how to formulate the MTL problem in the kernel-induced feature space and to obtain the best kernel space at the same time?* The question above motivates us to develop a novel framework to model task relationship in the high dimensional features and kernel functions simultaneously when handling multiple related regression tasks. Firstly, we exploit a nonlinear prediction model to capture the more complicated but more flexible relationship between MRI measures and cognitive outcomes using multi-kernel method as a framework.

Rather than conducting feature learning and kernel learning individually, we pose the problem of multi-task learning as that of simultaneously learning a shared representation from high dimensional features and kernels, to capture the kernel-wise relationships among multiple tasks without ignoring the feature-wise correlation within each kernel space. Specifically, we propose a multi-kernel based multi-task learning with a joint sparsity-inducing regularization $\ell_{2,1} - \ell_1$ norm, called $\ell_{2,1} - \ell_1$SMKMTL. The proposed formulation explicitly captures the task correlation structure with $\ell_{2,1}$-norm regularization on the high dimensional features in the RKHS space. Moreover, an ℓ_1-norm is simultaneously employed over the $\ell_{2,1}$-norm, which ensures a small subset of kernels will be selected across all the tasks, thus identifies the important kernel functions. An overview of the proposed cognitive scores prediction pipeline is illustrated in Fig. 1. Convincing experimental results show that exploiting the two kinds of correlations can significantly improve the prediction performance and help accurately identify biologically meaningful imaging predictors. Then, we derive an alternative optimization algorithm to solve the proposed mixed norm regularized formulation efficiently. Furthermore, the MKL model has the advantage of fusing multiple modalities [14]. We apply our SMKMTL on multiple data modalities (MRI, PET, ApoE and demographic information) in our study.

The rest of this paper is organized as follows. Section 2 presents the problem formulation. Section 3 briefly reviews $\ell_{2,1}$-norm regularized multi-task learning. Section 4 introduces the formulation of our proposed $\ell_{2,1} - \ell_1$SMKMTL method and optimization algorithm. Section 5 presents experimental results on comparison of different prediction methods using the Alzheimer's Disease Neuroimaging Initiative (ADNI) data [31]. A discussion and limitations are provided in Section 6. This paper is concluded in Section 7.

2. Problem formulation

The aim of our work is to predict subjects' cognitive scores (e.g. ADAS, MMSE) using their MRI features (e.g. volume, area and thickness) across the entire brain. It is a regression problem. In order to associate the imaging markers and the cognitive measures, the regularized multivariate regression model is adopted in our study, treating MRI features as inputs and cognitive outcomes as outputs. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$ be MRI features (e.g. the volume of hippocampus), where n and p are the number of training instances and dimensionality of $\boldsymbol{x}_i$, $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^T \in \mathbb{R}^{n\times t}$, where $\boldsymbol{y}_i$ is the target cognitive score for $\boldsymbol{x}_i$ and t is the number of tasks, $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t] \in \mathbb{R}^{p\times t}$, where $\boldsymbol{w}_h$ is the weight vector for the hth task.

In the regression method of each prediction task, a subject's cognitive score under a task is modeled as a linear function of the corresponding MRI features: $f(\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{w}_t$. For instance, for the ith subject the cognitive score under the hth task is model as:

$$y_h = w_{h1}x_{i1} + w_{h2}x_{i2} + \cdots + w_{hp}x_{ip} + \xi_h \tag{1}$$

where ξ_h denotes the residual error of the hth task.

In our study, multiple regression models for t tasks are simultaneously considered for learning. The objective function with re-

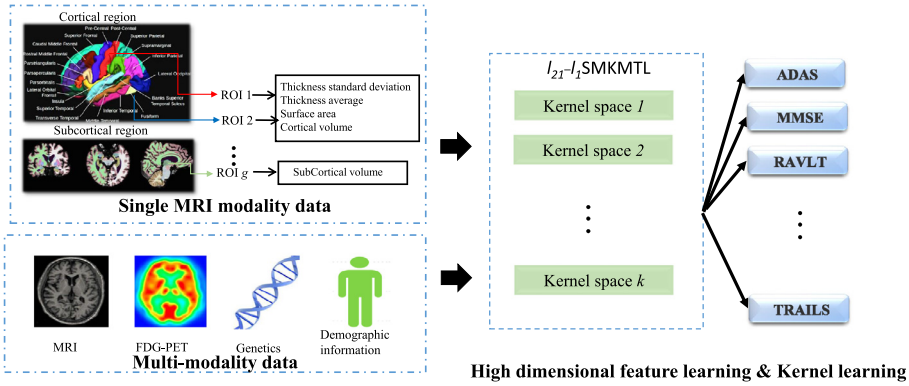


Fig. 1. Schematic illustration of our proposed framework.

spect to $\boldsymbol{W}$ can be formulated as:

$$\min_{\boldsymbol{W}} \quad L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}) + \lambda R(\boldsymbol{W}) \ , \tag{2}$$

where $L(\cdot)$ denotes the loss function (The square loss is considered: $L(\cdot) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_F^2$), $R(\cdot)$ is the regularizer and $\lambda > 0$ is the regularization parameter

Many regularization based MTL methods with different assumptions about how tasks are related have been proposed, leading to different regularization terms $R(\boldsymbol{W})$ in the formulation.

3. $\ell_{2,1}$-norm regularized multi-task learning

3.1. $\ell_{2,1}$-norm regularized linear multi-task learning

The biggest challenge in the prediction of inferring cognitive outcomes with MRI is the high dimensionality, which affects the computational performance and leads to a wrong estimation and identification of the relevant predictors. The commonly used ℓ_2-norm regularization leads to non-zero values for all parameters in $\boldsymbol{W}$. To reduce the high dimensionality of MRI features and identify some relevant biomarkers, some sparse methods with sparsity-inducing regularization have been employed [32], such as Lasso [33]. The Lasso formulation solves the following optimization problem:

$$\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_F^2 + \lambda \|\boldsymbol{W}\|_1, \tag{3}$$

where and $\|\cdot\|_F$ denotes the Frobenius norm. One major limitation of the Lasso above is that the tasks are assumed to be independent from each other.

Multi-task learning (MTL) [7] is a learning paradigm which seeks to improve the generalization performance of all tasks involved. The fundamental hypothesis of the MTL methods is to assume that if tasks are related then learning of one task can benefit from the learning of other tasks. Learning multiple related tasks simultaneously has been theoretically and empirically shown to often significantly improve the performance [22,34–36]. The key of the MTL is how to exploit the correlation among the tasks via an appropriate shared representation. Two popular shared representations for modeling task relatedness are model parameter sharing [17,37] and feature representation sharing [38–40]. It is known that there exist inherent correlations among different cognitive scores [5,6,41]. Therefore, the prediction of different types of cognitive scores is modeled as a MTL formulation, and the tasks are related in the sense that they all share a small set of features for all tasks, which is multi-task feature learning problem [42]. The assumption of feature representation sharing in AD is that only a subset of brain regions are relevant to each assessment, since multiple cognitive assessment are essentially influenced by the same important underlying pathology.

The $\ell_{2,1}$-norm was popularly used in multi-task feature learning [42]. Since the $\ell_{2,1}$-norm regularizer imposes the sparsity between all features and non-sparsity between tasks, the features that are discriminative for all tasks will get large weights. The objective function of the $\ell_{2,1}$-norm regularized MTL (called $\ell_{2,1}$MTL) is given by:

$$\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_F^2 + \lambda \|\boldsymbol{W}\|_{2,1} \tag{4}$$

One major challenge lies in the nonsmoothness of the $\ell_{2,1}$-norm regularization. The formulation of (4) can be solved by Nesterov's method with proximal gradient efficiently [43].

3.2. $\ell_{2,1}$-norm regularized multi-kernel multi-task learning

The limitation in this traditional $\ell_{2,1}$MTL model is that a subject's cognitive score under a task is modeled as a linear function of the MRI features. The kernel methods, e.g. SVM or SVR can model the nonlinear distribution of the data by mapping the input data into a nonlinear feature space by kernel embedding. In this section, we consider the case that $\ell_{2,1}$MTL is extended to learn the form of high-dimensional feature with nonlinear feature mapping. Let us define the kernel function $\phi_j(\boldsymbol{x}) : \mathbb{R}^p \to \mathbb{R}^{\hat{p}}$, that maps the data samples from an input space $\mathcal{X}$ to a feature space (RKHS) $\mathcal{H}$, where $\hat{p}$ denotes the dimensionality of the feature space. A kernel function k' is capable of attaining the inner product of two mapped data in $\mathcal{H}$: $k'(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{x}')$ without explicitly computing the high-dimensional data. The Gram matrix associated with the kernel function k' has entries $K(i, j) = k'(\boldsymbol{x}_i, \boldsymbol{x}_j)$. With the kernel mapping $\phi(\cdot)$, we can model the hth cognitive measure as a nonlinear function of the MRI features for the ith subject:

$$y_h = \hat{w}_{h1}\phi(\boldsymbol{x}_i)_1 + \hat{w}_{h2}\phi(\boldsymbol{x}_i)_2 + \cdots + \hat{w}_{h\hat{p}}\phi(\boldsymbol{x}_i)_{\hat{p}} + \xi \tag{5}$$

In the kernel methods, the most suitable types and parameters of the kernels for a particular task is often unknown. Instead of using only one specific kernel, Multiple Kernel Learning (MKL) attempts to achieve better results by combining several base kernels. MKL assumes that $\boldsymbol{x}_i$ can be mapped to k different Hilbert spaces, $\boldsymbol{x}_i \to \phi_j(\boldsymbol{x}_i)$, $j = 1, \ldots, k$, implicitly with k nonlinear mapping functions, and the objective of MKL is to seek the optimal kernel combination $\hat{k}'(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^k d_j k'_j(x, x')$, $d_j \geq 0$, $\sum_{j=1}^k d_j = 1$, where $\boldsymbol{d}$ is the kernel weight vector. The constraint of the kernel weight vector is also called simplex. We denote $\mathcal{S}_{\boldsymbol{d}} := \{\boldsymbol{d} \geq 0, \ \sum_{j=1}^k \boldsymbol{d}_j = 1\}$. The primal objective function of multiple kernel regression model

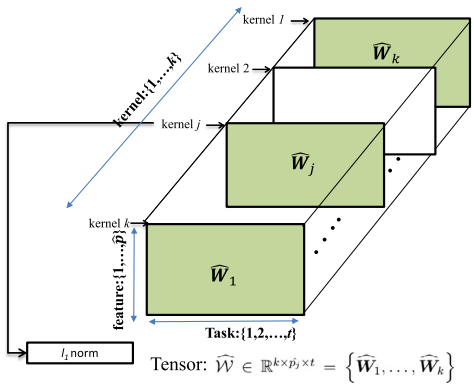


Fig. 2. The visualization of the learned weight tensor $\widehat{\mathcal{W}}$. The tensor is indexed by features, kernel and tasks. Each rectangle indicates a RKHS induced by a specific kernel function, and ℓ_1-norm regularizer is applied to promote the use of few kernels. The regression weights of the high dimensional features for a particular kernel are encouraged to be either zero or non-zero across all the tasks.

is written as:

$$\begin{aligned} \min_{\widehat{\boldsymbol{W}},\xi,\boldsymbol{d}\in\mathcal{S}_{\boldsymbol{d}}} \quad & \frac{1}{2}\sum_{j=1}^{k}\frac{\|\hat{\boldsymbol{w}}_j\|_2^2}{d_j}+\frac{\lambda}{2}\sum_{i=1}^{n}\xi_i^2 \\ \text{s.t.} \quad & \sum_{j=1}^{k}\hat{\boldsymbol{w}}_j^{\mathrm{T}}\phi_j(x_i)-y_i=\xi_i,\ \ i=1,\ldots,n \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{w}}_j$ is the normal of the separating hyperplane for the feature mapping ϕ_j, $\boldsymbol{\xi}=[\xi_1,\ldots,\xi_n]$ is the vector of slack variables, and λ is the regularization parameter.

The objective value of the dual problem of (6):

$$\begin{aligned} J(\boldsymbol{d}) = \ & \max_{\boldsymbol{\alpha}} \ -\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_t-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{K}}\boldsymbol{\alpha}-\frac{1}{2C}\boldsymbol{\alpha}^{*\mathrm{T}}\boldsymbol{\alpha} \\ \text{s.t.} \ & \sum_{j=1}^{k}d_j=1, \qquad d_j\geq 0 \end{aligned} \tag{7}$$

where $\hat{\boldsymbol{K}}=\sum_{j=1}^{k}d_j\boldsymbol{K}_j$, is the combined Gram matrix.

MKL learns both the weights of the kernel combination $\boldsymbol{d}$ and the parameters of the regression $\widehat{\boldsymbol{W}}$ by solving a single joint optimization problem. We follow the multiple kernel Learning scheme and use the $\ell_{2,1}$-norm to model the relationship among the tasks to learn a common kernel representation by imposing sparsity constraint on the kernel weight. The method, called $\ell_{2,1}$MKMTL, assumes that few base kernel is important for the tasks, and encourages a linear combination of only few kernels and assumes few selected kernels are similar across the tasks. The formulation of $\ell_{2,1}$MKMTL can be expressed as:

$$\begin{aligned} \min_{\widehat{\boldsymbol{W}},\xi} \quad & \frac{1}{2}\left(\sum_{j=1}^{k}\left(\sum_{h=1}^{t}\|\hat{\boldsymbol{w}}_{j.h}\|_2^2\right)^{\frac{1}{2}}\right)^2+\frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{ti}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{k}\hat{\boldsymbol{w}}_{j.h}^{\mathrm{T}}\phi_j(x_{ti})-y_{ti}=\xi_{hi},\ \ h=1,\ldots,t,\ \ i=1\ldots n_h \end{aligned} \tag{8}$$

where $\widehat{\mathcal{W}}\in\mathbb{R}^{k\times\hat{p}_j\times t}=\left\{\widehat{\boldsymbol{W}}_1,\ldots,\widehat{\boldsymbol{W}}_k\right\}$ is a weight tensor indexed by kernel, features and tasks (See Fig. 2), $\widehat{\boldsymbol{W}}_j\in\mathbb{R}^{\hat{p}\times t}$ denotes the parameter matrix, with row $\hat{\boldsymbol{w}}_{jl.}\in\mathbb{R}^{\hat{p}}$ corresponding to feature l, $l=1,\ldots,\mathbb{R}^{\hat{p}}$, and column $\hat{\boldsymbol{w}}_{j.h}\in\mathbb{R}^t$ corresponding to task h, $h=1,\ldots,t$, $\hat{w}_{jlh}$ is the lth feature weight of the hth task in the jth RKHS, $\boldsymbol{\xi}$ is the regression error, and λ adjusts the trade-off between the regression error and the regularization.

According to the Proposition 1 in [27], the formulation of Eq. (8) can be rewritten as:

$$\begin{aligned} \min_{\widehat{\boldsymbol{W}},\boldsymbol{D},\xi} \quad & \frac{1}{2}\sum_{j=1,h=1}^{k,t}\frac{\|\hat{\boldsymbol{w}}_{j.h}\|_2^2}{d_{hj}}+\frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{ti}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{k}\hat{\boldsymbol{w}}_{j.h}^{\mathrm{T}}\phi_j(x_{ti})-y_{hi}=\xi_{hi},\ \ h=1\ldots t,\ \ i=1\ldots n_h \\ & \sum_{j=1}^{k}\left(\sum_{h=1}^{t}d_{hj}^2\right)^{\frac{1}{2}}\leq 1, \quad d_{hj}\geq 0, \quad \forall h,k, \end{aligned} \tag{9}$$

where matrix $\boldsymbol{D}=[\boldsymbol{d}_1,\ldots,\boldsymbol{d}_t]^T\in\mathbb{R}^{t\times k}$ is the matrix of the kernel weights.

The formulation in Eq. (9) transfers the mixed $\ell_{2,1}$-norm on $\widehat{\boldsymbol{W}}$ to another mixed norm on the weights $\boldsymbol{D}$. The formulation promotes sparsity across kernels with an ℓ_1 norm and non-sparse combinations across tasks. The formulation can be solved by an iterative algorithm based on block-coordinate descent [27].

4. $\ell_{2,1}-\ell_1$ regularized sparse multi-kernel multi-task learning, $\ell_{2,1}-\ell_1$SMKMTL

4.1. Formulation of $\ell_{2,1}-\ell_1$ SMKMTL

The linearized $\ell_{2,1}$MTL assumed linear relationship between the MRI features and the cognitive outcomes. Although the multi-kernel $\ell_{2,1}$ MKMTL builds the nonlinear relationship for the features and tasks by mapping to high dimensional space, it only considers that tasks to be learned share a common subset of kernel representation without capturing the interrelationships among different cognitive measures over the feature space.

To overcome the weaknesses of the previous two methods, we project the original feature vectors to a high-dimensional space using multiple non-linear mapping functions for performing regression task in a nonlinear manner, and perform multi-task learning in the multiple kernel space for modeling the disease's cognitive scores with a joint $\ell_{2,1}-\ell_1$ sparsity-inducing regularizers. The assumptions of our model are that (1) only a small set of features are predictive for all the prediction tasks in the feature space (RKHS), (2) only a small set of kernels are common across all the tasks. For achieving this goal, an $\ell_{2,1}$-norm is applied on the high dimensional features, to encourages all tasks to shared a common set of the high dimensional features. Meanwhile a ℓ_1-norm is used to look for kernels that are useful for all tasks by promoting use of few kernels.

With the joint $\ell_{2,1}-\ell_1$ sparsity-inducing regularizer, we cast our problem as the following optimization problem:

$$\begin{aligned} \min_{\widehat{\mathcal{W}},\xi} \quad & \frac{1}{2}\left(\sum_{j=1}^{k}\left(\sum_{l=1}^{\hat{p}_j}\|\hat{\boldsymbol{w}}_{jl.}\|_2\right)\right)^2+\frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{hi}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{k}\hat{\boldsymbol{w}}_{j.h}^{\mathrm{T}}\phi_j(x_{hi})-y_{hi}=\xi_{ti}\ ,h=1\ldots t,\ \ i=1\ldots n_h \end{aligned} \tag{10}$$

The proposed multi-kernel framework utilizes tensor to capture the relatedness among tasks and transfer knowledge through high dimensional features and kernels, which cannot be achieved by any existing MTL formulations. In our formulation, the square least loss is chosen as the loss function. The above formulation can be easily generalized to other convex loss functions, such as ξ-insensitive loss for regression or hinge loss for classification.

Lemma 1. *Let $a_i\geq 0, i=1\ldots m$.*

$$\min\left\{\sum_{i=1}^{m}\frac{a_i}{\lambda_i}:\lambda_i\geq 0,\sum_{i=1}^{m}\lambda_i\leq 1\right\}=\left(\sum_{i=1}^{m}a_i^{\frac{1}{2}}\right)^2 \tag{11}$$

and the minimum is attained at

$$\lambda_i = \frac{a_i^{\frac{1}{2}}}{\sum_{i=1}^{m} a_i^{\frac{1}{2}}} \tag{12}$$

Proof. From the Cauchy–Schwarz inequality we have that

$$\sum_i a_i^{\frac{1}{2}} = \sum_i \frac{a_i^{\frac{1}{2}}}{\lambda_i^{\frac{1}{2}}} \lambda_i^{\frac{1}{2}} \le \left(\sum_i \frac{a_i}{\lambda_i}\right)^{\frac{1}{2}} \left(\sum_i \lambda_i\right)^{\frac{1}{2}} \le \left(\sum_i \frac{a_i}{\lambda_i}\right)^{\frac{1}{2}} \tag{13}$$

Using the result of the Lemma 1 and introducing new variables $\boldsymbol{d} = [d_1 \dots d_k]^T$, we have:

$$\left(\sum_{j=1}^{k}\left(\sum_{l=1}^{p_j} \|\hat{\boldsymbol{w}}_{jl.}\|_2\right)\right)^2 = \min_{\boldsymbol{d}\in\mathcal{S}_{\boldsymbol{d}}} \sum_{j=1}^{k} \frac{\left(\sum_{l=1}^{p_j} \|\hat{\boldsymbol{w}}_{jl.}\|_2\right)^2}{d_j}, \tag{14}$$

where $\mathcal{S}_{\boldsymbol{d}} := \left\{\boldsymbol{d} \ge 0,\ \sum_{j=1}^{k} \boldsymbol{d}_i = 1\right\}$.

Again using the lemma and introducing new variables $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_k]$, where $\boldsymbol{\theta}_j = [\theta_{j1} \dots \theta_{j\hat{p}_j}]^{\mathrm{T}}$, the regularizer can be written as:

$$\left(\sum_{l=1}^{\hat{p}_j} \|\hat{\boldsymbol{w}}_{jl.}\|_2\right)^2 = \min_{\boldsymbol{\theta}_j\in\mathcal{S}_{\boldsymbol{\theta}_j}} \frac{\hat{w}_{jlh}^2}{\theta_{jl}}, \tag{15}$$

where $\mathcal{S}_{\boldsymbol{\theta}_j} := \left\{\theta_{jl} \ge 0, \sum_{i=1}^{\hat{p}_j} \theta_{jl} \le 1\right\}$.

According to the Eqs. (14) and (15), we can rewrite the mixed norm regularization on $\widehat{\boldsymbol{W}}$ as:

$$\left(\sum_{j=1}^{k}\left(\sum_{l=1}^{\hat{p}_j} \|\hat{\boldsymbol{w}}_{jl.}\|_2\right)\right)^2 = \min_{\boldsymbol{d}\in\mathcal{S}_{\boldsymbol{d}}} \min_{\theta_j\in\mathcal{S}_{\boldsymbol{\theta}_j}} \sum_{h=1}^{t}\sum_{j=1}^{k}\sum_{l=1}^{\hat{p}_j} \frac{w_{jlh}^2}{\theta_{jl} d_j}. \tag{16}$$

□

The variable $\boldsymbol{d}$ is shared across all the tasks, indicating if a kernel is useful for any of the tasks. The d_j in the $\boldsymbol{d}$ indicates the weight of the jth kernel function, and the sparse of $\boldsymbol{d}$ indicates the sparse combination of kernels with ℓ_1-norm. The θ_{jl} in $\boldsymbol{\Theta}$ indicates the influence of the jth kernel function to the lth feature. That is, our method not only assigns proper weight to each kernel function by optimizing $\boldsymbol{d}$, but also considers the influence of each kernel to the high dimensional features in the feature space by optimizing $\boldsymbol{\Theta}$.

Now we perform a variable transformation: $\frac{\hat{w}_{jlh}}{\sqrt{\theta_{jl} d_j}} = \bar{w}_{jlh}, l = 1, \dots, d_j$, and let $\boldsymbol{D}_j$ is a diagonal matrix with entries as $\theta_{jl} d_j, l = 1, \dots, \hat{p}_j$. Then, we define $\mathcal{D}$:

$$\mathcal{D} = \begin{bmatrix} \boldsymbol{D}_1 & & \\ & \cdots & \\ & & \boldsymbol{D}_k \end{bmatrix} \tag{17}$$

Using this one can re-write the SMKMTL formulation as:

$$\begin{aligned} \min_{\mathcal{D}} \ & \sum_{h=1}^{t} \min_{\bar{\boldsymbol{W}}_h, \boldsymbol{\xi}} \frac{1}{2}\sum_{j=1}^{k} \|\bar{\boldsymbol{w}}_{j.h}\|_2^2 + \frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h} \xi_{hi}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{k} \bar{\boldsymbol{w}}_{j.h}^{\mathrm{T}} \boldsymbol{D}_j^{\frac{1}{2}} \phi_j(\boldsymbol{x}_{hi}) - y_{hi} = \xi_{hi}, h = 1 \dots t\ , i = 1 \dots n_h \\ & \boldsymbol{D}_j \succeq 0, \sum_{j=1}^{k} \mathrm{Tr}(\boldsymbol{D}_j) \le 1, j = 1 \dots k \end{aligned} \tag{18}$$

The primal formulation (18) can be seen as the following composite objective optimization problem:

$$\begin{aligned} \min_{\mathcal{D}} \quad & J(\mathcal{D}) = \sum_{h=1}^{t} J_h(\mathcal{D}) \\ \text{s.t.} \quad & \boldsymbol{D}_j \succeq 0, \sum_{j=1}^{k} \mathrm{Tr}(\boldsymbol{D}_j) \le 1, j = 1 \dots k \end{aligned} \tag{19}$$

with

$$\begin{aligned} J_h(\mathcal{D}) = & \min_{\bar{\boldsymbol{W}}_h, \boldsymbol{\xi}} \frac{1}{2}\sum_{j=1}^{k} \|\bar{\boldsymbol{w}}_{j.h}\|_2^2 + \frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h} \xi_{hi}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{k} \bar{\boldsymbol{w}}_{j.h}^{\mathrm{T}} \boldsymbol{D}_j^{\frac{1}{2}} \phi_j(\boldsymbol{x}_{hi}) - y_{hi} = \xi_{hi}, i = 1 \dots n_h \end{aligned} \tag{20}$$

For each $J_h(\mathcal{D})$, the Lagrange's theorem is applied to incorporates the constraints into the objective by introducing nonnegative Lagrangian multipliers $\boldsymbol{\alpha}$. The Lagrangian can be written as:

$$\begin{aligned} \mathcal{L}_h = & \frac{1}{2}\sum_{j=1}^{k} \|\bar{\boldsymbol{w}}_{j.h}\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{n_h} \xi_{hi}^2 \\ & + \sum_{i=1}^{n_h} \alpha_{hi}\left(\sum_{j=1}^{k} \bar{\boldsymbol{w}}_{j.h}^{\mathrm{T}} \boldsymbol{D}_j^{\frac{1}{2}} \phi_j(\boldsymbol{x}_{hi}) - y_{hi} - \xi_{hi}\right) \end{aligned} \tag{21}$$

We get:

$$\bar{\boldsymbol{w}}_{j.h}^* = -\boldsymbol{\alpha}_h^{\mathrm{T}} \boldsymbol{D}_j \Phi_{hj} \tag{22a}$$

$$\xi_{hi}^* = \frac{\alpha_{hi}}{\lambda} \tag{22b}$$

where $\boldsymbol{\Phi}_{hj} = [\phi_j(\boldsymbol{x}_{h1}), \dots, \phi_j(\boldsymbol{x}_{hn_h})]$ is the data matrix of the hth task in the jth feature space. Again, we substitute the above expressions for $\boldsymbol{x}_i$ and $\bar{\boldsymbol{w}}$. Thus, we derive the following associated dual problem:

$$\begin{aligned} J_h(\mathcal{D}) = & \max_{\boldsymbol{\alpha}_h} -\boldsymbol{\alpha}_h^{\mathrm{T}} \boldsymbol{y}_t - \frac{1}{2}\boldsymbol{\alpha}_h^{\mathrm{T}}\left(\sum_{j=1}^{k} \boldsymbol{\Phi}_{hj}^{\mathrm{T}} \boldsymbol{D}_j \boldsymbol{\Phi}_{hj}\right)\boldsymbol{\alpha}_h - \frac{1}{2C}\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{\alpha}_h \\ = & \max_{\boldsymbol{\alpha}_h} -\boldsymbol{\alpha}_h^{\mathrm{T}} \boldsymbol{y}_t - \frac{1}{2}\boldsymbol{\alpha}_h^{\mathrm{T}} \mathbf{G}_h(\mathcal{D})\boldsymbol{\alpha}_h - \frac{1}{2C}\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{\alpha}_h \\ \text{s.t.} \quad & \mathcal{D} \succeq 0, \mathrm{Tr}(\mathcal{D}) \le 1 \end{aligned} \tag{23}$$

where

$$\mathbf{G}_h(\mathcal{D}) = \sum_{j=1}^{k} \boldsymbol{\Phi}_{hj}^{\mathrm{T}} \boldsymbol{D}_j \boldsymbol{\Phi}_{hj} = \sum_{j=1}^{k} d_j \boldsymbol{G}_{hj} \tag{24}$$

and $\boldsymbol{G}_{hj}$ is called weighted Gram matrix of the h-task with the j-kernel, which is defined as:

$$\boldsymbol{G}_{hj}(m, n) = \sqrt{\boldsymbol{\theta}_j}\phi_j(\boldsymbol{x}_{hm}) \cdot \sqrt{\boldsymbol{\theta}_j}\phi_j(\boldsymbol{x}_{hn}) \tag{25}$$

where $\boldsymbol{\theta}_j = [\theta_{j1}, \theta_{j2}, \dots, \theta_{j\hat{p}_j}]$.

Our objective function can be transformed into a form similar to that in the MKL formulation in Eq. (6). The differences are: (1) t SVR tasks shared the same matrix$\mathcal{D}$ in our SMKMTL, whereas in $\ell_{2,1}$MKMTL the vector $\boldsymbol{d}$ is shared in t SVR tasks; (2) The $\boldsymbol{K}_j$ in $\ell_{2,1}$MKMTL is a Gram matrix, where each item is the similarity of two instances, whereas $\boldsymbol{G}_j$ in our SMKMTL is weighted Gram matrix, each item in which is the weighted similarity of two instances with a weight vector $\boldsymbol{\theta}_j$.

4.2. Optimization

The optimization of $\ell_{2,1}$SMKMTL is to learn the optimal regression weight of each SVR task and a shared parameter $\bar{\mathcal{D}}$ among the tasks simultaneously. To solve the formulation in Eq. (19), we employ an alternating minimization procedure, in which $\bar{\mathcal{D}}$ is held fix and optimize $\boldsymbol{\alpha_h}$ independently for each task (we call it α-step), and similarly fix $\boldsymbol{\alpha_h}$ in each task and optimize $\bar{\mathcal{D}}$ shared across these tasks(we call it $\bar{\mathcal{D}}$-step). This section describes how to solve the optimization problem for our proposed framework.

1. The $\boldsymbol{\alpha}_h$-step

The optimization of problem (10) with respect to $\boldsymbol{\alpha}_h$ consists in solving t single-task SVR problems while keeping the matrix $\bar{\mathcal{D}}$ fixed. The difficulty in working with this formulation is that the explicit mappings ϕ_js are required. We now describe a way of overcoming this problem and efficiently kernelizing the formulation. Let $\boldsymbol{\Phi}_j \equiv [\boldsymbol{\Phi}_{1j} \dots \boldsymbol{\Phi}_{tj}] \in \mathbb{R}^{\hat{p}\times n}$ and the compact SVD of $\boldsymbol{\Phi}_j$ be $\mathbf{P}_j\Sigma_j\mathbf{Q}_j^{\mathrm{T}}$. Then, $\boldsymbol{K}_j = \boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j = \mathbf{Q}_j\Sigma_j^2\mathbf{Q}_j^{\mathrm{T}}$. We can obtain the $\mathbf{Q}_j$ and Σ_j by the Gram matrix. Now, introduce new variables $\bar{\boldsymbol{D}}_j$ such that $\bar{\boldsymbol{D}}_j = \mathbf{P}_j\boldsymbol{D}_j\mathbf{P}_j^{\mathrm{T}}$. Here, $\bar{\boldsymbol{D}}_j \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix (PSD) of size same as rank of $\boldsymbol{\Phi}_j$. The objective function is:

$$\min_{\bar{\mathcal{D}}\succeq 0, \mathrm{Tr}(\bar{\mathcal{D}})\le 1} \quad J(\bar{\mathcal{D}}) = \sum_{h=1}^{t} J_h(\bar{\mathcal{D}}) \tag{26}$$

with

$$\begin{aligned} J_h(\bar{\mathcal{D}}) = &\max_{\boldsymbol{\alpha}_h} -\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{y}_t - \frac{1}{2}\boldsymbol{\alpha}_h^{\mathrm{T}}\left(\sum_{j=1}^{k}\boldsymbol{\Phi}_{hj}^{\mathrm{T}}\boldsymbol{D}_j\boldsymbol{\Phi}_{hj}\right)\boldsymbol{\alpha}_h - \frac{1}{2C}\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{\alpha}_h \\ = &\max_{\boldsymbol{\alpha}_h} -\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{y}_t - \frac{1}{2}\boldsymbol{\alpha}_h^{\mathrm{T}}\left(\sum_{j=1}^{k}\boldsymbol{\Phi}_{hj}^{\mathrm{T}}\boldsymbol{P}_j\bar{\boldsymbol{D}}_j\boldsymbol{P}_j^{\mathrm{T}}\right)\boldsymbol{\alpha}_h - \frac{1}{2\lambda}\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{\alpha}_h \\ = &\max_{\boldsymbol{\alpha}_h} -\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{y}_t - \frac{1}{2}\boldsymbol{\alpha}_h^{\mathrm{T}}\left(\sum_{j=1}^{k}\boldsymbol{\Phi}_{hj}^{\mathrm{T}}\boldsymbol{\Phi}_j\mathbf{Q}_j^{\mathrm{T}}\Sigma_j^{-1}\bar{\boldsymbol{D}}_j\Sigma_j^{-1}\mathbf{Q}_j\boldsymbol{\Phi}_j^{\mathrm{T}}\boldsymbol{\Phi}_{hj}\right) \\ &\boldsymbol{\alpha}_h - \frac{1}{2\lambda}\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{\alpha}_h \\ = &\max_{\boldsymbol{\alpha}_h} -\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{y}_t - \frac{1}{2}\boldsymbol{\alpha}_h^{\mathrm{T}}\left(\sum_{j=1}^{k}\mathbf{Q}_{hj}^{\mathrm{T}}\bar{\boldsymbol{D}}_j\mathbf{Q}_{hj}\right)\boldsymbol{\alpha}_h - \frac{1}{2\lambda}\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{\alpha}_h, \end{aligned} \tag{27}$$

where $\mathbf{Q}_{hj} = \Sigma_j^{-1}\mathbf{Q}_j^{\mathrm{T}}\boldsymbol{\Phi}_j^{\mathrm{T}}\boldsymbol{\Phi}_{hj}$, and the calculation of $\mathbf{Q}_{hj}$ does not require the kernel-induced features explicitly since $\boldsymbol{\Phi}_j^{\mathrm{T}}\boldsymbol{\Phi}_{hj}$ can be solved by kernel trick.

With the $\sum_{j=1}^{k}\mathbf{Q}_{hj}^{\mathrm{T}}\bar{\boldsymbol{D}}_j\mathbf{Q}_{hj}$, the objective function of each task J_h can be obtained by any SVR algorithm.

2. The $\bar{\mathcal{D}}$-step

Eq. (27) can be written as:

$$\begin{aligned} J_h(\bar{\mathcal{D}}) = &\max_{\boldsymbol{\alpha}_h} -\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{y}_t - \frac{1}{2}\mathbf{Tr}(\mathcal{B}\bar{\mathcal{D}}) - \frac{1}{2\lambda}\boldsymbol{\alpha}_h^{\mathrm{T}}\boldsymbol{\alpha}_h \\ = &-\boldsymbol{\alpha}_h^{\mathrm{T}*}\boldsymbol{y}_t - \frac{1}{2}\mathbf{Tr}(\mathcal{B}\bar{\mathcal{D}}) - \frac{1}{2\lambda}\boldsymbol{\alpha}_h^{*\mathrm{T}}\boldsymbol{\alpha}_h^* \\ = &\beta_h^* - \frac{1}{2}\mathbf{Tr}(\mathcal{B}\bar{\mathcal{D}}) \end{aligned} \tag{28}$$

where $\beta_h^* = -\boldsymbol{\alpha}_h^{*\mathrm{T}}\boldsymbol{y}_h - \frac{1}{2\lambda}\boldsymbol{\alpha}_h^{*\mathrm{T}}\boldsymbol{\alpha}_h^*$, and $\mathcal{B}$ is a block diagonal matrix with entries $\boldsymbol{B}_j = \mathbf{Q}_{hj}\boldsymbol{\alpha}_h^*\boldsymbol{\alpha}_h^{*\mathrm{T}}\mathbf{Q}_{hj}^{\mathrm{T}}$.

With the parameters β_h^* of each SVR model obtained, the $\bar{\mathcal{D}}$-step is to solve the following optimization problem:

$$\min_{\bar{\mathcal{D}}\in\mathcal{S}_{\bar{\mathcal{D}}}} \quad J(\bar{\mathcal{D}}) \tag{29}$$

where $J(\bar{\mathcal{D}}) = \sum_{h=1}^{t}\beta_h^* - \frac{1}{2}\mathbf{Tr}(\mathcal{B}\bar{\mathcal{D}})$, $\mathcal{S}_{\bar{\mathcal{D}}} := \left\{\bar{\mathcal{D}} \succeq 0,\ \mathrm{Tr}(\bar{\mathcal{D}}) \le 1\right\}$. The set $\mathcal{S}$ is called the Spectrahedron, and can be seen as a generalization of the unit simplex to symmetric matrices.

Therefore, the optimization of (29) is a convex optimization problem over the spectrahedron of PSD matrices. Minimizing a convex function over the spectrahedron is an important optimization problem with many applications in machine learning. A standard method to solve the problem is the projected gradient projection algorithm, which generates iteratively the sequence $\left\{\bar{\mathcal{D}}^{(m)}\right\}$ via

$$\begin{aligned} \bar{\mathcal{D}}^{(t+1)} = \Pi_{\mathcal{S}_{\bar{\mathcal{D}}}}(\bar{\mathcal{D}}^{(m)} - s_k\nabla J(\bar{\mathcal{D}})) = \arg\min_{\bar{\mathcal{D}}}\bar{\mathcal{D}}^T\nabla J(\bar{\mathcal{D}}) \\ + \frac{1}{2}\left\|\bar{\mathcal{D}} - \bar{\mathcal{D}}^{(m)}\right\|_2^2, \end{aligned} \tag{30}$$

where s_k a stepsize, and $\Pi_{\mathcal{S}_{\bar{\mathcal{D}}}}(\bar{\mathcal{D}})$ is the Euclidean projection onto $\mathcal{S}_{\bar{\mathcal{D}}}$.

In the procedure of projected gradient descent, the projection of PSD matrices requires the computation of a complete eigenvalue-decomposition, which is a costly step [44]. An alternative is the conditional gradient method (aka Frank–Wolfe algorithm) [44–46] considers the linearization of the objective function, and moves towards a minimizer of this linear function in each iteration. Specifically, it involves two steps in each iteration:

$$\begin{aligned} &\textbf{Compute} \quad \boldsymbol{v}^{(m)} := \mathbf{argmax}_{\boldsymbol{v}'\in\boldsymbol{S}}\left\langle \boldsymbol{v}', -\nabla f(\boldsymbol{x}^{(m)})\right\rangle \\ &\textbf{Update} \quad \boldsymbol{x}^{(t+1)} := (1-\gamma)\boldsymbol{x}^{(m)} + \gamma\boldsymbol{v}^{(m)} \end{aligned} \tag{31}$$

At a current position $\boldsymbol{x}^{(m)}$, the algorithm considers the linearization of the objective function, and moves towards a minimizer of this linear function, $\boldsymbol{v}^{(m)}$. Then, $\boldsymbol{v}^{(m)}$ is chosen as the next step-direction.

Lemma 2. *For any symmetric matrix $\boldsymbol{A} \in \mathbb{S}_{n+}$, it holds that*

$$\max_{\boldsymbol{X}\in\mathbb{S}_{n+}} \langle \boldsymbol{X}, \boldsymbol{A}\rangle = \lambda_{max}(\boldsymbol{A}) \tag{32}$$

For the optimization over the spectrahedron of PSD matrices, Frank–Wolfe algorithm simply requires a largest eigenvector computation instead of a complete SVD per iteration instead of the costly projection step according to Lemma 2 proposed in [44], which is much more efficient. Let $\boldsymbol{v}$:=**ApproxEV**($\boldsymbol{A}$, ξ) is an approximate eigenvalue solver to relax the requirement of exactly solving the linearized problem in each step, the function of **ApproxEV**($\boldsymbol{A}$, ξ) return a unit length vector $\boldsymbol{v}$ such that $\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} \ge \lambda_{\max}(\boldsymbol{A}) - \xi$. It approximates the maximum eigenvector to matrix $\boldsymbol{A}$ with the desired accuracy ξ. According to the Lemma 2, $\boldsymbol{v}$ approximates the linearized problem, that is

$$\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} = \left\langle \boldsymbol{v}\boldsymbol{v}^T, \boldsymbol{A}\right\rangle \ge \lambda_{max}(\boldsymbol{A}) - \xi, \tag{33}$$

Here, the Lanczos'algorithm [47] is used as the ApproxEV. $\boldsymbol{V}^{(m)} = \boldsymbol{v}^{(m)}\boldsymbol{v}^{(m)T}$ is the best descent direction on the linear approximation to $J(\cdot)$ at $\bar{\mathcal{D}}^{(m)}$. During the optimization procedure, the algorithm considers the linearization of the objective function at a current position $\bar{\mathcal{D}}^{(m)}$, and moves towards a minimizer of this linear function by the approximate function of **ApproxEV** over the domain of $\mathbb{S}_{n+}$. For our convex optimization problem over the spectrahedron, the procedure consists of the two steps as below:

$$\begin{aligned} &\textbf{Compute} \quad \boldsymbol{v}^{(m)} := \mathbf{ApproxEV}(\nabla J(-\bar{\mathcal{D}}^{(m)}), \xi^{(m)}), \quad \xi^{(m)} = \frac{C_f}{m^2} \\ &\textbf{Update} \quad \bar{\mathcal{D}}^{(m+1)} := (1-\eta_t)\bar{\mathcal{D}}^{(m)} + \eta_t\boldsymbol{V}^{(m)}, \\ &\qquad \boldsymbol{V}^{(m)} = \boldsymbol{v}^{(m)}\boldsymbol{v}^{(m)T}, \eta = \min\left\{1, \frac{2}{m}\right\}, \end{aligned} \tag{34}$$

where $\nabla J(\bar{\mathcal{D}}^{(m)}) = -\frac{1}{2}\mathcal{B}$, C_f is a curvature constant, the assumption of which is similar to a Lipschitz assumption on the gradient of f, η is a step-size.

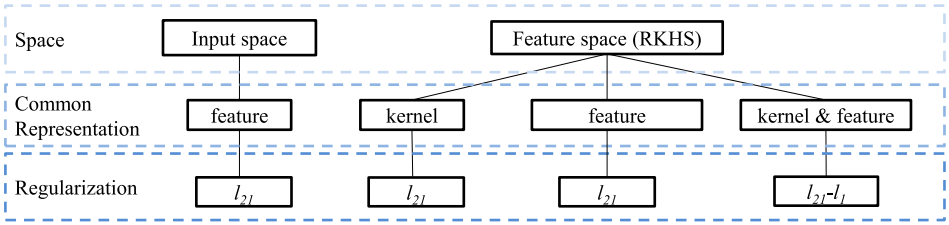


Fig. 3. The $\ell_{2,1}$ norm regularized multi-task learning with different schemes with respect to working space, common representation and regularization.

Instead of using the pre-defined step-sizes, we find the optimal point $\eta \in [0, 1]$ on the line segment between the current iterate $\bar{\mathcal{D}}$ and $\boldsymbol{V}^{(m)}$ to improve the numerical stability as follow:

$$\eta_m := \arg\min_{\eta_m \in [0,1]} J(\bar{\mathcal{D}}^{(m)} + \eta_m(\boldsymbol{V}^{(m)} - \bar{\mathcal{D}}^{(m)})) \tag{35}$$

The detailed algorithm is summarized in Algorithm 1. Given $\bar{\mathcal{D}}$,

Algorithm 1 The two step alternative optimization of $\ell_{2,1} - \ell_1$SMKMTL.

Input: Training Data $\boldsymbol{X}$ and $\boldsymbol{Y}$, regularization parameters λ
Output: $\boldsymbol{\alpha}^*$ and $\bar{\mathcal{D}}$

Initialize $\bar{\mathcal{D}}^{(0)} = \boldsymbol{v}_0\boldsymbol{v}_0^T$ (with trace one), $m = 1$;
repeat
  for $h = 1$ to t **do**
    With fixed $\bar{\mathcal{D}}$, compute $\boldsymbol{\alpha}_h^*$ by using an SVR solver
  end for
  Compute $\beta^* = -\boldsymbol{\alpha}_h^{*\mathrm{T}}\boldsymbol{y}_t - \frac{1}{2\lambda}\boldsymbol{\alpha}_h^{*\mathrm{T}}\boldsymbol{\alpha}_h^*$
  $\xi^{(m)} = \frac{C_f}{m^2}$
  Compute $\boldsymbol{v}^{(m)} := \mathbf{ApproxEV}(-\nabla J(\bar{\mathcal{D}}^{(m)}, \xi^{(m)})$
  $\boldsymbol{V}^{(m)} = \boldsymbol{v}^{(m)}\boldsymbol{v}^{(m)T}$
  Calculate $\eta_m := \arg\min_{\eta_m \in [0,1]} J(\bar{\mathcal{D}} + \eta_m(\boldsymbol{V}^{(m)} - \bar{\mathcal{D}}^{(m)}))$
  Update $\bar{\mathcal{D}}^{(m+1)} := (1 - \eta_m)\bar{\mathcal{D}}^{(m)} + \eta_m\boldsymbol{V}^{(m)}$
  $m = m + 1$
until convergence criterion is satisfied

the problem is equivalent to solving t SVR problems $J_h(\bar{\mathcal{D}})$ individually. The $\bar{\mathcal{D}}$ are learnt using Frank-Wolfe optimization and are shared across the tasks.

4.3. Connection to the existing methods

In this section, our aim is to review the popular $\ell_{2,1}$-norm regularized MTL methods to deal with this problem, as well as to present a taxonomy where these techniques can be categorized depending on three dimensions: space, common representation and regularization. Fig. 3 shows the proposed taxonomy.

Following the taxonomy, the introduced algorithms are distinguished into four families:

1. Linearized MTL (e.g. $\ell_{2,1}$MTL): only focuses on the exploitation of shared representation with respect to the MRI features in the original input space.

2. Single-kernel based MTL (e.g. CORNLIN): maps the original features into a high order features by a specific polynomial kernel function, and learn the relevant features by considering the intra-block correlations with $\ell_{2,1}$-norm.

3. Multi-kernel based MTL with kernel-wise correlation (e.g. $\ell_{2,1}$MKMTL): is a simple extension of the standard MKL to the case of multiple tasks. The method impose an $\ell_{2,1}$ norm on the kernel weights, to learn a optimal kernel combination suited for all the given tasks.

4. Multi-kernel based MTL with both feature-wise and kernel-wise correlation (e.g. $\ell_{2,1} - \ell_1$SMKMTL): assumes a common high dimensional features in the kernel space and kernel representation are shared simultaneously across the tasks.

The first two families belong to the multi-task learning with feature representation sharing. However, the limitations of them are that $\ell_{2,1}$MTL neglects the inherently nonlinear relationship between MRI and cognitive outcomes, and CORNLIN only considers one fixed kernel mapping and ignores the kernel selection. Hence, it is necessary to employ multiple kernels for multiple tasks. The third family belongs to the MTL with model parameter sharing by modeling only individual kernel-wise correlation. It neglects the correlation among multiple related tasks in the feature space. By contrast, our MTL model with kernel-wise and feature-wise correlation can learn the common representation (features and kernel parameters) shared across tasks, so as to maximize task relatedness.

Fig. 4 shows the difference of the schematic diagram of these families, and Fig. 5 illustrates the difference of regularization and space. The first three families only employ the $\ell_{2,1}$-norm on the features or kernel parameters. On the contrary, the proposed mixed sparsity-inducing norms emphasize structured sparsity from both kernel-wise and feature-wise in kernel induced space, to capture the kernel-wise relationships among multiple tasks without ignoring the feature-wise correlation within each kernel space. To the best of our knowledge, there is no sparsity-based algorithms exploiting both the high dimensional feature-wise and the kernel-wise correlation.

4.4. Extension of $\ell_{2,1} - \ell_1$ SMKMTL

Our proposed SMKMTL is a general framework. In this section, we extend the formulation to (1) classification model by displacing the loss with hinge loss; and (2) more choices of regularizers with different choice of kernel setting.

1. Extension to classification

Any suitable loss function can be used in the formulation (10), such as hinge loss for classification problems. In recent years, there has been a great interest in developing classification models to identify clinical labels such as AD, MCI, and Normal Control (NC). The modified primal formulation of classification is:

$$\begin{aligned} &\min_{\widehat{\boldsymbol{W}},\xi} \frac{1}{2}\left(\sum_{j=1}^{k}\left(\sum_{l=1}^{\hat{p}_j}\|\hat{\boldsymbol{w}}_{jl.}\|_2\right)\right)^2 + \frac{\lambda}{2}\sum_{t=1}^{T}\sum_{i=1}^{n_h}\xi_{hi}^2 \\ &\text{s.t. } y_{hi}\left(\sum_{j=1}^{k}\hat{\boldsymbol{w}}_{hj}^{\mathrm{T}}\phi_j(\boldsymbol{x}_{hi}) + b\right) \geq 1 - \xi_{hi}, h = 1\ldots t,\ i = 1,\ldots,n_h \end{aligned} \tag{36}$$

2. Extension to other kernel setting

Besides the above two variant methods, any other choices of kernel setting will derive into new formulations for SMKMTL. Different choices of kernel setting lead to different regularizers. We present several variations of SMKMTL, and we investigate their performance empirically. It is easy to see that some simple variant can induce several well-known algorithms

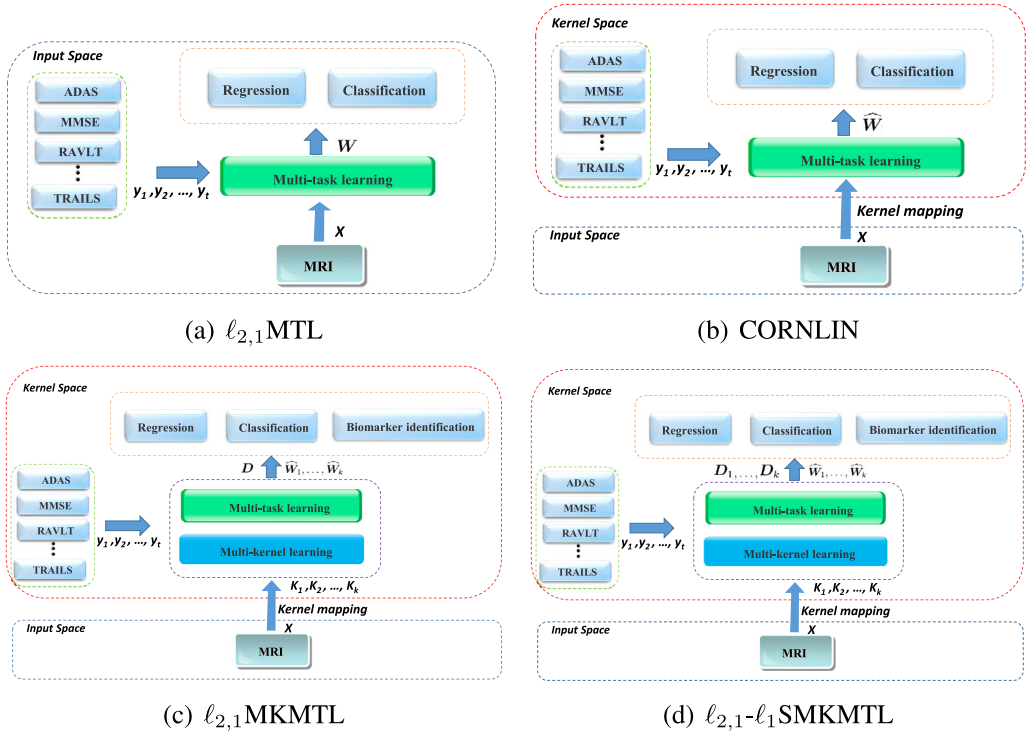


Fig. 4. The schematic diagram of the four MTL frameworks.

(i) Multi-kernel single task learning:

The objective function can be written as follows:

$$\begin{aligned} \min_{\widehat{\boldsymbol{W}},\xi} \ & \frac{1}{2}\left(\sum_{j=1}^{k}\|\hat{\boldsymbol{w}}_j\|_2\right)^2 + \frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{hi}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{k}\hat{\boldsymbol{w}}_j^{\mathrm{T}}\phi_j(\boldsymbol{x}_i) - y_i = \xi_i, \ \ i = 1\ldots n \end{aligned} \tag{37}$$

which is equivalent to the traditional multi-kernel regression model in [48].

(ii) Single-kernel multi-task learning with nonlinear kernel mapping:

When the amount of kernel is reduced to single one, the optimization problem in Eq. (10) is re-defined as:

$$\begin{aligned} \min_{\widehat{\boldsymbol{W}},\xi} \ & \frac{1}{2}\left(\sum_{l=1}^{\hat{p}}\|\hat{\boldsymbol{w}}_{l.}\|_2\right)^2 + \frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{hi}^2 \\ \text{s.t.} \quad & \hat{\boldsymbol{w}}_{.h}^{\mathrm{T}}\phi(\boldsymbol{x}_{hi}) - y_{hi} = \xi_{ti} \ , h = 1\ldots t, \ \ i = 1\ldots n \end{aligned} \tag{38}$$

The method is named as $\ell_{2,1}$KMTL. In the special case where the linear kernel is used, the formulation becomes:

$$\begin{aligned} \min_{\boldsymbol{W},\xi} \ & \frac{1}{2}\left(\sum_{l=1}^{p}\|\boldsymbol{w}_{l.}\|_2\right)^2 + \frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{hi}^2 \\ \text{s.t.} \quad & \boldsymbol{w}_{.h}^{\mathrm{T}}\boldsymbol{x}_{hi} - y_{hi} = \xi_{hi} \ , h = 1\ldots t, \end{aligned} \tag{39}$$

which reduce to the linearized $\ell_{2,1}$MTL.

(iii) Multi-kernel multi-task learning with feature-wise kernel:

We propose to apply the kernel function on each single feature. In this way, we can select the features that contribute most to constructing a better prediction model through ℓ_1 regularization on the kernels' weight vector.

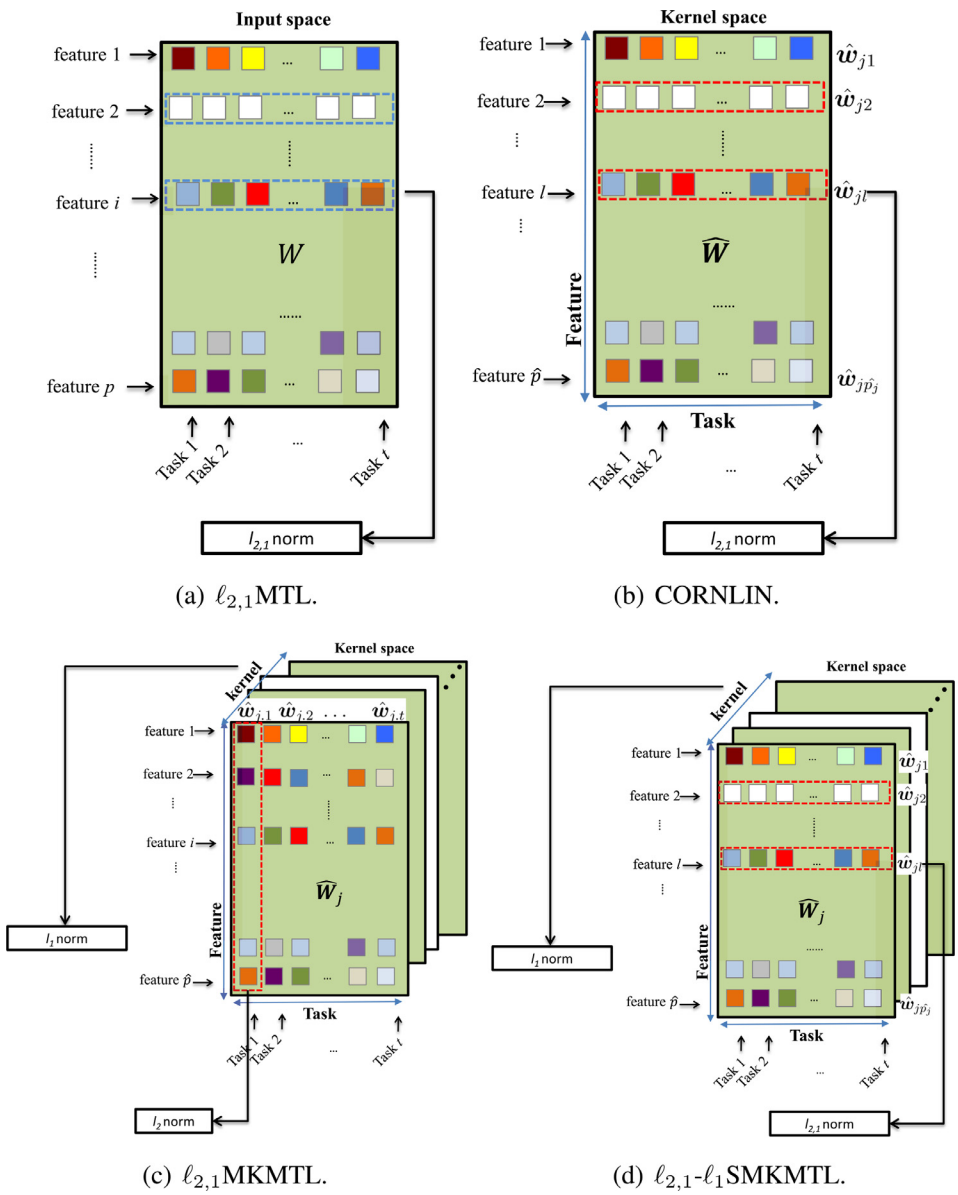
$$\begin{aligned} \min_{\widehat{\boldsymbol{W}},\xi} \ & \frac{1}{2}\left(\sum_{j=1}^{p}\left(\sum_{l=1}^{\hat{p}_j}\|\hat{\boldsymbol{w}}_{jl.}\|_2\right)\right)^2 + \frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{hi}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{k}\hat{\boldsymbol{w}}_{j.h}^{\mathrm{T}}\phi_j(\boldsymbol{x}_{hi}) - y_{hi} = \xi_{ti} \ , h = 1\ldots t, \ \ i = 1\ldots n_h \end{aligned} \tag{40}$$

(iv) Multi-kernel multi-task learning with ROI-wise kernel:

In our data, multiple specific geometric features (volume, area and thickness) are extracted to describe the same ROI in the brain, leading to that features exhibit certain intrinsic group structures. Our previous study proposed a prior knowledge guided multi-task feature learning model, using the group information to enforce the intra-group similarity, has been demonstrated the multiple shape measures tend to be selected together as joint predictors [41]. It is desired to explore and utilize such interrelation structures and select these important and structurally correlated features together.

Based on the above motivation, each individual ROI (brain region) is associated with a specific kernel function, thus, there is a total of g base kernels (g: the number of ROIs). For AD, the number of features in each group is 1 or 4, and the number of groups g can be in the hundreds. Let $\boldsymbol{x}_i' = [\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, \ldots, \boldsymbol{x}_i^{(g)}]$, where $\boldsymbol{x}^{(c)}$ indicates the cth representation of $\boldsymbol{x}$ from the cth ROI.

$$\begin{aligned} \min_{\hat{\boldsymbol{w}},\xi} \ & \frac{1}{2}\left(\sum_{j=1}^{g}\left(\sum_{l=1}^{\hat{p}_j}\|\hat{\boldsymbol{w}}_{jl.}\|_2\right)\right)^2 + \frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{ti}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{g}\hat{\boldsymbol{w}}_{j.h}^{\mathrm{T}}\phi_j(\boldsymbol{x}_{hi}') - y_{hi} = \xi_{hi} \ , h = 1\ldots t, \ \ i = 1\ldots n_h \end{aligned} \tag{41}$$



(a) $\ell_{2,1}$MTL. (b) CORNLIN.

(c) $\ell_{2,1}$MKMTL. (d) $\ell_{2,1}$-ℓ_1SMKMTL.

Fig. 5. The different sparsity-inducing norms of the four $\ell_{2,1}$-norm regularized MTL families.

In the special case where the linear kernel is used, the formulation becomes:

$$\min_{\hat{\boldsymbol{w}},\boldsymbol{\xi}} \frac{1}{2}\left(\sum_{j=1}^{g}\|\boldsymbol{w}_{j.}\|_2\right)^2 + \frac{\lambda}{2}\sum_{h=1}^{t}\sum_{i=1}^{n_h}\xi_{ti}^2$$
$$\text{s.t.} \quad \sum_{j=1}^{g}\boldsymbol{w}_{j.h}^{\mathrm{T}}\boldsymbol{x}'_{hi} - y_{hi} = \xi_{hi}\ , h = 1\ldots t, \tag{42}$$

where $\boldsymbol{w}_{j.}$ is a submatrix with 4 or 1 rows and h columns.

5. Experiment

In this experiment we evaluated the effectiveness of our proposed $\ell_{2,1} - \ell_1$ norm regularized SMKMTL algorithm. Around our research problem, we consider the following questions in our analysis, which are also the contributions of this paper: (1) What is the performance of the nonlinear method compared with the linear $\ell_{2,1}$-norm MTL, the nonlinear $\ell_{2,1}$MKMTL and other MTL methods with different assumption? No previous studies have systematically and extensively examined the prediction performance by linearized MTL and nonlinear kernelized MTL methods with the same $\ell_{2,1}$-norm. (2) How to nonlinearly identify the relevant biomarkers by the multi-kernel based methods? (3) How is the learning capacity of the multi-kernel framework on fusing multi-modality data? To investigate these questions, we first perform extensive experimental analysis to evaluate the performance of our proposed $\ell_{2,1} - \ell_1$ norm regularized SMKMTL.

5.1. Dataset and parameter settings

5.1.1. Dataset

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 affine registration, B1 bias correction, skull stripping and volumetric alignment with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) according to the atlas generated in [49]. During the MRI preprocessing, an iterative algorithm is per-

formed to compute the probabilities of each voxel until the probabilities do not change between two consecutive iterations. Totally, 48 cortical regions and 44 subcortical regions are generated. 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 region, subcortical volume was calculated as features. The SA of left and right hemisphere and total intracranial volume (ICV) were also included. This yielded a total of $p = 319$ MRI features extracted from cortical/subcortical ROIs in each hemisphere. (including 275 cortical and 44 subcortical features from 115 brain ROI totally, see Table S1). Details of the analysis procedure are available at http://adni.loni.ucla.edu/research/mri-post-processing/.

The ADNI project is a longitudinal study, where selected subjects are categorized into three baseline diagnostic groups: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD), repeatedly over a 6-month or 1-year interval. The date when the subjects are scheduled to perform the screening becomes baseline (BL) after approval and the time point for the follow-up visits is denoted by the duration starting from the baseline. In this work, we further performed the following preprocessing steps:

- remove features with more than 10% missing entries (for all patients and all time points);
- remove the ROI whose name is unknown;
- remove the instances with missing value of cognitive scores;
- exclude patients without baseline MRI records;
- complete the missing entries using the average value.

The preprocessing steps yield 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, $n_3 = 173$).

Ten widely used clinical/cognitive assessment scores [5,6] 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 of the first 5 learning trials (TOTAL), Trial 6 total number of words recalled(TOT6), 30 minutes delay score (T30) and 30 min delay recognition score (RECOG), FLU involving Animal Total score (ANIM) and Vegetable Total score (VEG), and TRAILS including Trail Making test A score and B score.

5.1.2. Parameter settings

Following the settings of previous MKL studies, the candidate kernels are: six different kernel bandwidths $(2^{-2}, 2^{-1}, \ldots, 2^3)$, polynomial kernels of degree 1–3, and a linear kernel, which totally yields 10 kernels. All these kernels are applied on all the features. Each base kernel matrices were pre-computed and normalized to have unit trace. Moreover, we take advantage of warm-start techniques for successive SVR retrainings. The gradient based approach produces estimates of $\bar{\mathcal{D}}$ on a smooth trajectory, so that the previous SVR solution provides a good guess for the current SVR training.

The training instances are normalized to be of zero mean and unit variance, and the test instances are also normalized using the same mean and variance of the training data. We use 10-fold cross valuation to evaluate our model and conduct the comparison. In each of trials, a 5-fold nested cross validation procedure for all the comparable methods in our experiments is employed to tune the regularization parameters. The regularization parameter of λ is chosen by nested cross-validation strategy on the training data (trying values $10^{-2}, 10^{-1}, \ldots, 10^2, 10^3$) in this study. The reported results were the best results of each method with the optimal parameter. Data was z-scored before applying regression methods. For the quantitative performance evaluation, we employed the metrics of Correlation Coefficient (CC) and Root Mean Squared Error (rMSE) between the predicted clinical scores and the target clinical scores for each regression task. Moreover, to evaluate the overall performance on all the tasks, the normalized mean squared error (nMSE) [11,17] and weighted R-value (wR) [50] are used. The nMSE and wR are defined as follows:

$$\mathrm{nMSE}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \frac{\sum_{h=1}^{t} \frac{\|\boldsymbol{Y}_h - \hat{\boldsymbol{Y}}_h\|_2^2}{\sigma(\boldsymbol{Y}_h)}}{n}, \tag{43}$$

$$\mathrm{wR}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \frac{\sum_{h=1}^{t} \mathrm{Corr}(\boldsymbol{Y}_h, \hat{\boldsymbol{Y}}_h) n_h}{n}, \tag{44}$$

where $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ are the ground truth cognitive scores and the predicted cognitive scores, respectively.

A smaller (higher) value of nMSE and rMSE (CC and wR) represent better regression performance. We report the mean and standard deviation based on 10 iterations of experiments on different splits of data for all comparable experiments.

5.2. Experiment I: the comparison with the baseline methods

In this section, we conduct empirical evaluation for the proposed methods by comparing with three single task learning methods: Lasso, Ridge and MKL, all of which are applied independently on each task. Moreover, we compare our method with linear $\ell_{2,1}$MTL and one variation of our $\ell_{2,1} - \ell_1$SMKMTL, $\ell_{2,1}$KMTL (single-kernel multi-task learning with $\ell_{2,1}$-norm) as the baseline comparable method. Moreover, we compare the state-of-the-art nonlinear multi-task learning methods, including $\ell_{2,1}$MKMTL in [27] and CORNLIN in [5], which are closest in spirit with our proposed model. The average and standard deviation of performance measures are calculated by 10 fold cross validation, and are shown in Table 1. It is worth noting that we use the same training and testing data across the experiments for all the methods for fair comparison.

Experimental results are reported in Table 1 where the best results are boldfaced. Additionally, a statistical analysis is performed on the results of nMSE and wR and reported in Table 1. 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) All the compared multi-task learning methods ($\ell_{2,1}$MTL, KMTL and $\ell_{2,1} - \ell_1$SMKMTL) improve the predictive performance over the independent regression algorithms(Ridge, Lasso and MKL). This justifies the motivation of learning multiple tasks simultaneously, and verifies that the tasks are not independent and capturing their relatedness can improve learning performance.
(2) Additionally, the proposed $\ell_{2,1} - \ell_1$SMKMTL achieved the highest prediction performance with respect to nMSE and CC, which demonstrates that using the nonlinear model regularized with the mixed norm has the potential to better capture the complex relationship between brain structure and cognitive decline, and the representation involving features and kernels shared by tasks are effectively captured by the mixed sparsity inducing norm.
(3) Compared with linearized regression model, regardless of single task learning or multi-task learning, kernelized regression model achieve better performance at the most cases, which demonstrates that this relationship between predictors (the MRI measures) the responses (cognitive scores) is nonlinear. It also means that the cognitive scores are derived from the high-order of original MRI features or interactions between features. The results show that the

Table 1
Performance comparison of various methods in terms of rMSE, CC, nMSE and wR on ten cognitive prediction tasks. The best results are bolded, and superscript symbols * indicates that $\ell_{2,1} - \ell_1$SMKMTL significantly outperformed that method on that score in terms of nMSE or wR. Student's *t*-test at a level of 0.05 was used.

Task	Metric	Ridge	Lasso	SVR	MKL	$\ell_{2,1}$KMTL	$\ell_{2,1}$MTL	$\ell_{2,1}$MKMTL	CORNLIN	$\ell_{2,1} - \ell_1$SMKMTL
ADAS	rMSE	7.55 ± 0.29	6.84 ± 0.36	7.12 ± 0.59	6.89 ± 0.52	6.91 ± 0.31	6.94 ± 0.43	6.91 ± 0.54	7.81 ± 0.68	**6.80 ± 0.44**
	CC	0.60 ± 0.03	0.65 ± 0.03	0.63 ± 0.07	0.65 ± 0.03	0.62 ± 0.01	0.66 ± 0.04	0.64 ± 0.01	0.49 ± 0.14	**0.69 ± 0.02**
MMSE	rMSE	2.65 ± 0.13	2.21 ± 0.09	2.24 ± 0.17	**2.21 ± 0.15**	2.19 ± 0.11	2.36 ± 0.17	**2.21 ± 0.15**	2.36 ± 0.17	2.37 ± 0.19
	CC	0.41 ± 0.04	0.53 ± 0.04	0.54 ± 0.03	0.54 ± 0.05	0.51 ± 0.02	0.56 ± 0.06	0.53 ± 0.06	0.42 ± 0.11	**0.57 ± 0.05**
TOTAL	rMSE	11.41 ± 0.50	10.02 ± 0.55	9.97 ± 0.58	9.91 ± 0.69	9.73 ± 0.51	**9.61 ± 0.45**	9.82 ± 0.53	10.25 ± 0.70	9.82 ± 0.77
	CC	0.40 ± 0.08	0.49 ± 0.08	0.45 ± 0.04	0.50 ± 0.06	0.53 ± 0.03	0.53 ± 0.08	0.49 ± 0.06	0.42 ± 0.13	**0.59 ± 0.07**
TOT6	rMSE	3.91 ± 0.24	3.32 ± 0.19	3.53 ± 0.32	3.42 ± 0.30	3.29 ± 0.34	3.34 ± 0.15	3.46 ± 0.23	3.51 ± 0.27	**3.11 ± 0.19**
	CC	0.36 ± 0.09	0.51 ± 0.10	0.50 ± 0.12	0.48 ± 0.10	**0.55 ± 0.13**	0.50 ± 0.11	0.43 ± 0.08	0.39 ± 0.16	**0.55 ± 0.10**
T30	rMSE	4.05 ± 0.22	3.44 ± 0.18	3.62 ± 0.37	3.57 ± 0.34	3.55 ± 0.32	3.44 ± 0.15	3.53 ± 0.16	3.72 ± 0.29	**3.42 ± 0.20**
	CC	0.38 ± 0.10	0.52 ± 0.10	0.44 ± 0.12	0.51 ± 0.10	0.52 ± 0.15	0.52 ± 0.10	0.48 ± 0.09	0.37 ± 0.14	**0.53 ± 0.10**
RECOG	rMSE	4.33 ± 0.29	3.64 ± 0.21	3.79 ± 0.22	3.74 ± 0.24	3.69 ± 0.18	3.72 ± 0.25	**3.67 ± 0.21**	3.77 ± 0.26	3.68 ± 0.18
	CC	0.26 ± 0.08	0.42 ± 0.09	0.40 ± 0.07	0.39 ± 0.07	0.43 ± 0.15	0.40 ± 0.09	0.40 ± 0.09	0.33 ± 0.14	**0.52 ± 0.08**
ANIM	rMSE	6.52 ± 0.42	5.35 ± 0.45	5.39 ± 0.40	5.34 ± 0.51	**5.28 ± 0.44**	5.30 ± 0.44	5.39 ± 0.37	5.45 ± 0.25	5.38 ± 0.64
	CC	0.18 ± 0.09	0.37 ± 0.10	0.34 ± 0.16	0.37 ± 0.07	**0.39 ± 0.02**	0.38 ± 0.07	0.35 ± 0.07	0.31 ± 0.13	**0.39 ± 0.07**
VEG	rMSE	4.32 ± 0.18	3.70 ± 0.09	3.88 ± 0.20	3.76 ± 0.14	3.56 ± 0.12	3.70 ± 0.09	3.71 ± 0.10	3.91 ± 0.16	**3.55 ± 0.19**
	CC	0.40 ± 0.07	0.51 ± 0.06	0.42 ± 0.09	0.50 ± 0.07	0.57 ± 0.13	0.57 ± 0.06	0.50 ± 0.06	0.39 ± 0.11	**0.59 ± 0.05**
TR-A	rMSE	27.18 ± 1.70	23.75 ± 1.40	24.77 ± 1.86	24.71 ± 1.78	23.91 ± 1.95	23.42 ± 1.09	23.71 ± 1.64	24.23 ± 1.84	**23.05 ± 1.69**
	CC	0.29 ± 0.10	0.36 ± 0.04	0.32 ± 0.11	0.37 ± 0.06	0.38 ± 0.08	0.40 ± 0.05	0.39 ± 0.07	0.29 ± 0.17	**0.44 ± 0.06**
TR-B	rMSE	83.72 ± 5.71	71.23 ± 2.81	78.87 ± 6.56	78.01 ± 6.92	67.53 ± 6.44	71.32 ± 2.93	72.67 ± 4.05	73.03 ± 4.91	**68.99 ± 1.21**
	CC	0.33 ± 0.11	0.47 ± 0.10	0.44 ± 0.04	0.46 ± 0.06	0.47 ± 0.12	0.47 ± 0.10	0.50 ± 0.08	0.40 ± 0.10	**0.51 ± 0.09**
	nMSE	16.44 ± 1.72*	12.05 ± 0.76*	14.47 ± 1.20*	13.56 ± 1.13*	12.28 ± 1.09*	11.90 ± 0.94*	12.30 ± 0.65*	12.84 ± 1.19*	**10.02 ± 0.34**
	wR	0.36 ± 0.04*	0.48 ± 0.05*	0.45 ± 0.07*	0.48 ± 0.05*	0.49 ± 0.04*	0.49 ± 0.05*	0.47 ± 0.03*	0.38 ± 0.12*	**0.61 ± 0.03**

nonlinear models have the ability to capture the relationship while the linearized models would not be able to detect such non-linear predictive information, thereby leading to limited predictive performance.

(4) With the same $\ell_{2,1}$ regularization to model correlation of tasks in the feature space, SMKMTL outperforms $\ell_{2,1}$KMTL. It indicates that employing multiple base kernels is beneficial in the case of multiple tasks.

(5) we observe that $\ell_{2,1} - \ell_1$SMKMTL obtains a better performance compared with the two nonlinear MTL methods. The results show that while $\ell_{2,1}$MKMTL improves over single-task learning, it suffers by requiring all tasks to share kernels, which may not a good solution to build the relatedness among the tasks. The $\ell_{2,1}$MKMTL method only considers the common kernel representation while not being able to take into account the feature representation. It indicates that simply modeling relatedness based on shared kernel is insufficient. Moreover, $\ell_{2,1} - \ell_1$SMKMTL shows superiority over CORNLIN which indicates the importance of incorporating multi-kernel framework for nonlinearly modeling the relationship. performance of kernelized MTL methods. Both of the two nonlinear kernelized MTL methods consider that tasks to be learned share a common subset of kernel representation or a common subset of the nonlinear features in RKHS by mapping with polynomial kernel function. The result demonstrate the advantages of the proposed multi-kernel multi-task learning by capturing the correlations between tasks from the features representation and kernel representation simultaneously in the tasks of prediction cognitive outcomes. It also suggests that modeling task relatedness using more representation and embedding the appropriate sparsity-inducing norm into the multi-task learning can improve the performance.

We also show the scatter plots of actual values versus predicted values for the score of ADAS, MMSE RAVLT-TOTAL and FLU-VEG on testing data of 10-fold cross validation in Fig. 6.

5.3. Experiment II: kernel setting

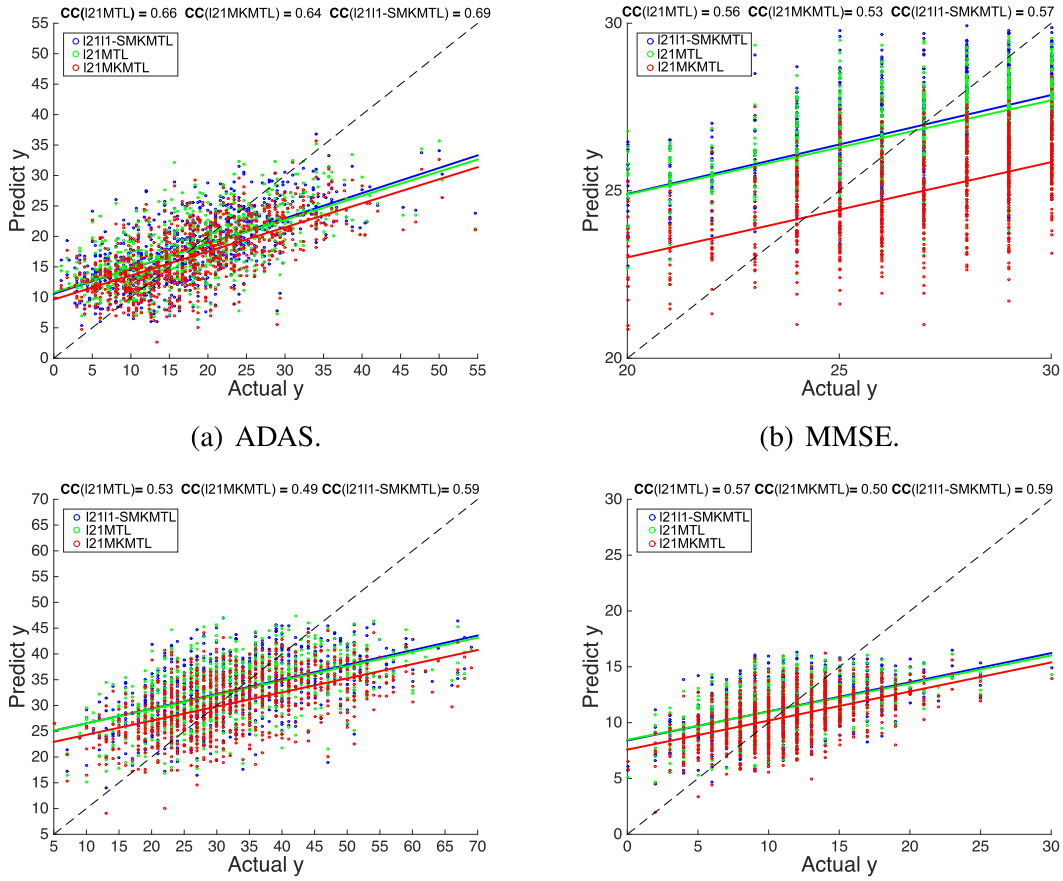
Many empirical studies have shown that the choice of kernel often affects the resulting performance of a kernel method significantly. To evaluate the influence of kernel setting in the SMKMTL model, we investigate the performance with respect to different kernel setting involving kernel scheme, kernel type and kernel regularization. Prediction performance results, measured by nMSE and wR of $\ell_{2,1} - \ell_1$SMKMTL.

Table 2
The performance of $\ell_{2,1} - \ell_1$SMKMTL with respect to the kernel scheme (All or ROI) and kernel function.

Kernel setting	Kernel type	Kernel size	nMSE	wR
ALL	RBF	6	10.55 ± 0.26	0.57 ± 0.06
	Polynomial	3	11.07 ± 0.24	0.45 ± 0.02
	Linear	1	11.01 ± 0.31	0.48 ± 0.04
	Hybrid	10	**10.02 ± 0.34**	**0.61 ± 0.03**
ROI	Linear	115	**9.81 ± 0.26**	**0.66 ± 0.03**
	RBF	115	9.83 ± 0.39	0.59 ± 0.04
	Polynomial	115	10.55 ± 0.45	0.51 ± 0.12

In this section, we employ a wide variety of kernels to evaluate the performance of multi-kernel framework and investigate the proposed variant of our method. There are two choice in the kernel setting: All-scheme and ROI-scheme. In the All-scheme, all the kernel matrix is calculated on the whole set of features; while in the ROI-scheme or ROI-wise kernel, each ROI generates one kernel matrix with the specific features within this group through kernel functions, which totally yields g kernels ($g = 115$ is the number of ROI, please see Table S1). The ROI-scheme can result in sparsity in terms of ROI since each kernel corresponds to each ROI. In the above experiment, the All-scheme was chosen.

The result is shown in Table 2, and we found $\ell_{2,1} - \ell_1$SMKMTL is sensitive with respect to kernel setting. The ROI-scheme consistently outperform the ones ALL-scheme at the most cases, which was expected, as (1) more kernel functions carry sufficient information, and (2) many ROIs are not informative, ROI-scheme is capable of select the discriminative ROIs, making the results easier to interpret. Moreover, besides the hybrid kernel, RBF achieve best performance while polynominal kernel clearly has the worst result in the ALL-scheme. In the ROI-scheme, the linear kernel is the best. The result indicates that the kernel setting in the proposed methods is very crucial for multi-task learning. The inappropriate kernels usually result in sub-optimal or even poor performance.



(a) ADAS. (b) MMSE.

(c) RAVLT-TOTAL. (d) FLU-VEG.

Fig. 6. Scatter plots of actual scores versus predicted scores on testing data for cross-sectional analysis using $\ell_{2,1} - \ell_1$SMKMTL, $\ell_{2,1}$MTL and $\ell_{2,1}$MKMTL based on MRI features. The black dashed line is a reference of perfect correlation (predicted value exactly equals to actual value).

Table 3
Performance comparison of various multi-task learning methods on ten cognitive prediction tasks. The best results are bolded, and superscript symbols * indicates that $\ell_{2,1} - \ell_1$SMKMTL significantly outperformed that method on that score in terms of nMSE or wR. Student's *t*-test at a level of 0.05 was used.

Task	Metric	RMTL	CMTL	Trace	SRMTL	G-SMuRFS	$\ell_{2,1} - \ell_1$SMKMTL
ADAS	rMSE	7.65 ± 0.44	7.64 ± 0.37	8.17 ± 0.60	6.88 ± 0.32	6.71 ± 0.52	$\mathbf{6.80 \pm 0.44}$
	CC	0.58 ± 0.02	0.60 ± 0.02	0.54 ± 0.03	0.65 ± 0.03	0.67 ± 0.05	$\mathbf{0.69 \pm 0.02}$
MMSE	rMSE	3.32 ± 0.26	3.08 ± 0.46	6.11 ± 2.03	2.33 ± 0.27	2.19 ± 0.16	2.37 ± 0.19
	CC	0.33 ± 0.08	0.38 ± 0.04	0.14 ± 0.09	0.52 ± 0.05	0.55 ± 0.08	$\mathbf{0.57 \pm 0.05}$
RAVLT-TOTAL	rMSE	11.01 ± 0.58	11.56 ± 0.51	13.09 ± 3.12	9.96 ± 0.56	9.78 ± 0.50	9.82 ± 0.77
	CC	0.42 ± 0.09	0.39 ± 0.07	0.34 ± 0.17	0.52 ± 0.06	$\mathbf{0.59 \pm 0.08}$	$\mathbf{0.59 \pm 0.07}$
RAVLT-TOT6	rMSE	3.57 ± 0.23	3.90 ± 0.26	3.78 ± 0.49	3.31 ± 0.15	3.39 ± 0.17	$\mathbf{3.11 \pm 0.19}$
	CC	0.43 ± 0.09	0.36 ± 0.09	0.39 ± 0.15	0.50 ± 0.09	0.50 ± 0.059	$\mathbf{0.55 \pm 0.10}$
RAVLT-T30	RMS	3.70 ± 0.17	4.03 ± 0.24	3.9063 ± 0.43	3.44 ± 0.11	3.53 ± 0.21	$\mathbf{3.42 \pm 0.20}$
	CC	0.44 ± 0.09	0.38 ± 0.09	0.40 ± 0.14	0.52 ± 0.10	0.51 ± 0.06	$\mathbf{0.53 \pm 0.10}$
RAVLT-RECOG	rMSE	3.85 ± 0.30	4.38 ± 0.22	4.52 ± 0.85	3.63 ± 0.26	3.59 ± 0.24	3.68 ± 0.18
	CC	0.35 ± 0.10	0.25 ± 0.06	0.25 ± 0.13	0.41 ± 0.09	0.42 ± 0.08	$\mathbf{0.52 \pm 0.08}$
FLU-ANIM	rMSE	5.94 ± 0.39	6.60 ± 0.56	6.74 ± 1.42	5.32 ± 0.33	5.39 ± 0.38	5.38 ± 0.64
	CC	0.25 ± 0.09	0.18 ± 0.08	0.21 ± 0.14	0.36 ± 0.09	0.34 ± 0.11	0.39 ± 0.07
FLU-VEG	rMSE	3.98 ± 0.08	4.39 ± 0.28	4.67 ± 0.77	3.71 ± 0.08	3.67 ± 0.32	$\mathbf{3.55 \pm 0.19}$
	CC	0.44 ± 0.05	0.39 ± 0.07	0.33 ± 0.11	0.50 ± 0.06	0.49 ± 0.05	$\mathbf{0.59 \pm 0.05}$
TRAILS-A	RMS	27.77 ± 1.92	27.45 ± 1.97	28.82 ± 3.27	25.09 ± 1.42	22.91 ± 3.56	$\mathbf{23.05 \pm 1.69}$
	CC	0.28 ± 0.11	0.2871 ± 0.11	0.26 ± 0.11	0.34 ± 0.06	0.42 ± 0.07	$\mathbf{0.44 \pm 0.06}$
TRAILS-B	rMSE	90.12 ± 7.09	83.65 ± 5.41	89.68 ± 7.83	80.00 ± 4.63	72.01 ± 4.39	$\mathbf{68.99 \pm 1.21}$
	CC	0.29 ± 0.12	0.33 ± 0.11	0.29 ± 0.1215	0.36 ± 0.09	0.44 ± 0.073	$0.51 \pm \mathbf{0.09}$
	nMSE	$17.67 \pm 2.30^*$	$16.67 \pm 1.91^*$	$20.22 \pm 5.21^*$	$14.01 \pm 1.16^*$	$13.11 \pm 1.13^*$	$\mathbf{10.02 \pm 0.34}$
	wR	$0.38 \pm 0.03^*$	$0.35 \pm 0.03^*$	$0.31 \pm 0.08^*$	$0.46 \pm 0.04^*$	$0.47 \pm 0.03^*$	$\mathbf{0.61 \pm 0.03}$

5.4. Experiment III: the comparison with the MTL with other assumption

To illustrate how well our $\ell_{2,1} - \ell_1$SMKMTL works by means of modeling the correlation among the tasks, we comprehensively compare our proposed methods with several popular state-of-the-art related methods. Representative multi-task learning algorithms includes:

1. Robust multi-Task Feature Learning (RMTL) [51]: RMTL ($\min_{\boldsymbol{W}} \ L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}) + \lambda_1 \|\boldsymbol{P}\|_* + \lambda_2 \|\boldsymbol{S}\|_{2,1}$ subject to $\boldsymbol{W} = \boldsymbol{P} + \boldsymbol{S}$), which assumes that the model $\boldsymbol{W}$ can be decomposed into two components: a shared feature structure P capturing task relatedness and a group-sparse structure S detecting outliers.
2. Clustered multi-Task Learning (CMTL) [52]: CMTL($\min_{\boldsymbol{W}, M: M^{\mathrm{T}}M = I_c} \ L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}) + \lambda_1 (\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}) - \mathrm{Tr}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{M})) + \lambda_2 \mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})$, where $\boldsymbol{M} \in \mathbb{R}^{c \times t}$ is an orthogonal cluster indicator matrix, and the tasks are clustered into $c < t$ clusters) incorporates a regularization term to induce clustering between tasks and then share information only to tasks belonging to the same cluster. In the CMTL, the number of clusters is set to 5 since the 10 tasks belong to 5 sets of cognitive functions.
3. Trace-Norm Regularized multi-Task Learning (Trace) [53]: Assume that all models share a common low-dimensional subspace ($\min_{\boldsymbol{W}} \ L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}) + \lambda \|\boldsymbol{W}\|_*$.
4. Sparse regularized multi-task learning formulation (SRMTL) [54]: SRMTL ($\min_{\boldsymbol{W}} \ L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}) + \lambda_1 \|\boldsymbol{W}\mathcal{Z}\|_F^2 + \lambda_2 \|\boldsymbol{W}\|_1$, where $\mathcal{Z} \in \mathbb{R}^{k \times k}$) contains two regularization processes: (1) all tasks are regularized by their mean value, and therefore knowledge from one task can be utilized by other tasks via the mean value; (2) sparsity are enforced in the learning with ℓ_1 norm.
5. G-SMuRFS [13]: G-SMuRFS ($\min_{\boldsymbol{W}} \ L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}) + \lambda_1 \|\boldsymbol{W}\|_{2,1} + \lambda_2 \sum_{l=1}^{q} w_l \sqrt{\sum_{j \in \mathcal{G}_l} \|\boldsymbol{w}_{j.}\|_2}$) takes into account coupled feature and group sparsity across tasks.

From the results, we can find that compared with the other multi-task learning with different assumptions, our proposed method belongs to the multi-task feature learning methods with sparsity-inducing norms, have an advantage over the other comparative nonsparse multi-task learning methods. Since not all the brain regions are associated with AD, many of the features are irrelevant and redundant. The results reveal that sparse based MTL methods are appropriate for the task of prediction cognitive measures and better than the non-sparse based MTL method. To our surprise, RMTL, CMTL and Trace performs worse than Ridge method which tells us that these assumptions in these methods may be inappropriate in the cognitive performance prediction.

5.5. Experiment IV: discriminative ROI identification

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. We, then, turn our analysis now to the identification of MRI biomarkers. One of the strengths of the $\ell_{2,1}$MTL and $\ell_{2,1} - \ell_1$SMKMTL formulation is that they facilitate the identification of biomarkers due to the sparse property of $\ell_{2,1}$-norm or ℓ_1-norm. Our $\ell_{2,1} - \ell_1$SMKMTL with ROI-scheme is a ROI-sparse model which is able to identify a compact set of relevant neuroimaging biomarkers from the region level due to the ℓ_1-norm over the kernels, which would provide us with better interpretability of the brain region. The linear kernel is chosen as base kernel function. In order to visualize the distribution of the kernel weight of each ROI, we plot the weights of the g base kernels corresponding to each ROI of $\ell_{2,1} - \ell_1$SMKMTL and $\ell_{2,1}$MTL in Fig. 7. The weight of selected ROI is calculated based on the weight parameters of $\boldsymbol{W}$ and $\tilde{\mathcal{D}}$ from $\ell_{2,1}$MTL and $\ell_{2,1} - \ell_1$SMKMTL, respectively. The weight values indicates the contribution of different ROIs. For $\ell_{2,1}$MTL, the weight is calculated by $\mu_q \sqrt{\sum_{q=1}^{g} \|\boldsymbol{w}_{q.}\|_2}$ for the q-th MRI brain region, where $\mu_g = 1$ for subcortical regions and $\mu_g = 1/2$ for cortical regions.

Table 4
Top 20 ROIs selected and weights optimized by $\ell_{2,1}$MTL and $\ell_{2,1} - \ell_1$SMKMTL.

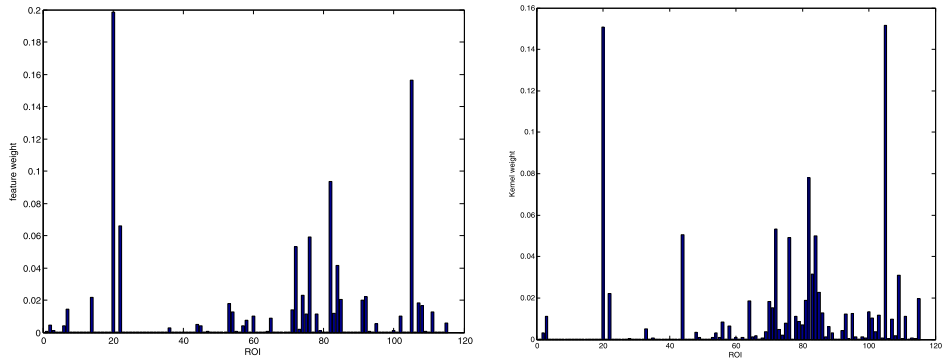
$\ell_{2,1}$MTL		$\ell_{2,1} - \ell_1$SMKMTL	
ROI	Weight	ROI	Weight
L.MidTemporal	0.198	L.MidTemporal	0.156
R.Entorhinal	0.151	R.Entorhinal	0.154
L.Hippocampus	0.093	L.Hippocampus	0.084
L.SupFrontal	0.062	L.InfTemporal	0.052
OpticChiasm	0.048	R.TransvTemporal	0.052
R.LatVent	0.047	R.Fusiform	0.051
R.PostCing	0.046	L.IsthmCing	0.050
L.SupParietal	0.042	L.Precuneus	0.036
WMHypoInt	0.038	L.RostAntCing	0.033
L.InfParietal	0.037	R.PostCing	0.031
L.Insula	0.035	L.SupFrontal	0.028
R.BanksSTS	0.024	R.BanksSTS	0.022
R.TransvTemporal	0.023	L.CaudAntCing	0.021
L.IsthmCing	0.022	R.Paracentral	0.021
R.Lingual	0.021	R.InfTemporal	0.018
R.InfParietal	0.017	L.InfLatVent	0.017
R.MidTemporal	0.016	L.LatVent	0.015
L.InfTemporal	0.014	R.MidTemporal	0.011
R.FrontalPole	0.011	L.Fusiform	0.009
L.TransvTemporal	0.010	L.CaudAntCing	0.009

For $\ell_{2,1} - \ell_1$SMKMTL, contrast to calculation of variable $\boldsymbol{w}$ in the linearized MTL method, the weight in $\ell_{2,1} - \ell_1$SMKMTL is calculated by $\mathrm{Tr}(\tilde{\boldsymbol{D}}_q)$ for the q-th MRI brain region. The multiple Kernel Learning based techniques for nonlinear feature selection have been explored and have been shown to be effective [55]. In this experiment, we empirically investigate the effectiveness of nonlinear ROI selection in our $\ell_{2,1} - \ell_1$SMKMTL. The weights of ROIs in both methods are obtained by calculating the overall weights for all the cognitive tasks.

From Fig. 7, we see that weights obtained by both $\ell_{2,1}$MTL and $\ell_{2,1} - \ell_1$SMKMTL are sparse, which are able to identify a compact set of relevant neuroimaging biomarkers from the region level. The optimized weights corresponding to the region obtained by different formations indicate the importance contribution for the classification. The top-20 region names and the corresponding weights for both methdos are shown in Table 4. From Table 4, we can see that the ROIs selected from linear and kernelized methods are different since different formulations prefer different brain region. However, some regions are common for both methods, such as L.MidTemporal, R.Entorhinal, L.Hippocampus, and L.InfTemporal. These findings are in accordance with the known knowledge that in the pathological pathway of AD. These identified brain regions have been pointed out in the previous literatures and have been also shown to be highly related to clinical functions.

5.6. Experiment V: classification

In this experiment, we extend our regression model to a classification task. In recent years, there has been a great interest in computer-aided diagnosis of AD and its prodromal stage, Mild Cognitive Impairment (MCI). The heterogeneous MCI group involves a mix of individuals, some who will convert to AD (named progressive MCI, pMCI) within 36 months and others who will still be



(a) $\ell_{2,1}$MTL (b) $\ell_{2,1}$-ℓ_1SMKMTL

Fig. 7. The kernel weight of each ROI.

Table 5
Performance comparison of comparable methods on classification in terms of accuracy. Superscript symbols * and † indicate that $\ell_{2,1} - \ell_1$SMKMTL-All and $\ell_{2,1} - \ell_1$SMKMTL-ROI significantly outperformed that method with respect to accuracy. Student's *t*-test at a level of 0.05 was used.

	AD vs. NC	MCI vs. NC	AD vs. MCI	pMCI vs. sMCI
Lasso	$0.80 \pm 0.03^{\dagger *}$	$0.71 \pm 0.02^{\dagger *}$	$0.72 \pm 0.03^{\dagger *}$	$0.66 \pm 0.05^{\dagger *}$
Ridge	$0.78 \pm 0.01^{\dagger *}$	$0.70 \pm 0.02^{\dagger *}$	$0.70 \pm 0.01^{\dagger *}$	$0.65 \pm 0.03^{\dagger *}$
MKL	$0.84 \pm 0.02^{\dagger *}$	$0.70 \pm 0.01^{\dagger *}$	$0.71 \pm 0.02^{\dagger *}$	$0.65 \pm 0.02^{\dagger *}$
$\ell_{2,1}$MTL	$0.83 \pm 0.02^{\dagger *}$	$0.72 \pm 0.03^{\dagger *}$	$0.73 \pm 0.02^{\dagger *}$	$0.69 \pm 0.03^{\dagger}$
$\ell_{2,1}$MKMTL	$0.84 \pm 0.03^{\dagger *}$	0.72 ± 0.02	$0.72 \pm 0.01^{\dagger *}$	$0.67 \pm 0.02^{\dagger}$
$\ell_{2,1} - \ell_1$SMKMTL-All	$0.85 \pm 0.04^{\dagger}$	0.73 ± 0.02	$\mathbf{0.76 \pm 0.03}$	$0.68 \pm 0.02^{\dagger}$
$\ell_{2,1} - \ell_1$SMKMTL-ROI	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.75 \pm 0.03}$	$\mathbf{0.76 \pm 0.02}$	$\mathbf{0.74 \pm 0.02}$

stable (named stable MCI, sMCI). The diagnosis of AD can be formulated as a binary or multi-class classification problem. In this experiment, we considered four binary classification problems: AD vs. NC, MCI vs. NC, AD vs. MCI, and pMCI vs. sMCI, and consider each binary classification problem as a task. Then, the multi-class problem is formulated as a multi-task paradigm that exploits the correlations amongst multiple tasks by learning them simultaneously rather than individually. The previous works have shown that learning multiple related tasks simultaneously can get better results than learning these tasks independently [56,57] for AD diagnosis. To evaluate the performance of our proposed method in disease diagnosis, we compare our proposed methods with other baseline methods.

From the results of Table 5, $\ell_{2,1} - \ell_1$SMKMTL consistently improved the performance of the linearized MTL in all the test cases except the classification of pMCI vs. sMCI, which verifies the benefits of jointly learning from the classification tasks and implies that considering the correlation over the high dimensional feature and kernel function at the same time is capable of uncovering the structure information shared by multiple tasks for multi-task learning.

To show the superior performance of our algorithm, we selected several state-of-the-art classification methods for comparison:

(1) Subspace-based linearized MTL [57]: It combines feature selection and subspace learning in a unified $\ell_{2,1}$MTL framework. It also formulates the multi-class problem as multi-task paradigm that exploits the correlations amongst multiple tasks by learning them simultaneously rather than individually.
(2) MKMFA [58]: It incorporates the Marginal Fisher Analysis with $\ell_{2,1}$-norm based MKL to simultaneously select a subset of the relevant brain regions and learn a dimensionality transformation.

Moreover, few previous works in a recent study use the entire data from ADNI-1 for AD/MCI classification. Tong et al. [59] proposed to use a multiple instance learning method (mi-Graph) with local intensity patches as features for the detection of AD and its MCI. Coup et al. proposed the SNIPE (Scoring by Nonlocal Image Patch Estimator) method with a large amount of non-local patches [60]. Wolz et al. [61] proposed a multi-method to combine multiple features (Hippocampal volume, cortical thickness, manifold-based features, tensor-based morphometry features) on the same dataset. In this paper, we compared the three state-of-the-art methods. To make a more fair comparison, the classification results are obtained using leave-one-out cross validation, the evaluation scheme of which is the same as the works in [59–61]. Note that MKMFA and Subspace-based linearized MTL use the same ROI based feature set as our method. Table 6 shows the performance of each algorithm for two binary classifications with respect to accuracy, sensitivity and specificity.

Experiments indicate that our algorithm is able to attain higher classification performance than other methods. In particular, compared with MKMFA which also uses MKL framework, $\ell_{2,1} - \ell_1$SMKMTL obtained a slightly better result due to the consideration of the useful correlation among multiple classification tasks. Compared with the ROI-based morphological features used by our method, MiGraph and SNIPE use the patch based features which could provide much richer information for the disease. Although direct comparison with the aforementioned studies is not appropriate due to the use of different MRI features, the obtained results validate the promising performance of our method for classification with less features.

5.7. Experiment VI: multi-modalities fusion

The MKL model has been successfully applied to combine multiple modalities in AD studies [20]. Now, we extend $\ell_{2,1} - \ell_1$SMKMTL to the multi-modal case for fusing multiple modali-

Table 6
Comparison of the state-of-the-art methods for AD diagnosis on the dataset from ADNI-1.

Methods	Features	AD vs. NC			pMCI vs. sMCI		
	ACC	SEN	SPE	ACC	SEN	SPE	
$\ell_{2,1} - \ell_1$SMKMTL	ROI-based morphological features	**0.91**	**0.87**	**0.93**	**0.74**	**0.69**	0.71
MKMFA [58]	ROI-based morphological features	0.89	0.86	0.92	0.71	0.68	**0.72**
Subspace-based linearized MTL [57]	ROI-based morphological features	0.82	0.79	0.88	0.64	0.61	0.68
MiGraph [59]	patch based features	0.89	0.85	**0.93**	0.69	0.66	0.71
SNIPE [60]	patch based features	0.89	0.84	**0.93**	0.70	0.69	0.71
Multi-Method [61]	ROI-based morphological features	0.87	0.78	0.95	0.67	**0.69**	0.66

ties data in ADNI dataset. Clinical and research studies commonly demonstrate that complementary brain images for a more accurate and rigorous assessment of the disease status and cognitive function [14,62–64]. To estimate the effect of combining multi-modality data with our $\ell_{2,1} - \ell_1$SMKMTL, we further perform some experiments, which are (1) employing only MRI modality, (2) employing only PET modality, (3) combining two modalities: PET and MRI (MP), and (4) combining three modalities: PET, MRI and demographic information including age, gender, years of education and ApoE genotyping (MPD). We compare the performance of $\ell_{2,1}$MTL and $\ell_{2,1}$MKMTL on the fusing multi-modalities. For $\ell_{2,1}$MTL, the features from multi-modalities are concatenated into a long vector features. In the $\ell_{2,1}$MKMTL and $\ell_{2,1} - \ell_1$SMKMTL, ten different kennel functions described in the first experiment are used for each modality.

Different than the previous experiments, the samples from ADNI-2 are used instead of ADNI-1, since the amount of the patients with PET is sufficient. From the ADNI-2, we obtained all the patients with both MRI and PET, totally 756 samples. The PET imaging data are from the ADNI database processed by the UC Berkeley team, who corrected the raw scans for partial volume effects using the geometric transfer matrix approach and coregistered each florbetapir scan to the corresponding MRI using SPM5. With the same Freesurfer-defined regions as MRI, the mean florbetapir uptake is calculated 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 7 and 8. Note that the amount of tasks is 9 due to the no acquisition of VEG measure for patients in ADNI-2. From the results, it is clear that the method with multi-modality outperforms the methods using one single modality of data. Furthermore, the results show that the most of the improvement of $\ell_{2,1} - \ell_1$SMKMTL is statistically significant. This validates our assumption that the complementary information among different modalities is helpful for cognitive function prediction. Regardless of two or three modalities, the proposed $\ell_{2,1} - \ell_1$SMKMTL achieved better performances than the linear based multi-task learning for the most cases, same as for the single modality learning task above.

Besides the baseline methods, we also compare the closest competing techniques to ours: Multi-modal multi-task learning (M3T) [30], Manifold regularized multi-task feature selection (M2TFS) [28] and the Inter-modality relationship constrained multi-modality multi-task learning (I-M3T) [29]. All of the above methods utilize the multi-kernel learning combined with multi-task learning for multi-modality data. Note that both the M2TFS and I-M3T consider each modality rather than cognitive outcome as a task to feature selection, thus require the same number of features computed from MRI and PET modalities. In order to meet it, we select the same regions from MRI and PET, totally 70 regions remained. M3T employs an $\ell_{2,1}$MTL to selects the common subset of relevant features for multiple features from each modality, and uses a multi-modal support vector machine to fuse the above-selected features from all modalities to predict multiple prediction tasks. Since M2TFS and I-M3T require that each modality contains the same feature dimensionality, we applied them to only MRI and PET (MP) in our experiments. Regardless of the two or three modalities, the proposed method achieves better performances than the three closely related works. This poor regression performance can be attributed to the fact that: (1) M2TFS and I-M3T do not take the correlation of multiple prediction tasks into account; (2) Although M3T incorporates the correlation into MTL to select the discriminative feature subset across multiple tasks, the feature learning and multi-modality fusion are conducted individually. (3) The two-stage strategy employed by the three methods tends to result in inconsistencies and can not achieve a optimally global solution, since the feature learning works in the input feature space while multi-modality fusing operates in the input feature space. The unified framework can be illustrated in Fig. 8. Moreover, we also illustrate the multi-modality data fusing scheme proposed by M3T in Fig. 9 and M2TFS (I-M3T) in Fig. 10, which employ the combination of MKL and MTL. We can clearly find that

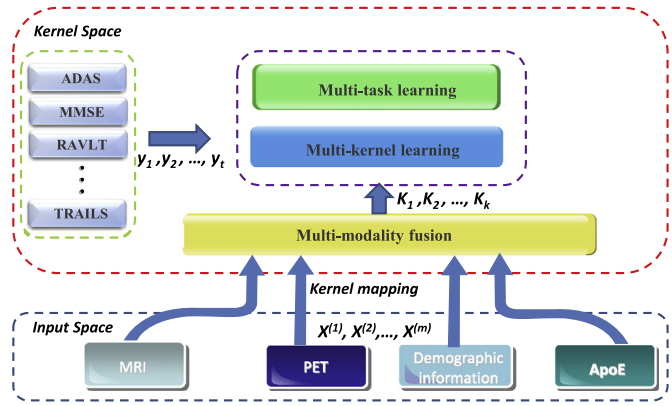


Fig. 8. The illustration of multi-modality data fusing and multi-task learning in our proposed unified framework.

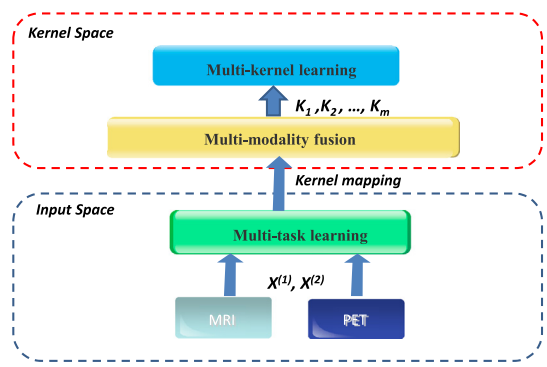


Fig. 9. The illustration of multi-modality data fusing and multi-task learning in the M3T method [30].

Table 7
Performance comparison of various methods with fusing multiple modalities data in terms of rMSE and nMSE. The superscript symbols * indicates that $\ell_{2,1} - \ell_1$SMKMTL-All significantly outperformed that method on that score in terms of nMSE. Student's *t*-test at a level of 0.05 was used.

Method	ADAS	MMSE	FLU	TRAILS	
			ANIM	A	B
$\ell_{2,1}$MTL-MRI	6.49 ± 1.02	1.96 ± 0.30	4.91 ± 0.25	16.39 ± 2.90	55.82 ± 7.68
$\ell_{2,1}$MTL-PET	6.94 ± 1.24	2.11 ± 0.29	5.19 ± 0.14	16.56 ± 3.53	56.88 ± 9.44
$\ell_{2,1}$MTL-MP	6.21 ± 1.03	2.06 ± 0.29	4.92 ± 0.26	16.09 ± 2.76	53.70 ± 7.14
$\ell_{2,1}$MTL-MPD	6.17 ± 0.97	2.06 ± 0.27	**4.78 ± 0.20**	15.97 ± 2.78	53.37 ± 7.24
$\ell_{2,1}$MKMTL-MRI	6.36 ± 0.94	2.07 ± 0.29	4.99 ± 0.23	16.18 ± 3.08	55.95 ± 9.47
$\ell_{2,1}$MKMTL-PET	6.81 ± 1.15	2.06 ± 0.36	5.15 ± 0.22	16.61 ± 3.58	57.85 ± 11.24
$\ell_{2,1}$MKMTL-MP	6.11 ± 0.88	2.00 ± 0.25	4.96 ± 0.26	16.13 ± 2.98	54.13 ± 9.45
$\ell_{2,1}$MKMTL-MPD	**5.96 ± 0.83**	1.95 ± 0.25	4.82 ± 0.22	16.00 ± 3.06	53.48 ± 9.59
$\ell_{2,1} - \ell_1$SMKMTL-MRI	6.42 ± 0.95	1.95 ± 0.30	4.88 ± 0.26	16.11 ± 2.93	54.96 ± 7.49
$\ell_{2,1} - \ell_1$SMKMTL-PET	6.78 ± 1.06	2.05 ± 0.32	5.10 ± 0.25	16.52 ± 3.51	55.51 ± 9.56
$\ell_{2,1} - \ell_1$SMKMTL-MP	6.08 ± 0.98	1.91 ± 0.29	4.85 ± 0.24	15.95 ± 2.99	52.44 ± 8.07
$\ell_{2,1} - \ell_1$SMKMTL-MPD	6.03 ± 0.98	**1.90 ± 0.29**	4.80 ± 0.24	**15.88 ± 3.02**	**52.20 ± 8.12**
M3T-MP [30]	6.20 ± 1.45	2.12 ± 0.24	4.78 ± 0.19	16.45 ± 2.83	53.57 ± 7.26
M3T-MPD [30]	6.15 ± 1.22	2.01 ± 0.18	4.71 ± 0.33	15.70 ± 2.25	52.31 ± 6.96
M2TFS-MP [28]	6.32 ± 1.58	2.25 ± 0.31	4.77 ± 0.15	16.31 ± 2.68	54.45 ± 7.11
I-M3T-MP [29]	6.39 ± 1.66	2.38 ± 0.36	4.84 ± 0.22	16.22 ± 2.18	55.66 ± 7.82
Method	RAVLT				nMSE
	TOTAL	TOT6	T30	RECOG	
$\ell_{2,1}$MTL-MRI	10.18 ± 0.64	3.53 ± 0.14	3.73 ± 0.19	3.16 ± 0.30	10.24 ± 0.73*
$\ell_{2,1}$MTL-PET	10.41 ± 0.44	3.62 ± 0.14	3.79 ± 0.17	3.25 ± 0.36	10.72 ± 1.16*
$\ell_{2,1}$MTL-MP	10.01 ± 0.55	3.50 ± 0.14	3.69 ± 0.19	3.16 ± 0.31	9.71 ± 0.62*
$\ell_{2,1}$MTL-MPD	9.75 ± 0.57	3.45 ± 0.15	3.64 ± 0.20	3.17 ± 0.31	9.52 ± 0.60*
$\ell_{2,1}$MKMTL-MRI	10.09 ± 0.60	3.53 ± 0.08	3.73 ± 0.25	3.20 ± 0.30	10.21 ± 1.01*
$\ell_{2,1}$MKMTL-PET	10.30 ± 0.43	3.59 ± 0.14	3.75 ± 0.23	3.20 ± 0.35	10.82 ± 1.45*
$\ell_{2,1}$MKMTL-MP	9.78 ± 0.37	3.47 ± 0.08	3.66 ± 0.19	3.15 ± 0.30	9.71 ± 0.96*
$\ell_{2,1}$MKMTL-MPD	**9.35 ± 0.46**	3.40 ± 0.03	3.60 ± 0.22	3.19 ± 0.29	9.41 ± 0.98*
$\ell_{2,1} - \ell_1$SMKMTL-MRI	9.98 ± 0.52	3.47 ± 0.13	3.67 ± 0.20	3.14 ± 0.31	9.93 ± 0.75*
$\ell_{2,1} - \ell_1$SMKMTL-PET	10.19 ± 0.41	3.56 ± 0.14	3.74 ± 0.21	3.19 ± 0.35	10.31 ± 1.10*
$\ell_{2,1} - \ell_1$SMKMTL-MP	9.72 ± 0.46	3.39 ± 0.13	3.59 ± 0.16	3.11 ± 0.32	9.28 ± 0.86
$\ell_{2,1} - \ell_1$SMKMTL-MPD	9.56 ± 0.44	3.36 ± 0.12	3.55 ± 0.17	**3.10 ± 0.32**	**9.16 ± 0.86**
M3T-MP [30]	9.84 ± 0.41	3.33 ± 0.19	3.52 ± 0.20	3.32 ± 0.39	9.55 ± 0.58*
M3T-MPD [30]	9.77 ± 0.52	**3.21 ± 0.24**	**3.49 ± 0.12**	3.29 ± 0.31	9.42 ± 0.55*
M2TFS-MP [28]	9.91 ± 0.35	3.37 ± 0.28	3.76 ± 0.24	3.12 ± 0.44	9.49 ± 0.41*
I-M3T-MP [29]	9.80 ± 0.30	3.45 ± 0.30	3.79 ± 0.21	3.33 ± 0.47	9.59 ± 0.54*

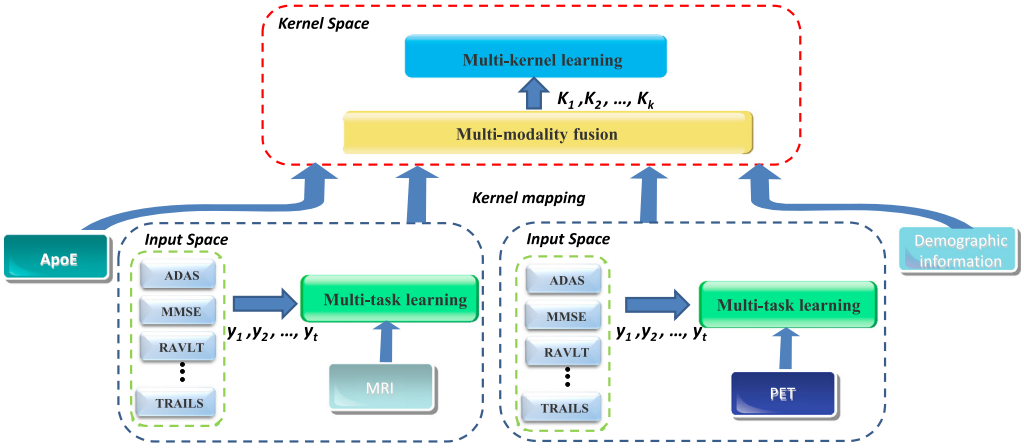


Fig. 10. The illustration of multi-modality data fusing and multi-task learning in the M2TFS [28] (I-M3T [29]) method.

the MKL and MTL are conducted individually and performed in the different space.

6. Discussion and future directions

This paper presents a novel mixed sparsity-inducing norm regularized nonlinear multi-task learning approach for brain cognitive performance prediction. In short, the main contributions of this paper are summarized below:

1. We propose a general unifying multi-kernel framework to build nonlinear relationship the MRI features and cognitive outcomes, and model the relationship among the prediction tasks in the kernel induced feature space. Our method not only assigns proper weights to each kernel function but also considers the importance of the features in the feature space induced by these base kernel functions.

2. We propose a mixed sparsity-Inducing $\ell_{2,1} - \ell_1$ norm to regularize the multiple kernel multi-task learning formulation.

Table 8
Performance comparison of various methods with fusing multiple modalities data in terms of CC and wR. Significantly outperformed that method on that score in terms of wR. Student's *t*-test at a level of 0.05 was used.

Method	ADAS	MMSE	FLU	TRAILS	
			ANIM	A	B
$\ell_{2,1}$MTL-MRI	0.67 ± 0.09	0.53 ± 0.11	0.48 ± 0.11	0.41 ± 0.11	0.52 ± 0.07
$\ell_{2,1}$MTL-PET	0.61 ± 0.05	0.48 ± 0.08	0.39 ± 0.10	0.38 ± 0.12	0.50 ± 0.06
$\ell_{2,1}$MTL-MP	0.70 ± 0.07	0.54 ± 0.10	0.48 ± 0.11	0.43 ± 0.11	0.56 ± 0.07
$\ell_{2,1}$MTL-MPD	0.70 ± 0.06	0.56 ± 0.09	$\mathbf{0.52 \pm 0.10}$	$\mathbf{0.45 \pm 0.11}$	0.57 ± 0.06
$\ell_{2,1}$MKMTL-MRI	0.67 ± 0.09	0.51 ± 0.11	0.46 ± 0.09	0.41 ± 0.11	0.52 ± 0.09
$\ell_{2,1}$MKMTL-PET	0.63 ± 0.05	0.49 ± 0.10	0.41 ± 0.13	0.37 ± 0.09	0.47 ± 0.06
$\ell_{2,1}$MKMTL-MP	0.71 ± 0.06	0.53 ± 0.10	0.47 ± 0.11	0.42 ± 0.10	0.56 ± 0.08
$\ell_{2,1}$MKMTL-MPD	0.72 ± 0.06	0.55 ± 0.11	0.51 ± 0.09	0.44 ± 0.09	0.58 ± 0.06
$\ell_{2,1} - \ell_1$SMKMTL-MRI	0.67 ± 0.09	0.54 ± 0.12	0.49 ± 0.09	0.42 ± 0.13	0.52 ± 0.10
$\ell_{2,1} - \ell_1$SMKMTL-PET	0.63 ± 0.05	0.48 ± 0.10	0.41 ± 0.11	0.38 ± 0.09	0.52 ± 0.06
$\ell_{2,1} - \ell_1$SMKMTL-MP	0.71 ± 0.06	0.56 ± 0.10	0.49 ± 0.09	0.43 ± 0.12	0.58 ± 0.07
$\ell_{2,1} - \ell_1$SMKMTL-MPD	$\mathbf{0.72 \pm 0.06}$	$\mathbf{0.57 \pm 0.10}$	0.51 ± 0.09	0.44 ± 0.12	$\mathbf{0.58 \pm 0.07}$
M3T-MP [30]	0.69 ± 0.11	0.52 ± 0.12	0.49 ± 0.12	0.39 ± 0.17	0.55 ± 0.08
M3T-MPD [30]	0.71 ± 0.08	0.54 ± 0.13	0.51 ± 0.11	0.43 ± 0.10	0.56 ± 0.09
M2TFS-MP [28]	0.62 ± 0.15	0.49 ± 0.18	0.45 ± 0.09	0.41 ± 0.22	0.51 ± 0.13
I-M3T-MP [29]	0.62 ± 0.11	0.48 ± 0.18	0.41 ± 0.11	0.38 ± 0.32	0.50 ± 0.11
Method	RAVLT				wR
	TOTAL	TOT6	T30	RECOG	
$\ell_{2,1}$MTL-MRI	0.57 ± 0.07	0.53 ± 0.08	0.51 ± 0.04	0.44 ± 0.07	$0.52 \pm 0.08^*$
$\ell_{2,1}$MTL-PET	0.54 ± 0.10	0.49 ± 0.12	0.49 ± 0.09	0.40 ± 0.09	$0.48 \pm 0.08^*$
$\ell_{2,1}$MTL-MP	0.59 ± 0.07	0.54 ± 0.08	0.52 ± 0.03	0.45 ± 0.07	$0.54 \pm 0.07^*$
$\ell_{2,1}$MTL-MPD	0.61 ± 0.07	0.56 ± 0.07	0.54 ± 0.02	0.44 ± 0.08	$0.55 \pm 0.06^*$
$\ell_{2,1}$MKMTL-MRI	0.58 ± 0.06	0.53 ± 0.09	0.51 ± 0.04	0.43 ± 0.07	$0.51 \pm 0.07^*$
$\ell_{2,1}$MKMTL-PET	0.55 ± 0.11	0.50 ± 0.11	0.50 ± 0.08	0.43 ± 0.08	$0.48 \pm 0.07^*$
$\ell_{2,1}$MKMTL-MP	0.61 ± 0.08	0.56 ± 0.10	0.54 ± 0.05	0.46 ± 0.07	$0.54 \pm 0.07^*$
$\ell_{2,1}$MKMTL-MPD	0.65 ± 0.07	0.57 ± 0.08	0.56 ± 0.03	0.44 ± 0.08	$0.56 \pm 0.06^*$
$\ell_{2,1} - \ell_1$SMKMTL-MRI	0.59 ± 0.07	0.55 ± 0.08	0.53 ± 0.03	0.45 ± 0.07	$0.53 \pm 0.08^*$
$\ell_{2,1} - \ell_1$SMKMTL-PET	0.56 ± 0.10	0.51 ± 0.11	0.50 ± 0.08	0.43 ± 0.09	$0.49 \pm 0.07^*$
$\ell_{2,1} - \ell_1$SMKMTL-MP	0.62 ± 0.07	0.58 ± 0.08	0.56 ± 0.04	0.47 ± 0.07	$0.56 \pm 0.07^*$
$\ell_{2,1} - \ell_1$SMKMTL-MPD	$\mathbf{0.63 \pm 0.06}$	$\mathbf{0.59 \pm 0.07}$	$\mathbf{0.57 \pm 0.04}$	$\mathbf{0.49 \pm 0.08}$	$\mathbf{0.57 \pm 0.06}$
M3T-MP [30]	0.61 ± 0.08	0.52 ± 0.07	0.52 ± 0.09	0.41 ± 0.08	$0.51 \pm 0.08^*$
M3T-MPD [30]	0.62 ± 0.09	0.55 ± 0.11	0.54 ± 0.06	0.43 ± 0.11	$0.52 \pm 0.05^*$
M2TFS-MP [28]	0.58 ± 0.13	0.50 ± 0.15	0.51 ± 0.11	0.42 ± 0.08	$0.50 \pm 0.09^*$
I-M3T-MP [29]	0.55 ± 0.11	0.47 ± 0.12	0.52 ± 0.13	0.40 ± 0.09	$0.51 \pm 0.06^*$

3. We design an efficient optimization algorithm to solve this non-smooth formulation.

4. To highlight the advantages of our proposed MTL method, we thoroughly investigate and evaluate the proposed method to demonstrate our methods along various dimensions including prediction performance on the cognitive outcomes prediction, classification, biomarkers identification and multi-modal data fusion.

We conducted extensive experiments using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to demonstrate our methods along various dimensions including prediction performance on the baseline cognitive outcomes, biomarkers identification and multi-modal data fusion.

While there has been a steady growth in machine learning research for computer aided diagnosis or healthcare, there are some fundamental problems for medical predictive modeling:

(1) *How to fuse the multi-modality data of patients.* The multi-modality clinical data contains important information for the accurate and effective diagnosis of diseases. Thus, effectively fusing various modalities of each patient can provide supplementary clinical information which is not apparent in each individual modality alone.

(2) *How to leverage the correlation of multiple related clinical tasks.* The ability to learn an accurate model for predicting patient outcomes hinges on the amount of training data available. When training samples from a task are limited, it is not enough to learn an accurate model. Many clinical tasks for the same disease are relevant, it is important to collaboratively learn multiple clinically-related tasks by studying the inherent interactions and correlations among these tasks. Besides, a recent study formulated mortality prediction as multi-task learning in which a task corresponds to a disease [65]. Because patients are usually associated with multiple diseases, it would be necessary to consider the correlation of different diseases each patient is associated with. Moreover, Xu et al. developed a personalized model by leveraging the shared information among patients, to improve performance of each task [66].

(3) *How to identify the sensitive biomarkers from high dimensional features.* In the high dimensional clinical data, only a small number of variables are relevant to the disease. Hence, it is critical to intelligently and automatically extract the useful information from data.

To our best knowledge, our model is the first method to jointly fusing multi-modality data and multi-task learning. This approach exploits both the task relationship and the complementary nature of different modalities for an effective and strong predictive model. Next, we focus our discussion on the important components of the unified framework.

- *Multi-kernel learning framework*

Although the linearized method can be optimized efficiently, it cannot be directly applied to capture the high order statistics and the underlying structures of complex data spaces. The kernel based models enable us to capture nonlinear associations between MRI and cognitive outcomes. As the kernel plays an essential role in the formulation, inappropriate kernels may not accurately capture

the correlation structure of data. The multi-kernel learning framework aims to solve the kernel selection problem in a principled way kernel selection, therefore it can provide a solution to obtain an optimally combined kernel representation of candidate kernel functions. In the Experiment I, we compared with the single kernel based MTL methods ($\ell_{2,1}$KMTL and CORNLIN), $\ell_{2,1} - \ell_1$SMKMTL achieved a better performance, which indicates the importance of kernel learning. Moreover, the result in Experiment II also indicates that the kernel selection for kernelized methods is very crucial for multi-task learning. The inappropriate kernels usually result in sub-optimal or even poor performance. Furthermore, through learning common kernel functions across multiple tasks, it can facilitate to look for more discriminative and robust kernel functions. Furthermore, the multi-kernel learning provides a general framework for regression and classification. The promising classification performance obtained in Experiment V also indicates a generalization ability of our proposed framework.

- *Multi-task learning with kernel-wise and feature-wise correlation*

It is known that there exist inherent correlations among multiple clinical cognitive variables of a subject. The key of MTL is to identify how the tasks are related. In the Experiment I, we compared MTL with feature representation sharing ($\ell_{2,1}$MTL and CORNLIN), MTL with kernel parameters sharing ($\ell_{2,1}$MKMTL) and our MTL method with both feature and kernel sharing. The results demonstrate that neither MTL with feature representation sharing nor MTL with kernel parameters sharing can model well relatedness among the prediction tasks of the cognitive outcomes. By contrast, joint integration of feature-level and kernel-level analysis is more favorable by both feature representation and model parameter sharing.

- *Multi-modality fusion*

Several recent studies using multi-modality data in ADNI have demonstrated that it is important and beneficial to build prediction models by leveraging multi-modality data. Depending on the task, different fusion methods lead to different classification performances. In our Experiment VI, we conducted a comparison between feature-level fusion and kernel-level fusion. The multiple kernels can come from different sources of feature spaces. More specifically, the data of each modality is represented using a base kernel, and by the optimal weight parameters for multiple modality are optimized. In the proposed multi-modality learning formulation, both feature and modality level learning are performed in a unified multi-kernel multi-task learning framework. It beats the comparable methods proposed in [30] which perform feature-level and modality-level learning separately. Furthermore, as shown in Tables 7 and 8, our method is better than two other recent studies [28,29], even though they used the combination of MKL and MTL, which further shows the advantage of our proposed method due to considering the multi-modality data and multi-prediction tasks at the same time. Our new model is designed for multi-modality data fusion, which can also be naturally extended to deal with data with multi-view [67] or heterogeneous feature subset [68].

- *Effect of nonlinear biomarker selection*

The interpretability is highly desired in clinical diagnosis. The sparse representations enhance the interpretability of the model. Our proposed method is a sparse model that is able to identify a compact set of relevant neuroimaging biomarkers. Different from the linearized sparse models, $\ell_{2,1} - \ell_1$SMKMTL performs non-linear feature selection by associating a base kernel for each feature (ROI). Henceforth, it poses this problem of non-linear feature selection as that of optimizating the kernel parameters of the kernels. There has been much recent progress in non-linear feature selection where the predictor is a non-linear function of the input features. Since the complex relationship between the MRI features and cognitive outcomes, the traditional linear feature learning methods are not able to identify the underlying important features. However, no previous work investigate the AD-relevant imaging markers non-linear biomarker selection by a non-linear scheme. From Table 4 in the Experiment IV, we can observe that some ROIs identified by $\ell_{2,1} - \ell_1$SMKMTL, such as R.Fusiform [69], L.Precuneus, [70] and Paracentral [71], are consistent with results reported in previous studies on neuroimaging and cognition. However, they are not identified by the $\ell_{2,1}$MTL.

In summary, if the data contains multi-modalities or there exists some other clinical prediction tasks relevant to the current task, properly modeling the underlying correlation among the modalities and tasks could improve the prediction performance, especially for limited training samples from a single modality. MKL is a good choice for a predictive model for multi-modality fusion and multi-task learning due to its nonlinearity and generalization ability. Moreover, MKL can be used to perform nonlinear feature selection with the help of sparsity-inducing norm. Furthermore, although the training time of optimization is longer relative to the linear based methods. For the MKL methods, the time complexity of the gradient calculation is ignorable compared to the SVR solver. Therefore, the warm-start technique is used for successive SVR retrainings to accelerate the training process in our $\ell_{2,1} - \ell_1$SMKMTL method.

Although our proposed method demonstrates a good performance, some limitations should be considered in future studies.

(1) In our current work, the $\ell_{2,1}$-norm is not extended to a more general $\ell_{2,q}$ norm ($q \leq 1$) . The ℓ_q norm over the ROI can obtain different sparsities by varying and optimizing the value of q, which can achieve more interpretable results. As the next step, we will extend $\ell_{2,1}$-norm to $\ell_{2,q}$-norm, and design an efficient optimization algorithm to solve it.
(2) The $\ell_{2,1} - \ell_1$ regularized SMKMTL promotes sparse kernel combinations to support interpretability. However, the sparse MKL with ℓ_1-norm relies on the assumption that some kernels are irrelevant for predicting cognitive outcomes. Enforcing sparse combination may lead to unexpected models [72]. Various MKL formulations have been developed including ℓ_p-norm ($p \geq 1$) regularization over the kernel weights. To address the limitation, we extend it to $\ell_{2,1} - \ell_p$ SMKMTL by generalizing MKL to arbitrary ℓ_p-norm ($p \geq 1$) allowing for non-sparse solutions.
(3) When correlating the multiple prediction models, we assume that all the tasks shared the same feature subset in the input space or kernel induced feature space, which is not realistic sometimes. Different assessments evaluate the subjects' different cognitive functions, which results in different tasks preferring different brain regions, such as the tasks in TRAILS aimed to test a combination of visual, motor and executive functions, while the test of RAVLT aimed to test verbal learning memory. It is reasonable to assume that the correlation among the tasks are not equal, some tasks may be more closely related than others in the assessment tests of cognitive outcomes. We will extend our model to exploit more complex correlation structure inherent in the correlation among the tasks.
(4) Our method is a supervised model requiring a large amount of samples with target values or labels. However, in most applications, labeled data are expensive to collect but unlabeled data are abundant. In some MTL applications, the training set of each task consists of both labeled and unlabeled data, hence we hope to exploit useful information

contained in the unlabeled data to further improve the performance of MTL.

7. Conclusion

Many multi-task learning with sparsity-inducing regularization for modeling cognitive outcomes in AD have been proposed in the past decade, current formulations remain restricted to linear models and can't capture the relationship between the MRI features and cognitive outcomes. To address these shortcomings, we propose a multi-kernel multi-task learning with a joint sparsity-inducing regularization to model the more complicated but more flexible relationship between MRI features and cognitive outcomes. Our algorithm performs multi-task learning in the multiple kernel space and optimizes the combination of kernel function simultaneously for modeling the disease's cognitive scores. Is capable of uncovering the structure information shared by multiple tasks in the RKHS. Extensive experiments on ADNI dataset illustrate that proposed method not only yields superior performance on prediction performance of cognitive measures, but also is a powerful tool for discovering a small set of imaging biomarkers. The proposed approach is not restricted to cognitive performance prediction of subjects in AD, but also can be applied to other multi-task problem in many other applications. We will extend our methods to arbitrary ℓ_p-norm MKL with $p > 1$ allowing for non-sparse solutions.

Acknowledgment

This research was supported by the the National Natural Science Foundation of China (No. 61502091), the Fundamental Research Funds for the Central Universities (Nos. N161604001 and N150408001), the National Science Foundation for Distinguished Young Scholars of China under Grant (No. 71325002), Major International (Regional) Joint Research Project of NSFC under Grant (No. 71620107003) and the Foundation for Innovative Research Groups of National Science Foundation of China under Grant (No. 61621004).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2018.01.028.

References

[1] Alzheimer's Association, et al., 2016 Alzheimer'S disease facts and figures, Alzheimer's Dement. 12 (4) (2016) 459–509.
[2] N. Batsch, M. Mittelman, World Alzheimer Report 2012: Overcoming the Stigma of Dementia, 5, Alzheimer's Disease International (ADI), 2015.
[3] R.J. Castellani, R.K. Rolston, M.A. Smith, Alzheimer disease, Disease-a-month: DM 56 (9) (2010) 484–546.
[4] G.B. Frisoni, N.C. Fox, C.R. Jack, P. Scheltens, P.M. Thompson, The clinical use of structural MRI in alzheimer disease, Nat. Rev. Neurol. 6 (2) (2010) 67–77.
[5] 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 Trans. Med. Imaging 33 (7) (2014) 1475–1487.
[6] J. Wan, Z. Zhang, J. Yan, T. Li, B.D. Rao, S. Fang, S. Kim, S.L. Risacher, A.J. Saykin, L. Shen, Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 940–947.
[7] R. Caruana, Multitask learning, Mach. Learn. 28 (1) (1997) 41–75.
[8] A.F. Marquand, M. Brammer, S.C. Williams, O.M. Doyle, Bayesian multi-task learning for decoding multi-subject neuroimaging data, Neuroimage 92 (2014) 298–311.
[9] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, L. Shen, High--order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1277–1285.
[10] J. Xu, J. Zhou, P.-N. Tan, Formula: Factorized multi-task learning for task discovery in personalized medical models, in: Proceedings of the 15th SIAM International Conference on Data Mining, SIAM, 2015, pp. 496–504.
[11] J. Zhou, J. Liu, V.A. Narayan, J. Ye, Alzheimer's Disease Neuroimaging Initiative, et al., Modeling disease progression via multi-task learning, Neuroimage 78 (2013) 233–248.
[12] Z. Huo, D. Shen, H. Huang, New multi-task learning model to predict Alzheimer's disease cognitive assessment, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 317–325.
[13] J. Yan, T. Li, H. Wang, H. Huang, J. Wan, K. Nho, S. Kim, S.L. Risacher, A.J. Saykin, L. Shen, et al., Cortical surface biomarkers for predicting cognitive outcomes using group $\ell_{2,1}$ norm, Neurobiol. Aging 36 (2015) S185–S193.
[14] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A.D.N. Initiative, et al., Multimodal classification of Alzheimer's disease and mild cognitive impairment, Neuroimage 55 (3) (2011) 856–867.
[15] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A.J. Saykin, L. Shen, ADNI, Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance, in: Proceedings of the International Conference on Computer Vision, 2011, pp. 6–13.
[16] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, L. Shen, High--order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2012.
[17] A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, Mach. Learn. 73 (3) (2008) 243–272.
[18] B. Gu, V.S. Sheng, A robust regularization path algorithm for ν-support vector classification, IEEE Trans. Neural Netw. Learn. Syst. PP (2016) 1–8.
[19] B. Gu, V.S. Sheng, K.Y. Tay, W. Romano, S. Li, Incremental support vector learning for ordinal regression, IEEE Trans. Neural Netw. Learn. Syst. 26 (7) (2015) 1403–1416.
[20] F. Liu, L. Zhou, C. Shen, J. Yin, Multiple kernel learning in the primal for multimodal Alzheimer's disease classification, IEEE J. Biomed. Health Inf. 18 (3) (2014) 984–990.
[21] O.B. Ahmed, J. Benois-Pineau, M. Allard, G. Catheline, C.B. Amar, A.D.N. Initiative, et al., Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning, Neurocomputing 220 (2017) 98–110.
[22] T. Evgeniou, C.A. Micchelli, M. Pontil, Learning multiple tasks with kernel methods, J. Mach. Learn. Res. 6 (2005) 615–637.
[23] T. Kato, H. Kashima, M. Sugiyama, K. Asai, Multi-task learning via conic programming, in: Proceedings of the Advances in Neural Information Processing Systems, 2008, pp. 737–744.
[24] J.C. Caicedo, F.A. González, E. Romero, Content-based histopathology image retrieval using a kernel-based semantic annotation framework, J. Biomed. Inf. 44 (4) (2011) 519–528.
[25] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, J. Mach. Lear. Res. 12 (2011) 2211–2268.
[26] C. Widmer, M. Kloft, V.T. Sreedharan, G. Rätsch, Framework for multi-task multiple kernel learning and applications in genome analysis, preprint arXiv:1506.09153 (2015).
[27] A. Rakotomamonjy, R. Flamary, G. Gasso, S. Canu, Lp-lq penalty for sparse linear and sparse multiple kernel multitask learning., IEEE Trans. Neural Netw. 22 (8) (2011) 1307.
[28] B. Jie, D. Zhang, B. Cheng, D. Shen, Manifold regularized multitask feature learning for multimodality disease classification, Hum. Brain Mapp. 36 (2) (2015) 489–507.
[29] F. Liu, C.-Y. Wee, H. Chen, D. Shen, Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification, NeuroImage 84 (2014) 466–475.
[30] D. Zhang, D. Shen, A.D.N. Initiative, 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.
[31] K. Ito, B. Corrigan, Q. Zhao, J. French, R. Miller, H. Soares, E. Katz, T. Nicholas, B. Billing, R. Anziano, T. Fullerton, Alzheimer's Disease Neuroimaging Initiative, Disease progression model for cognitive deterioration from Alzheimer's disease neuroimaging initiative database, Alzheimer's Dement. 7 (2011) 151–160.
[32] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Optimization with sparsity-inducing penalties, Found. Trends® Mach. Learn. 4 (1) (2012) 1–106.
[33] J. Liu, J. Ye, Efficient euclidean projections in linear time, in: Proceedings of the International Conference on Machine Learning (ICML), 2009.
[34] R.K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, J. Mach. Learn. Res. 6 (2005) 1817–1853.
[35] S. Ben-David, R. Schuller, Exploiting Task Relatedness for Multiple Task Learning, in: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines. Lecture Notes in Computer Science, vol 2777, Springer, Berlin, Heidelberg, 2003, pp. 567–580.
[36] J. Baxter, et al., A model of inductive bias learning, J. Artif. Intell. Res. 12 (149–198) (2000) 3.
[37] T. Jebara, Multitask sparsity via maximum entropy discrimination, J. Mach. Learn. Res. 12 (2011) 75–110.
[38] T. Evgeniou, M. Pontil, Regularized multi–task learning, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 109–117.
[39] K. Yu, V. Tresp, A. Schwaighofer, Learning gaussian processes from multiple tasks, in: Proceedings of the 22nd International Conference on Machine Learning, ACM, 2005, pp. 1012–1019.
[40] Y. Xue, X. Liao, L. Carin, B. Krishnapuram, Multi-task learning for classification with dirichlet process priors, J. Mach. Learn. Res. 8 (2007) 35–63.

[41] P. Cao, X. Shan, D. Zhao, M. Huang, O. Zaiane, Sparse shared structure based multi-task learning for mri based cognitive performance prediction of Alzheimer's disease, Pattern Recognit. 72 (2017) 219–235.
[42] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in: Proceedings of the Advances in Neural Information Processing Systems, 2007, pp. 41–48.
[43] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization, in: Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 339–348.
[44] M. Jaggi, Convex optimization without projection steps, preprint arXiv:1108.1170 (2011).
[45] M. Jaggi, Revisiting frank-wolfe: projection-free sparse convex optimization., in: Proceedings of the ICML, 2013, pp. 427–435.
[46] E. Hazan, Sparse approximate solutions to semidefinite programs, Lect. Notes Comput. Sci. 4957 (2008) 306–316.
[47] J. Kuczyński, H. Woźniakowski, Estimating the largest eigenvalue by the power and lanczos algorithms with a random start, SIAM J. Matrix Anal. Appl. 13 (4) (1992) 1094–1122.
[48] A. Rakotomamonjy, F.R. Bach, S. Canu, Y. Grandvalet, Simplemkl, J. Mach. Learn. Res. 9 (2008) 2491–2521.
[49] 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.
[50] C.M. Stonnington, C. Chu, S. Klöppel, C.R. Jack, J. Ashburner, R.S. Frackowiak, Alzheimer Disease Neuroimaging Initiative, et al., Predicting clinical scores from magnetic resonance scans in Alzheimer's disease, Neuroimage 51 (4) (2010) 1405–1413.
[51] 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, ACM, 2011, pp. 42–50.
[52] J. Zhou, J. Chen, J. Ye, Clustered multi-task learning via alternating structure optimization, in: Proceedings of the Advances in Neural Information Processing Systems, 2011, pp. 702–710.
[53] S. Ji, J. Ye, An accelerated gradient method for trace norm minimization, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 457–464.
[54] J. Zhou, Multi-task learning in crisis event classification, Technical Report, http://www.public.asu.edu/jzhou29.
[55] P. Jawanpuria, M. Varma, S. Nath, On p-norm path following in multiple kernel learning for non-linear feature selection, in: Proceedings of the International Conference on Machine Learning, 2014, pp. 118–126.
[56] H. Suk, S.W. Lee, D. Shen, Subclass-based multi-task learning for Alzheimer's disease diagnosis, Front. Aging Neurosci. 6 (6) (2014) 168.
[57] X. Zhu, H.-I. Suk, S.-W. Lee, D. Shen, Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification, IEEE Trans. Biomed. Eng. 63 (3) (2016) 607–618.
[58] P. Cao, X. Liu, J. Yang, D. Zhao, M. Huang, J. Zhang, O. Zaiane, Nonlinearity-aware based dimensionality reduction and over-sampling for ad mci classification from mri measures, Comput. Biol. Med. 91 (2017).
[59] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J.V. Hajnal, D. Rueckert, A.D.N. Initiative, et al., Multiple instance learning for classification of dementia in brain mri, Med. Image Anal. 18 (5) (2014) 808–818.
[60] P. Coupé, S.F. Eskildsen, J.V. Manjón, V.S. Fonov, J.C. Pruessner, M. Allard, D.L. Collins, Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease, Neuroimage Clin. 1 (1) (2012) 141.
[61] R. Wolz, V. Julkunen, J. Koikkalainen, E. Niskanen, D.P. Zhang, D. Rueckert, H. Soininen, Multi-method analysis of mri images in early diagnostics of alzheimer's disease, PLoS ONE 6 (10) (2011) e25446.
[62] M. Zhang, Y. Yang, H. Zhang, F. Shen, D. Zhang, $L_{2,p}$-Norm and sample constraint based feature selection and classification for AD diagnosis, Neurocomputing 195 (2016) 104–111.
[63] L. Xu, X. Wu, K. Chen, L. Yao, Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment, Comput. Methods Progr. Biomed. 122 (2) (2015) 182–190.
[64] L. Xu, X. Wu, R. Li, K. Chen, Z. Long, J. Zhang, X. Guo, L. Yao, Prediction of progressive mild cognitive impairment by multi-modal neuroimaging biomarkers, J. Alzheimers Dis. 51 (4) (2016) 1045–1056.
[65] N. Nori, H. Kashima, K. Yamashita, H. Ikai, Y. Imanaka, Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 855–864.
[66] J. Xu, J. Zhou, P.-N. Tan, Formula: F act or ized mu lti-task l e a rning for task discovery in personalized medical models, in: Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM, 2015, pp. 496–504.
[67] H. Wang, F. Nie, H. Huang, Multi-view clustering and feature learning via structured sparsity, in: Proceedings of the International Conference on Machine Learning, 2013, pp. 352–360.
[68] P. Cao, X. Liu, J. Yang, D. Zhao, W. Li, M. Huang, O. Zaiane, A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules, Pattern Recognit. 64 (2017) 327–346.
[69] Y. Liu, T. Paajanen, Y. Zhang, E. Westman, L.O. Wahlund, A. Simmons, C. Tunnard, T. Sobow, P. Mecocci, M. Tsolaki, Combination analysis of neuropsychological tests and structural mri measures in differentiating ad, mci and control groups–the addneuromed study, Neurobiol. Aging 32 (7) (2011) 1198.
[70] E. Niskanen, M. Könönen, S. Määttä, M. Hallikainen, M. Kivipelto, S. Casarotto, M. Massimini, R. Vanninen, E. Mervaala, J. Karhu, New insights into alzheimer's disease progression: a combined tms and structural mri study, PLoS ONE 6 (10) (2011) e26113.
[71] 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, Am. J. Neuroradiol. 30 (5) (2009) 893–899.
[72] M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien, Lp-norm multiple kernel learning, J. Mach. Learn. Res. 12 (2011) 953–997.

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.

Xiaoli Liu is a Ph.D. candidate in Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, China. Her research interests include sparse learning, optimization and medical data mining.

Jinzhu Yang is a professor in the Department of Computer Science and Engineering, Northeastern University, China. His research interests include imaging processing and image segmentation.

Dazhe Zhao is a 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 interested in software engineering, and medical image computing.

Min Huang is a professor in the Department of Systems Engineering at Northeastern University. 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.