Computers in Biology and Medicine 140 (2022) 105090

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Dual feature correlation guided multi-task learning for Alzheimer's disease prediction

Shanshan Tang [a], Peng Cao [b,c,*], Min Huang [a,**], Xiaoli Liu [d], Osmar Zaiane [e]

[a] College of Information Science and Engineering, State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, Liaoning, 110819, China
[b] College of Computer Science and Engineering, Northeastern University, Shenyang, China
[c] Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang, China
[d] Department of Chemical and Biomolecular Engineering, Faculty of Engineering, National University of Singapore, Singapore
[e] Department of Computing Science, University of Alberta, Edmonton, Canada

ARTICLE INFO

Keywords:
Alzheimer's disease
Regression
Multi-task learning
Sparse learning
Feature correlation
Biomarker identification

ABSTRACT

Alzheimer's disease (AD) is a gradually progressive neurodegenerative disease affecting cognition functions. Predicting the cognitive scores from neuroimage measures and identifying relevant imaging biomarkers are important research topics in the study of AD. Despite machine learning algorithms having many successful applications, the prediction model suffers from the so-called curse of dimensionality. Multi-task feature learning (MTFL) has helped tackle this problem incorporating the correlations among multiple clinical cognitive scores. However, MTFL neglects the inherent correlation among brain imaging measures. In order to better predict the cognitive scores and identify stable biomarkers, we first propose a generalized multi-task formulation framework that incorporates the task and feature correlation structures simultaneously. Second, we present a novel feature-aware sparsity-inducing norm (FAS-norm) penalty to incorporate a useful correlation between brain regions by exploiting correlations among features. Three multi-task learning models that incorporate the FAS-norm penalty are proposed following our framework. Finally, the algorithm based on the alternating direction method of multipliers (ADMM) is developed to optimize the non-smooth problems. We comprehensively evaluate the proposed models on the cross-sectional and longitudinal Alzheimer's disease neuroimaging initiative datasets. The inputs are the thickness measures and the volume measures of the cortical regions of interest. Compared with MTFL, our methods achieve an average decrease of 4.28% in overall error in the cross-sectional analysis and an average decrease of 7.97% in the Alzheimer's Disease Assessment Scale cognitive total score longitudinal analysis. Moreover, our methods identify sensitive and stable biomarkers to physicians, such as the hippocampus, lateral ventricle, and corpus callosum.

1. Introduction

Alzheimer's disease (AD) is one of the most common progressive neurodegenerative diseases and the number of AD patients has been about 50 million according to World Health Organization (WHO) report (2019) [1]. Note that, China bears a heavy burden of AD costs, which greatly changes the estimates of AD costs worldwide [2]. Despite the fact that there is no cure for AD, early diagnosis of AD allows effective measures to prevent the disease from worsening [3].

Early diagnosis of AD usually starts with multiple cognitive tests [4]. The most commonly used cognitive tests include the Alzheimer's Disease Assessment Scale cognitive total score (ADAS) [5], the Mini-Mental State Exam score (MMSE) [6], and the Rey Auditory Verbal Learning Test (RAVLT) [7]. But researchers have found that the cognitive test results could be influenced by environmental factors, such as education, sociocultural biases of testing content, and the testing process. For example, African Americans scored significantly lower than White Non-Hispanics on the MMSE in an analysis that controlled for traditional demographics, including age, sex, and years of formal education, which suggests that differences in quality of education impact cognitive performance [8]. Fortunately, previous works suggest that brain atrophy may be present for years before the appearance of AD

* Corresponding author. College of Computer Science and Engineering, Northeastern University, Shenyang, China.
** Corresponding author.
E-mail addresses: caopeng@cse.neu.edu.cn (P. Cao), mhuang@mail.neu.edu.cn (M. Huang).

https://doi.org/10.1016/j.compbiomed.2021.105090
Received 27 May 2021; Received in revised form 22 November 2021; Accepted 26 November 2021
Available online 1 December 2021

symptoms [9,10]. Therefore the neuroimaging technology has been widely employed in the early diagnosis of AD because it can provide more sensitive and stable biomarkers [11]. In particular, magnetic resonance imaging (MRI) is one of the most informative techniques, which has been the first choice for the diagnosis of suspected AD [12].

This paper focuses on the problem of diagnosing AD by using MRI. The related works generally include two aspects: (1) the identification of relevant biomarkers and (2) predicting cognitive scores from MRI. For the first aspect, the identification of sensitive and stable biomarkers could facilitate disease diagnosis and prognosis [13]. Several studies apply statistical learning methods to selecting a set of neuroanatomic measures for AD diagnosis, such as principal component analysis and factor analysis [14–16]. For the second aspect, predicting cognitive scores can benefit screening and tracking the disease progression [17]. Some researchers performed linear regression models to predict clinical cognitive scores [17–19]. All these methods have a common characteristic: the biomarker identification and the clinical cognitive score prediction are separately performed, which will limit their prediction abilities.

Several studies have developed models that jointly realize the cognitive score prediction and the biomarker discovery. Sparse learning is one of the most popular techniques that are capable of simultaneously building predictive models from training data and performing biomarker identification via embedded feature selection [20]. It is well known that the ℓ_1-norm penalty leads to a sparse model, i.e., it can shrink many entries of the model to be exactly zero to achieve feature selection [21]. Sparse learning methods based on the ℓ_1-norm penalty have attracted a great number of research efforts due to their sparsity-inducing property, convenient convexity, and strong theoretical guarantees [20,22]. F. Bunea et al. discussed the most popular methods of predictor selection in regression models and presented that concurrently learning of cognitive score prediction and biomarker identification achieves a better performance than the individual component [23]. Despite the theoretical and empirical success, these models only predict clinical scores at a single time point or a single cognitive test, and their performances are far from satisfactory to be clinically useful for AD prognosis [24].

Multi-task learning (MTL) formulations are proposed to address the aforementioned challenges. In MTL, multiple tasks are learned simultaneously to improve the performance by utilizing task relatedness [25]. One appealing feature of the $\ell_{2,1}$-norm penalty is that it encourages multiple tasks to share similar sparsity patterns [26], and it has been commonly used in regression models and performs the joint feature selection on the multiple tasks [24,25,27,28]. Although multi-task learning methods that penalize the $\ell_{2,1}$-norm have achieved great success in many applications, they ignore the intrinsic useful correlation information among the features in a group structure. In AD research studies, the features that come from the brain regions can be divided into a set of non-overlapping groups [29]. Fig. 1 illustrates the intrinsic feature group structure. It can be observed that the brain is segmented into the regions of interest (ROIs), i.e., brain regions, according to the brain atlas (such as the Desikan-Killany), and the thickness measures and the volume measures of ROIs are used as the input features to predict the cognitive scores. Therefore, the thickness average (TA), thickness standard deviation (TS), surface area (SA), and cortical volume (CV) from the same cortical ROI could be seen as a group. Following this line, several studies have constructed multi-task learning models that group the relevant features together [30–32]. However, the previous studies only consider the intrinsic feature group structure, which is a restrictive assumption. The correlation between ROIs (feature groups) [33] has been ignored.

The ROIs correlations are denoted as the correlations between features in different brain regions in this work, and the Pearson correlation coefficient (PCC) is utilized to calculate the correlation coefficient for each pair of features. Then the correlation coefficients between ROIs are the sum of their feature correlation coefficients. Note that the correlation coefficients over 0.7 have been reserved to guarantee the sparsity. For example, X. Chen et al. only used the label correlation coefficients with cutoff 0.4 and the feature correlation coefficients with cutoff 0.6 [34]. Fig. 2 illustrates the correlations between ROIs by a chord diagram. It can be intuitively observed that there exist correlations between ROIs, which is called cross-regional feature correlation in this paper. Therefore, we consider the intrinsic group structure of the features as the explicit feature correlation and consider the cross-regional feature correlation as the implicit feature correlation. Inspired by the above analysis, we propose a novel regularization that incorporates the implicit feature correlation and construct multi-task learning models to predict clinical cognitive scores and identify biomarkers. The regularization penalty and multi-task learning models are presented in Section 4.

The main contributions of our study are summarized as follows:

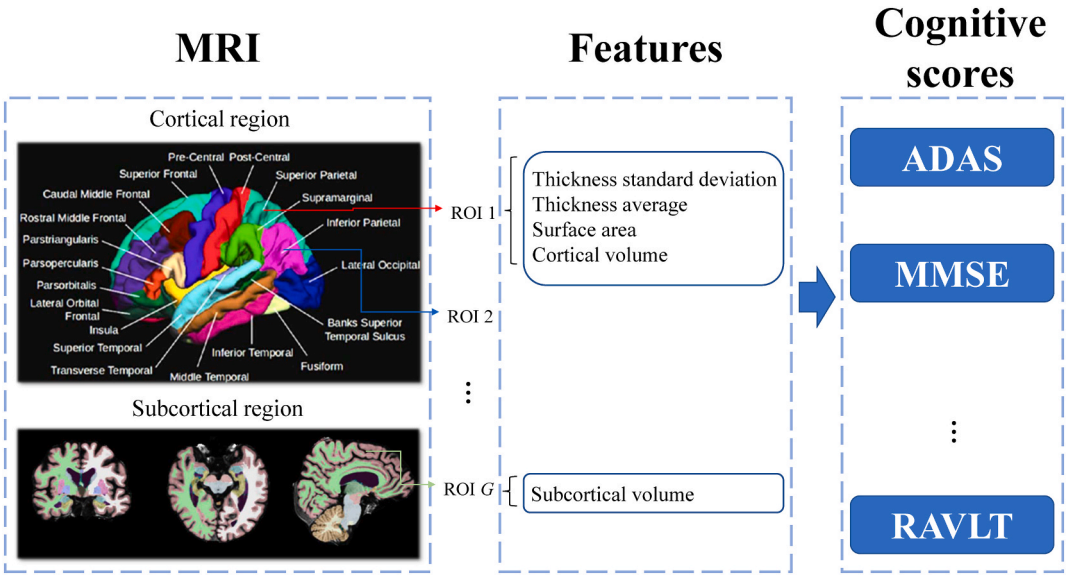


Fig. 1. The procedure of predicting cognitive scores using the features extracted from the brain MRI data. The brain is segmented into the regions of interest (ROIs) according to the brain atlas (such as the Desikan-Killany). Some ROIs (cortical regions) include four features: thickness average (TA), thickness standard deviation (TS), surface area (SA), and cortical volume (CV). Some ROIs (subcortical regions) include one feature: subcortical volume (SV).

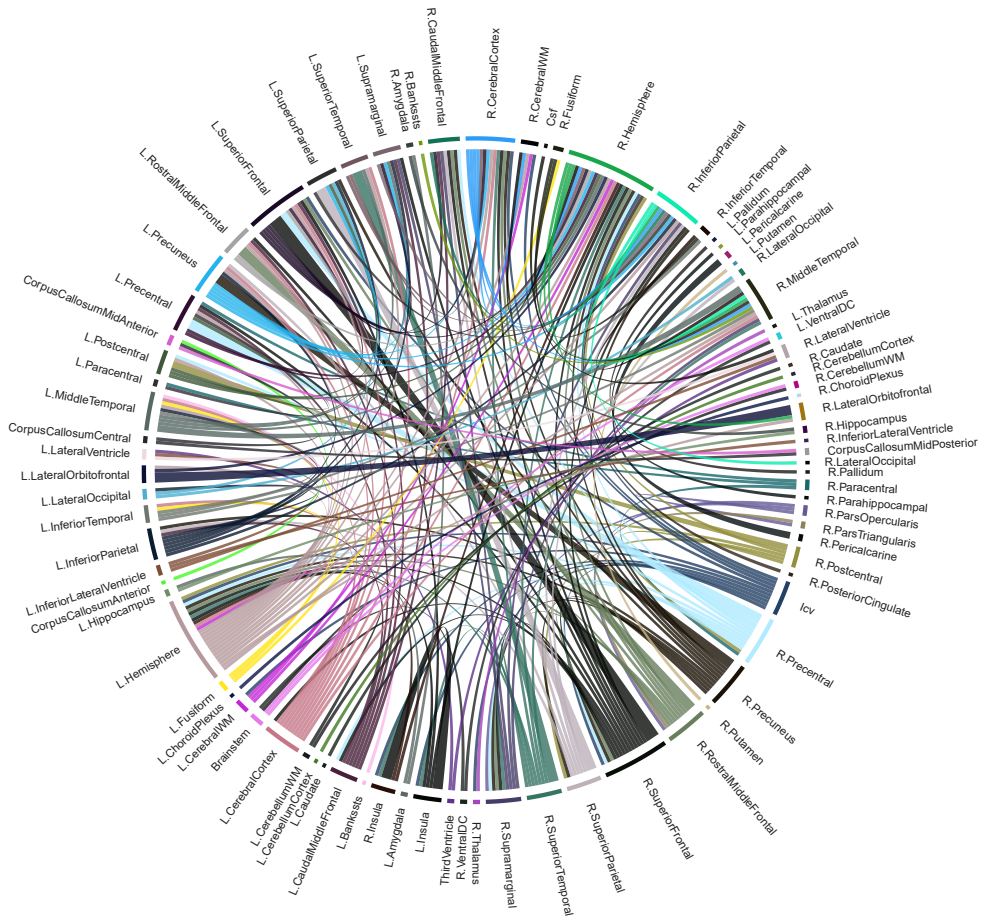


Fig. 2. The illustration of the correlations between ROIs. The arcs indicate the ROIs. The connections represent the correlations between ROIs, and the width of each connection indicates the correlation strength between the connected ROIs.

(1) **Formulation of a multi-task learning framework for disease prediction.** According to whether the regularization items incorporate the task or feature structure information, this paper summarizes the common regularization items proposed and applied in AD research studies. Then, we propose a generalized multi-task formulation framework simultaneously incorporating task correlation structure and feature correlation structure to improve regression performance and help identify important biomarkers.

(2) **Designing a new regularization.** We propose a feature-aware sparsity-inducing norm (FAS-norm) penalty, which incorporates a useful correlation between brain regions by exploiting correlations among features. Then, the FAS-norm penalty is extended to three multi-task learning models.

(3) **Development of efficient optimization algorithm.** We develop an optimization algorithm based on the alternating direction method of multipliers (ADMM) to solve the non-smooth convex problem caused by our proposed regularization penalty. The optimization method could be extended to other problems incorporating the FAS-norm penalty.

(4) **Comprehensive experiments to validate the effectiveness of the proposed models.** We first carry out simulations to assess the effectiveness of our methods in the scenario where tasks have a different number of features. Then, we evaluate the proposed methods using both cross-sectional and longitudinal Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets. For comparison, we implement a broad range of other algorithms. The experimental results on the ADNI datasets demonstrate that the proposed models consistently outperform competing methods and could identify stable biomarkers. We also conduct experiments on the dataset from the UCI data archive to evaluate the effectiveness of the proposed generalized multi-task formulation framework applied to the other conventional structural data.

The rest of this paper is organized as follows. Section 2 reviews the related work for jointly predicting the cognitive scores and selecting biomarkers based on regularized multi-task learning. Section 3 presents the preliminaries including the summaries of the regularization with different prior knowledge. Section 4 introduces our proposed generalized multi-task learning formulation framework, the FAS-norm penalty, and its three extended models, as well as the optimization algorithm based on the ADMM. Section 5 presents the experimental results and analysis on both synthetic and real datasets. Section 6 discusses the effectiveness of the proposed methods on another multi-view dataset, clinic score prediction, and biomarker identification. Section 7 concludes the paper.

2. Related work

Regarding the prediction of cognitive scores and identification of relevant biomarkers for AD study, a number of techniques have been presented using regularized multi-task learning approaches to realize them jointly, and this is also the focus of this paper. Thus, we only review the existing regularized multi-task learning methods for AD study.

The idea of regularized multi-task learning is to utilize the intrinsic relationships among multiple related tasks in order to improve the prediction performance, i.e., properly introducing the prior tasks correlation structure knowledge into regularization could help improve the

prediction accuracy. One of the key issues in regularized multi-task learning is to build learning models to capture such prior tasks correlation structure knowledge. There are two types of analysis that have been commonly studied in the literature about AD study to incorporate tasks correlation structure: cross-sectional analysis and longitudinal analysis. In the cross-sectional analysis, the prediction of different types of cognitive scores can be modeled as an MTL formulation, and the $\ell_{2,1}$-norm is the most commonly used regularization to incorporate the prior tasks correlation structure knowledge [27,28,35]. Specifically, H. Wang et al. employed an ℓ_1-norm penalty to impose sparsity among all elements and proposed the use of a combined $\ell_{2,1}$-norm and ℓ_1-norm penalties to select features [28]. D. Zhang et al. proposed a multi-task learning with $\ell_{2,1}$-norm to select a common subset of relevant features for multiple variables from each modality [27]. In the longitudinal analysis, researches formulate the prediction of clinical scores at a sequence of time points as a multi-task problem, where each task concerns the prediction of a clinical score at one time point. A great amount of work is devoted to capture the intrinsic relationship among tasks at different time points [25,36–38]. For example, J. Zhou et al. assumed that the multiple regression models from different time points satisfy the smoothness property, and proposed a temporal smoothness term [24,25]. B. Jie et al. incorporated two smoothness regularization terms into the objective function, fused smoothness term that penalties the differences between two successive weight vectors and output smoothness term that penalties the differences between outputs of two successive models [37]. H. Wang et al. imposed the low rank regularization denoted as the trace norm to exploit task correlations among the learning tasks at different time points [39]. M. Wang et al. utilized a relationship induced regularization to automatically capture the intrinsic relationship among tasks at different time points for estimating clinical scores based on longitudinal imaging data [40].

Although the aforementioned studies have performed outstanding results, they only consider the task correlation structure, ignoring the interrelated structures within neuroimaging measures, and thus may have limited power to generate optimal solutions [30]. In order to address this issue, a number of researches have been presented to take into account the feature group correlation structure [31,32,41,42]. For example, J. Wan exploited not only inter-vector correlation among regression coefficient vectors but also an intra-block correlation in each regression coefficient vector [41]. J. Yan et al. proposed a group-level $\ell_{2,1}$-norm strategy to group relevant features together in an anatomically meaningful manner and used this prior knowledge to guide the learning process, and the proposed model is called Group-sparse Multitask Regression and Feature Selection (G-SMuRFS) [30]. G-SMuRFS allows learning a common subset of feature groups across all the tasks simultaneously. This assumption is too restrictive since different tasks may prefer different feature groups. In order to solve this limitation, X. Liu et al. proposed a multi-task sparse group lasso (MT-SGL) method which encourages individual feature group selection with sparsity-inducing norm [32]. Despite the above achievements, studies have demonstrated that AD is closely related to the structure change of the connectivity among different brain regions, and the connectivity patterns will provide useful prior knowledge to guide the learning process [43].

In order to solve the issues above, we propose a generalized multi-task formulation framework that simultaneously incorporates the task and feature correlation structures. We also present a novel regularization penalty to incorporate a useful correlation between brain regions by exploiting correlations among the cross-regional features.

3. Preliminaries

A lowercase character denotes scalar (e.g., a). An uppercase character denotes matrix (e.g., A). A bold lowercase character denotes vector (e.g., $\boldsymbol{a}$). The i-th entry of $\boldsymbol{a}$ is denoted as a_i, the i-th row of A as $\boldsymbol{a}^i$ or $A_{i,\cdot}$, the j-th column of A as $\boldsymbol{a}_j$ or $A_{\cdot,j}$, the transpose of A as A^T, the trace of A as $tr(A)$ if A is a square matrix, and the inverse of A as A^{-1}.

3.1. Problem setup

In machine learning methods, the regression relationship between $\boldsymbol{y}$ and X is usually denoted as $\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{\xi}$. $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ denotes the training data where n and p are the number of the training instances and the dimensionality respectively. $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}} \in \mathbb{R}^{n\times 1}$ denotes the labels and $\boldsymbol{w} \in \mathbb{R}^{p\times 1}$ denotes the parameter vector of the model. $\boldsymbol{\xi} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ denotes the prediction error. The regression problem can be constructed as estimating the parameters based on a suitable regularized loss function:

$$\min_{\boldsymbol{w}} \quad \mathcal{L}(\boldsymbol{y}, X, \boldsymbol{w}) + \lambda R(\boldsymbol{w}) \ , \tag{1}$$

where the loss function can be $\mathcal{L}(\boldsymbol{y}, X, \boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{w}\|_2^2$. The regularization term $R(\cdot)$ penalizes the complexity of a learning model and alleviates overfitting by adding prior structural knowledge to it. $\lambda > 0$ is a regularization parameter controlling the tradeoff between the loss and the penalty.

MTL setting with *t tasks is considered in this paper. The input of the* j-th task can be denoted as $X_j = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_j}]^{\mathrm{T}} \in \mathbb{R}^{n_j\times p_j}$, where $j = 1, 2, \ldots, t$. The output of the j-th task can be denoted as $\boldsymbol{y}_j \in \mathbb{R}^{n_j\times 1}$. The model of the j-th task is denoted as $\boldsymbol{w}_j \in \mathbb{R}^{p_j\times 1}$. Then, the regularized least square loss function for the MTL can be formulated as:

$$\min_{W} \quad \sum_{j=1}^{t} \frac{1}{2}\|\boldsymbol{y}_j - X_j\boldsymbol{w}_j\|_2^2 + \lambda R(W) \ , \tag{2}$$

where the j-th column of W is $\boldsymbol{w}_j$. Note that the related studies usually assume that all tasks have the same number of features, i.e., $p_1 = p_2 = \ldots = p_t = p$. Therefore $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t] \in \mathbb{R}^{p\times t}$.

3.2. The cross-sectional and longitudinal analysis

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. ADNI database is collected through regular hospital visits of patients after their first screening (baseline). Two families of multi-task learning problems are studied on the ADNI datasets in this work: the cross-sectional analysis and the longitudinal analysis. It is assumed that the input of each task is identical in the remainder of this section, which means $X_j = X \in \mathbb{R}^{n\times p}$, $j = 1, 2, \ldots, t$. Consequently, the least square loss function can be denoted as $\mathcal{L}(Y, X, W) = \frac{1}{2}\|Y - XW\|_F^2$.

The cross-sectional analysis is shown in Fig. 3 (*a*). $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ denotes the thickness measures and the volume measures of ROIs at patients' first screening. $Y = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times t}$ denotes the scores of patients corresponding to different cognitive tests. The $\boldsymbol{w}_j$ in $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t] \in \mathbb{R}^{p\times t}$ denotes the model parameter vector of the j-th cognitive score. In the cross-sectional analysis, the models are constructed for multiple cognitive scores at a time point. It is assumed that the cognitive scores are correlated with each other.

The longitudinal analysis is shown in Fig. 3 (*b*), where $\boldsymbol{w}_j$ denotes the model parameter vector of one cognitive score at the j-th time point. The first screening of the patient is called the baseline, and the time point for the follow-up visits is denoted by the duration starting from the baseline [25]. Therefore, time points are denoted as baseline (time = 0), six months later (time = 6), 12 months later (time = 12), 24 months later (time = 24), and 36 months later (time = 36). $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ is still the thickness measures and the volume measures of ROIs at the baseline. $Y = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times t}$ is the scores of patients corresponding to a cognitive test at different time points. In the longitudinal analysis, we predict future scores of the specific cognitive test using baseline MRI data. Note that the longitudinal models are applied independently on each cognitive score, and we do not assume that the cognitive scores are correlated.

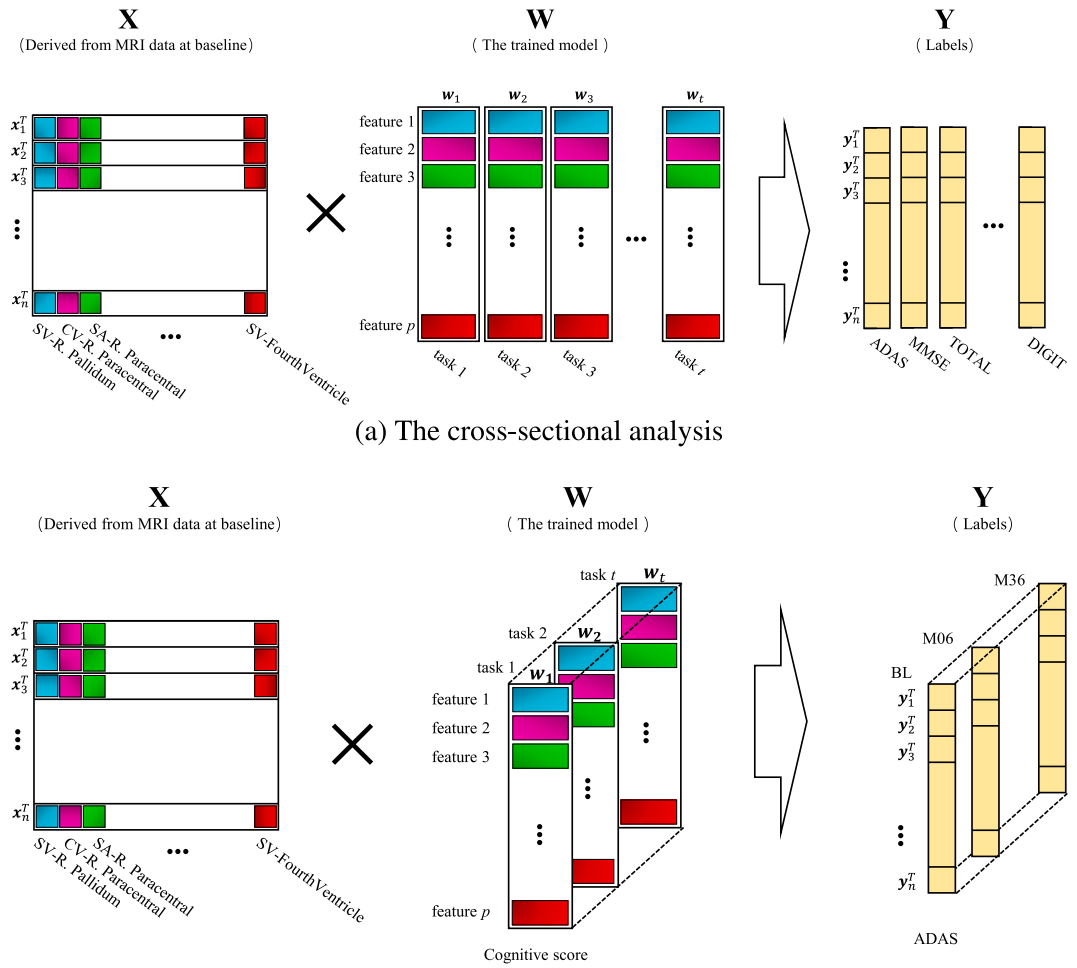
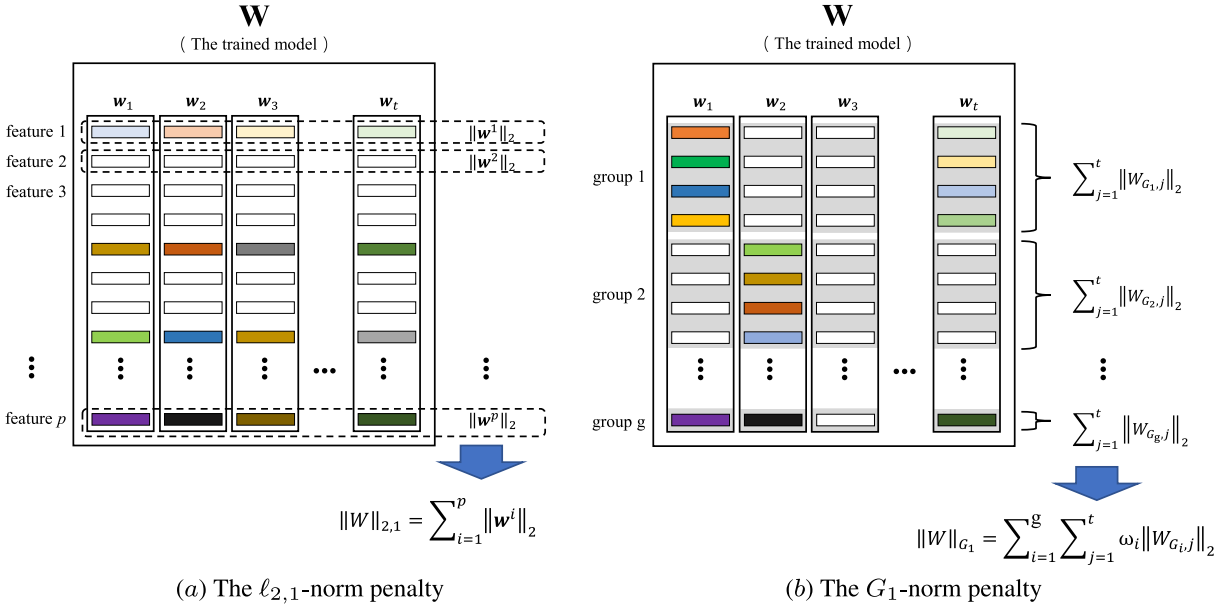


Fig. 3. (*a*) The cross-sectional analysis: to predict multiple cognitive scores at the baseline. (*b*) The longitudinal analysis: to predict scores of the specific cognitive test at multiple time points.



(*a*) The $\ell_{2,1}$-norm penalty (*b*) The G_1-norm penalty

Fig. 4. An illustration of the $\ell_{2,1}$-norm penalty and the G_1-norm penalty. The non-zero weights are colored. In subfigure (*a*), the weights framed by each dashed line are shrunk together. In subfigure (*b*), the weights in each gray background are shrunk together.

3.3. Regularization with different prior structural knowledge

In this section, we summary the regularization terms that are commonly applied in AD research studies from the perspective that the regularization could incorporate different prior structural knowledge. The $\ell_{2,1}$-norm penalty is one of the most common regularization terms for MTL, which encourages multiple tasks to share similar sparsity patterns [26]. It is formulated as

$$\|W\|_{2,1} = \sum_{i=1}^{p} \|\boldsymbol{w}^i\|_2 \ , \tag{3}$$

where $\boldsymbol{w}^i$ is the weights of one feature over all tasks, and the illustration of the $\ell_{2,1}$-norm penalty is in Fig. 4 (*a*).

Although the $\ell_{2,1}$-norm penalty has performed outstanding results, it does not consider the feature correlation, which is inconsistent with the reality [33]. Previous studies assume the p covariates to be divided into g disjoint groups G_i, $i = 1, \ldots, g$, with each group having v_i covariates respectively [46]. In the context of AD research studies, each group corresponds to an ROI in the brain, and the covariates in each group correspond to specific features of that region. Therefore, the number of features in each group, v_i, ranges from 1 to 4, and the number of groups g can be in the hundreds. Then the G_1-norm penalty is introduced according to the relationship between the brain regions (ROIs) to encourage a task-specific subset of ROIs [42]. The formulation of the G_1-norm penalty is given as

$$\|W\|_{G_1} = \sum_{i=1}^{g} \sum_{j=1}^{t} \omega_i \|W_{G_i j}\|_2 \ , \tag{4}$$

which is shown in Fig. 4 (*b*). ω_i denotes the weight of the i-th feature group, where $\omega_i = \sqrt{v_i}$. $W_{G_i j}$ denotes the weights of features that consist in the i-th feature group for the j-th task. Brand et al. proposed a method applying the group-level ℓ_1-norm penalty (the G_1-norm [47]) to capture the relationships that are intrinsic in the input modalities [42].

Following the line that whether the regularization items incorporate the task or feature structure knowledge, this paper divides them into four categories: no structural knowledge, task structure, feature structure, and both task and feature structure. Specifically, the temporal penalty and the fused lasso penalty are employed to incorporate the temporal smoothness within tasks for the longitudinal analysis [25]. The exclusive lasso penalty assumes a competitive nature among the features shared by all the tasks. That is to say, if a feature was assigned a very large weight in one task, the weights of this feature in other tasks were expected to be small or even zero [40]. In the relationship induced term, Ω denotes the task covariance matrix that will benefit learning the models by inducing the correct relationship among tasks [40]. The robust regularization term is aimed at identifying irrelevant (outlier) tasks when learning from multiple tasks [45]. Note that, the $G_{2,1}$-norm penalty is different from the G_1-norm penalty because the $G_{2,1}$-norm penalty introduces the group feature structure across all tasks whereas the G_1-norm penalty introduces that for each task. That is, the $G_{2,1}$-norm penalty aims to select task-shared features but the G_1-norm penalty aims to select task-specific features [30,42]. The details of the regularization terms that are often applied in the AD research studies are listed in Table 1.

Table 1
Regularization with different prior structural knowledge.

Prior structural knowledge	Name	Formulation	Cite
No structural knowledge	ℓ_2-norm	$\|W\|_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{t} W_{ij}^2$	[24]
	ℓ_1-norm	$\|W\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{t} \|W_{ij}\|$	[44]
Task structure	$\ell_{2,1}$-norm	$\|W\|_{2,1} = \sum_{i=1}^{p} \|\boldsymbol{w}^i\|_2$	[44]
	Temporal penalty	$\|WH\|_F^2 = \sum_{j=1}^{t-1} \|\boldsymbol{w}_j - \boldsymbol{w}_{j+1}\|_2^2$	[25]
	Fused lasso	$\|WH\|_1 = \sum_{j=1}^{t-1} \|\boldsymbol{w}_j - \boldsymbol{w}_{j+1}\|_1$	[25]
	Exclusive lasso	$\sum_{i=1}^{p} \|\boldsymbol{w}^i\|_1^2$	[40]
	Relationship induced term	$tr(W\Omega^{-1}W^T)$ *s.t.* $\Omega \succcurlyeq 0$, $tr(\Omega) = 1$	[40]
	Robust	$\|W^T\|_{2,1} = \sum_{j=1}^{t} \|\boldsymbol{w}_j\|_2$	[45]
Feature structure	G_1-norm	$\|W\|_{G_1} = \sum_{i=1}^{g}\sum_{j=1}^{t} \omega_i \|W_{G_i j}\|_2$	[42]
Task and feature structure	$G_{2,1}$-norm	$\|W\|_{G_{2,1}} = \sum_{i=1}^{g} \omega_i \|W_{G_i,\cdot}\|_F$	[30]

4. Methodology

4.1. The generalized multi-task formulation

Regularization is a technique to prevent the model from overfitting by adding prior structural knowledge to it, and incorporating proper prior knowledge into it will benefit learning the models. Although the regularized multi-task learning methods have performed outstanding results, multiple prediction models only consider the prior task correlation structure knowledge, ignoring the interrelated structures within features, and thus may have limited power to generate optimal solutions. To address this limitation, we design a general multi-task learning formulation. Mathematically, we minimize the following joint objective:

$$\min_{W} \quad \mathcal{L}(Y, X, W) + \lambda_t R_t(W) + \lambda_f R_f(W) \ , \tag{5}$$

where $R_t(\cdot)$ and $R_f(\cdot)$ incorporate the task correlation structure knowledge and the feature correlation structure knowledge respectively. We assume that adding the feature correlation structure information can help to improve regression performance and identify important biomarkers. The least square loss function is studied in this work. The j-th column of W is $\boldsymbol{w}_j$. X and Y denote input and output data respectively, which are described in Section 3.2.

4.2. Feature-aware sparsity-inducing regularization

The G_1-norm penalty only incorporates the intrinsic group structure of the features (the explicit feature correlation), and the cross-regional feature correlation (the implicit feature correlation) is neglected. This may cause bias in biomarker identification. According to the analysis of the cross-regional feature correlation in Fig. 2, it can be seen that the correlation among different regions is existent. We use the correlation between features in different regions to denote the region correlations. It is reasonable to assume that the difference of the importance between two strong correlated features is small.

Firstly, we construct a feature correlation matrix $C \in \mathbb{R}^{p\times p}$, and the off-diagonal entries donate feature correlation coefficients. $c_{m,l}$ denotes the correlation coefficient of the m-th feature and the l-th feature, which is calculated as:

$$c_{m,l} = \frac{cov(X_{\cdot,m}, X_{\cdot,l})}{\sigma_{X_{\cdot,m}}\sigma_{X_{\cdot,l}}} = \frac{\mathbb{E}\left[\left(X_{\cdot,m} - \mu_{X_{\cdot,m}}\right)\left(X_{\cdot,l} - \mu_{X_{\cdot,l}}\right)\right]}{\sigma_{X_{\cdot,m}}\sigma_{X_{\cdot,l}}} \ , \tag{6}$$

where $m, l = 1, \ldots, p$, $c_{m,l} \in [-1, 1]$. The $X_{\cdot,m}$ denotes the m-th feature values of all samples. cov is the covariance, and $\sigma_{X_{\cdot,m}}$ is the standard deviation of $X_{\cdot,m}$. $\mathbb{E}$ is the expectation. $\mu_{X_{\cdot,m}}$ is the mean of $X_{\cdot,m}$. $c_{m,l} > 0$ denotes positive correlation between features and $c_{m,l} < 0$ denotes negative correlation between features. The bigger $|c_{m,l}|$ denotes the stronger correlation between feature m and l.

Next, the distribution of the feature correlation coefficient has been shown in Fig. 5. It can be observed that the strength of the feature correlation is unstable. To build a more stable correlation matrix, a threshold technique is applied to connect only highly correlated features. Two features are considered highly correlated if the absolute value of their correlation coefficient is above a given threshold τ. Therefore, the useful connectivity among different brain regions can be identified

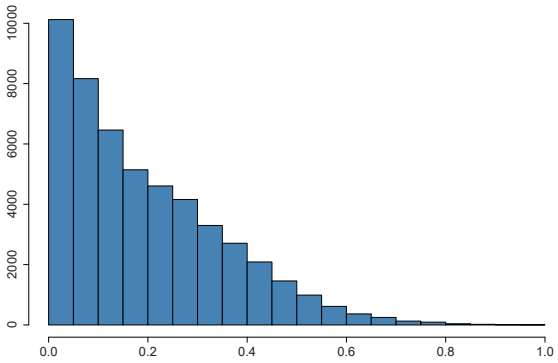


Fig. 5. The distribution of the feature correlation coefficient. (The horizontal axis is the absolute value of the correlation coefficient. The vertical axis is the amount of the feature pairs corresponding to the correlation coefficient.)

from the estimated feature correlation matrix.

The cross-regional feature correlation can be developed as a type of graph regularization. It can be constructed as an undirected graph $\mathbb{G} = (V,E)$, where the vertexes in V denote features and the undirected edges in E connect feature pairs whose correlation coefficients are reserved. $e(m, l) \in E$ corresponds to an edge between the m-th feature and the l-th feature. Finally, the matrix C is normalized into S:

$$s_{m,l} = \begin{cases} -\dfrac{c_{m,l}}{k} & (m,l) \in E, m \neq l \\ \dfrac{\sum_{m=1,m\neq l}^{p} |r_{m,l}|}{k} & (m,l) \in E, m = l \\ 0 & otherwise\ , \end{cases} \tag{7}$$

where $k = |E|$ is the number of edges in E.

Inspired by the above analysis, to capture the implicit feature correlation across different ROIs, an implicit feature correlation regularization, the feature-aware sparsity-inducing norm (FAS-norm) penalty, is designed. The FAS-norm penalty penalizes large deviations between features of high correlation, resulting in the following formulation:

$$\|SW\|_1 = \sum_{e(m,l)\in E} |s_{m,l}| \|\boldsymbol{w}^m - sign(s_{m,l})\boldsymbol{w}^l\|_1\ , \tag{8}$$

where $\boldsymbol{w}^i$ is the i-th row of the matrix W, and the weight $|s_{m,l}|$ is considered (See Fig. 6). The FAS-norm penalty encourages the correlated features to take similar values by shrinking the difference between them toward zero.

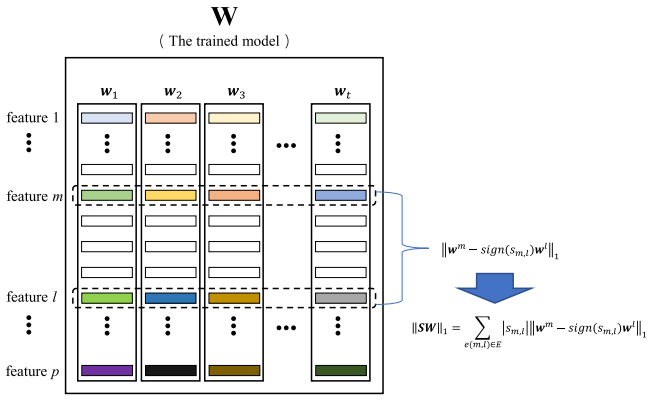


Fig. 6. An illustration of the feature-aware sparsity-inducing norm (FAS-norm) penalty. There are p features and t tasks. The non-zero weights are colored.

4.3. Feature-aware sparse multi-task feature learning

Bias can arise in the biomarker identification of the conventional MTFL model because it only incorporates the intrinsic correlation among tasks without considering the important correlation among features. It is thus desired to develop a model which incorporates task and feature correlations simultaneously. To this end, a feature-aware sparse multi-task feature learning model (FAS-MTFL) is proposed, which incorporates the FAS-norm penalty into the conventional MTFL model and helps to identify more stable biomarkers. In particular, the following multi-task learning model is considered:

$$\min_W \ \frac{1}{2}\|Y - XW\|_F^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|SW\|_1\ . \tag{9}$$

The objective contains two regularization penalties: (1) the $\ell_{2,1}$-norm penalty allows joint feature selection for all tasks, and (2) the FAS-norm penalty is enforced on the features, which incorporates the implicit feature correlation.

4.4. Dual feature correlation guided multi-task feature learning for the cross-sectional analysis

In the FAS-MTFL model above, the implicit feature correlation is incorporated by the FAS-norm penalty but it has not considered the explicit feature correlation. Recall that the explicit feature correlation and the implicit feature correlation exist simultaneously in ROIs. It is reasonable to incorporate complete feature structure information, thus the more accurate and stable biomarkers for AD can be identified. To this end, a dual feature correlation guided multi-task feature learning model (dMTLc) for the cross-sectional analysis is developed, which incorporates the G_1-norm penalty to model the explicit feature correlation:

$$\min_W \ \frac{1}{2}\|Y - XW\|_F^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|SW\|_1 + \lambda_3 \|W\|_{G_1}\ . \tag{10}$$

The G_1-norm penalty assumes that the features coming from a brain region can be divided into a group and selects task-specific feature groups.

4.5. Dual feature correlation guided multi-task fused learning for the longitudinal analysis

The models above incorporate the relatedness among different tasks by utilizing the $\ell_{2,1}$-norm penalty. The assumption for it is that these tasks are equally related to each other, which is improper for the case in the longitudinal analysis. In the course of disease progression, it is reasonable to assume that the difference of the cognitive scores between two successive time points is relatively small [25]. Therefore, we incorporate the fused lasso penalty to formulate a dual feature correlation guided multi-task fused learning model (dMTLl) for longitudinal analysis. The fused lasso penalty induces the temporal smoothness within tasks, and the dMTLl model is defined as:

$$\min_W \ \frac{1}{2}\|Y - XW\|_F^2 + \lambda_1 \|WH\|_1^2 + \lambda_2 \|SW\|_1 + \lambda_3 \|W\|_{G_1}\ , \tag{11}$$

where $\|WH\|_1^2$ is the fused lasso penalty, and $H \in \mathbb{R}^{t\times(t-1)}$, $H_{i,j} = 1$ if $i = j$, $H_{i,j} = -1$ if $i = j + 1$, $i, j = 1, 2, \ldots, t$.

4.6. Optimization

The formulation is challenging to optimize due to the use of non-smooth penalties. The ADMM algorithm simplifies the process of the non-smooth penalty by introducing a new variable, then uses the corresponding operation to solve the minimization problem. Motivated by the success of applying the ADMM algorithm to parallelizing distributed convex problems, we propose an optimization method based on that to solve Eq. (9).

By introducing slack variables $Q = W$ and $P = SW$, Eq. (9) can be rewritten in ADMM form as Eq. (12).

$$\begin{aligned} \min_{W,Q,P} \quad & \frac{1}{2}\|Y - XW\|_F^2 + \lambda_1\|Q\|_{2,1} + \lambda_2\|P\|_1 \\ \text{s.t.} \quad & W - Q = 0,\ SW - P = 0\ . \end{aligned} \tag{12}$$

The augmented Lagrangian of Eq. (12) is:

$$\begin{aligned} \mathcal{L}_\rho(W,Q,P,U_1,U_2) = \ & \frac{1}{2}\|Y - XW\|_F^2 + \lambda_1\|Q\|_{2,1} + \lambda_2\|P\|_1 \\ + < U_1, W - Q > + & \frac{\rho}{2}\|W - Q\|^2 + < U_2, SW - P > + \frac{\rho}{2}\|SW - P\|^2\ , \end{aligned} \tag{13}$$

where U_1 and U_2 are augmented lagrangian multipliers. Thus, the scaled form in Eq. (13) can be solved by the following problems.

Update *W*: The update function of W at the $(t+1)$-th iteration is carried out by Eq. (14),

$$\begin{aligned} W^{(t+1)} = \ & \arg\min_W \frac{1}{2}\|Y - XW\|_F^2 + < U_1^{(t)}, W - Q^{(t)} > + \frac{\rho}{2}\|W - Q^{(t)}\|^2 \\ & + < U_2^{(t)}, SW - P^{(t)} > + \frac{\rho}{2}\|SW - P^{(t)}\|^2\ . \end{aligned} \tag{14}$$

Note that Eq. (14) is the closed form and a closed solution can be obtained by setting its derivative to zero. Eq. (15) is derived:

$$0 = -X^{\mathrm{T}}(Y - XW) + U_1^{(t)} + \rho\left(W - Q^{(t)}\right) + SU_2^{(t)} + \rho S\left(SW - P^{(t)}\right)\ . \tag{15}$$

Therefore the $W^{(t+1)}$ can be updated efficiently employing Cholesky factorization. The optimal solution is given by $W^{(t+1)} = F^{-1}B^{(t)}$, where

$$F = X^{\mathrm{T}}X + \rho I + \rho SS\ . \tag{16}$$

$$B^{(t)} = X^{\mathrm{T}}Y - U_1^{(t)} + \rho Q^{(t)} - SU_2^{(t)} + \rho SP^{(t)}\ . \tag{17}$$

Update *Q*: According to Eq. (13), the update of Q can be solved as the following:

$$Q^{(t+1)} = \arg\min_Q \lambda_1\|Q\|_{2,1} + < U_1^{(t)}, W^{(t+1)} - Q > + \frac{\rho}{2}\|W^{(t+1)} - Q\|^2\ , \tag{18}$$

which is equivalent to the following problem:

$$Q^{(t+1)} = \arg\min_Q \frac{1}{2}\|Q - \Lambda_1^{(t+1)}\|^2 + \frac{\lambda_1}{\rho}\|Q\|_{2,1}\ , \tag{19}$$

where $\Lambda_1^{(t+1)} = W^{(t+1)} + \frac{U_1^{(t)}}{\rho}$.

It is clear that Eq. (19) can be decoupled into

$$q_{i,\cdot}^{(t+1)} = \arg\min_{q_{i,\cdot}} \frac{1}{2}\|q_{i,\cdot} - {\alpha_1}_{i,\cdot}^{(t+1)}\|^2 + \frac{\lambda_1}{\rho}\|q_{i,\cdot}\|_1\ , \tag{20}$$

where $q_{i,\cdot}$ and ${\alpha_1}_{i,\cdot}$ are the i-th row of $Q^{(t+1)}$ and $\Lambda_1^{(t+1)}$ respectively. Since the update of $q_{i,\cdot}^{(t+1)}$ is strictly convex, it can be observed that $q_{i,\cdot}^{(t+1)}$ is its unique minimizer. Then we can employ the following lemma to update $Q^{(t+1)}$ according to Ref. [26].

Lemma 1. For any $\lambda_1 \geq 0$, we can calculate Eq. (20) by the following:

$$q_{i,\cdot}^{(t+1)} = \frac{\max\left(\left\|{\alpha_1}_{i,\cdot}^{(t+1)}\right\|_2 - \frac{\lambda_1}{\rho}, 0\right)}{\left\|{\alpha_1}_{i,\cdot}^{(t+1)}\right\|_2}{\alpha_1}_{i,\cdot}^{(t+1)}\ . \tag{21}$$

Update *P*: According to Eq. (13), the update of P can be solved as follows:

$$P^{(t+1)} = \arg\min_P \lambda_2\|P\|_1 + < U_2^{(t)}, SW^{(t+1)} - P > + \frac{\rho}{2}\|SW^{(t+1)} - P\|^2\ , \tag{22}$$

which is equivalent to the following problem:

$$P^{(t+1)} = \arg\min_P \frac{1}{2}\|P - \Lambda_2^{(t+1)}\|^2 + \frac{\lambda_2}{\rho}\|P\|_1\ , \tag{23}$$

where $\Lambda_2^{(t+1)} = SW^{(t+1)} + \frac{U_2^{(t)}}{\rho}$. Then we can use the following lemma to update $P^{(t+1)}$ [48].

Lemma 2. For any $\lambda_1 \geq 0$, we can calculate Eq. (23) by the following:

$$p_{ij}^{(t+1)} = sign\left(\alpha_{2ij}^{(t+1)}\right)\max\left(\left|\alpha_{2ij}^{(t+1)}\right| - \frac{\lambda_1}{\rho}, 0\right)\ , \tag{24}$$

where $p_{ij}^{(t+1)}$ and $\alpha_{2ij}^{(t+1)}$ denote the element of matrix $P^{(t+1)}$ and $\Lambda_2^{(t+1)}$ respectively.

Update $U_1^{(t+1)}$ and $U_2^{(t+1)}$: According to the standard ADMM, the updates of augmented lagrangian multipliers are as follows:

$$\begin{aligned} U_1^{(t+1)} = \ & U_1^{(t)} + \rho\left(W^{(t+1)} - Q^{(t+1)}\right)\ , \\ U_2^{(t+1)} = \ & U_2^{(t)} + \rho\left(SW^{(t+1)} - P^{(t+1)}\right)\ . \end{aligned} \tag{25}$$

Taken together, the proposed method can be summarized in the Algorithm 1.

With Eq. (26), we can solve the G_1-norm problem by lemma 3 [49].

$$R_{G_i,\cdot}^{(t+1)} = \arg\min_{R_{G_i,\cdot}} \frac{1}{2}\|R_{G_i,\cdot} - \Lambda_{3G_i,\cdot}^{(t+1)}\|^2 + \frac{\lambda_3}{\rho}\|R_{G_i,\cdot}\|\ , \tag{26}$$

where $R_{G_i,\cdot}$ and $\Lambda_{3G_i,\cdot}$ are the rows of R and Λ_3 respectively, corresponding to the features in the group G_i.

Algorithm 1. ADMM optimization of FAS-MTFL

Algorithm 1 ADMM optimization of FAS-MTFL

Input: Feature matrix X, Target matrix Y, λ_1, λ_2, ρ.
Output: W.

1: Initialization: $W^{(0)} \leftarrow 0$, $Q^{(0)} \leftarrow 0$, $P^{(0)} \leftarrow 0$, $U_1^{(0)} \leftarrow 0$, $U_2^{(0)} \leftarrow 0$.
2: Compute the Cholesky factorization of F.
3: **repeat**
4: Update $W^{(t+1)}$ according to Eq. (14).
5: Update $Q^{(t+1)}$ according to Eq. (18).
6: Update $P^{(t+1)}$ according to Eq. (22).
7: Update $U_1^{(t+1)}$, $U_2^{(t+1)}$ according to Eq. (25).
8: **until** Convergence.

Lemma 3. For any $\lambda_3 \geq 0$, we have

$$R_{G_l,\cdot}^{(t+1)} = \frac{\max\left(\left\|\Lambda_{3G_l,\cdot}\right\|_2 - \frac{\lambda_3 v_l}{\rho}, 0\right)}{\left\|\Lambda_{3G_l,\cdot}\right\|_2}\Lambda_{3G_l,\cdot}, \tag{27}$$

where v_l is the weight of the l-th feature group. Then the dMTLc model in Eq. (10) could be optimized by the algorithm similar to the Algorithm 1. With the detailed analysis in Ref. [38], the dMTLl model in Eq. (11) could also be optimized by the similar algorithm.

5. Experiment

In this section, we first show the results on a synthetic dataset, and then the proposed models are extensively evaluated on the Alzheimer's disease neuroimaging initiative (ADNI) datasets.

5.1. Simulation study

In this section, we carry out simulations to demonstrate that: (1) the proposed generalized multi-task formulation framework that considers both the task and feature correlation outperforms MTFL that only considers the task correlation; (2) our FAS-norm penalty could address the too restrictive assumption issue of the $\ell_{2,1}$-norm penalty and improve the prediction performance; (3) the proposed methods are still effective in the scenario where tasks have a different number of features.

5.1.1. Data generation

In a real-life scenario, each task could have a different number of features. For ease of analysis, we construct a simplified case where six related tasks ($t = 6$) are trained together. The first three tasks have D_1 features, and the last three tasks have D_2 features. The dimensionalities of tasks are denoted by (D_1, D_2). We consider three possible setups, (D_1, D_2) $\in$ {(400, 300), (500, 400), (600, 500)}, and generate synthetic data by the linear model $\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{\xi}$.

First, we generate 50 samples for each setup. The input feature matrix $X \in \mathbb{R}^{50\times D_1}$ is a randomly generated matrix that follows the Gaussian distribution with zero mean and standard deviation of 1. Then the feature matrix X is multiplied by a matrix to make sure that only the correlation coefficient of $0.1*D_1$ pairs of features are above 0.5. Thus, the strong correlation among features is sparse. The first three tasks feed the whole X and the last three tasks feed the first D_2 features of the X. Second, we generate the true models for each setup. The weight vectors for the first three tasks are in D_1 dimension consisting of g blocks of five, i.e. $\boldsymbol{w} \in \mathbb{R}^{D_1}$, and $v_i = 5$, $g = D_1/5$ for the disjoint feature groups G_i, $i = 1$, 2, …, g. The weight vectors for the last three tasks are similar as the above. The number of the non-zero block is 15 for each weight vector. Then, the weights of feature pairs that the correlation coefficient up than 0.5 are nonzero, and the weights are identical for each feature pair. The non-zero weights are generated following the Gaussian distribution with zero mean and standard deviation of 1. Finally, the input label $\boldsymbol{y}$ is the product of X and $\boldsymbol{w}$ with an additive Gaussian noise $\boldsymbol{\xi}$ as $\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{\xi}$, where the Gaussian noise $\boldsymbol{\xi}$ is generated following the Gaussian distribution with zero mean and standard deviation of 0.1.

5.1.2. Settings

The penalty that incorporates the task correlation is not suitable directly in the scenario where multiple tasks have a different number of features. To address this issue, we partition the models into multiple blocks according to the dimensionality of tasks and the type of the penalty. We update the weights for the different penalties based on the blocks, as illustrated in the red boxes in Fig. 7. For this particular case, we partition the models into eight blocks, where the first block includes D_2 dimension across all the tasks, the second block includes D_1 dimension across the first three tasks, while each of the third to eighth blocks includes a whole weight vector. The penalty $R_t(W)$ that incorporates the task correlation denotes the $\ell_{2,1}$-norm for this particular case, and the $R_f(W)$ includes the FAS-norm and the G_1-norm. Specifically, for the third to fifth blocks, we introduce the feature correlation matrix $C_1 \in \mathbb{R}^{D_1\times D_1}$ constructed by D_1 features of X. For the sixth to eighth blocks, we

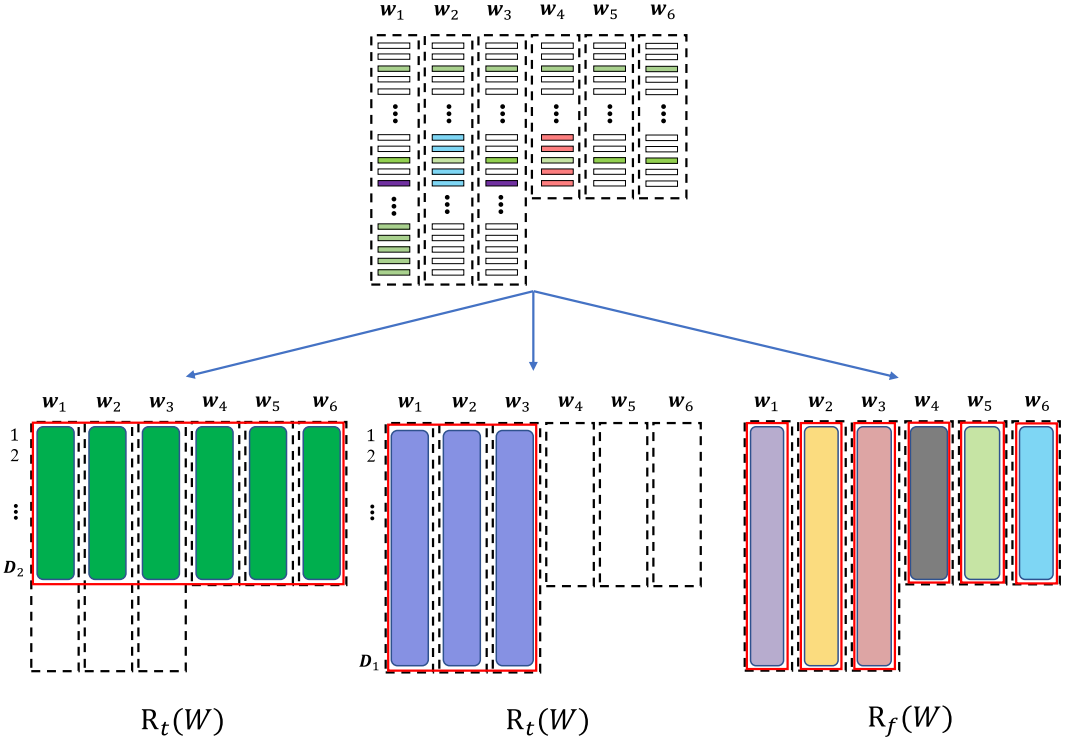


Fig. 7. Illustration of the proposed algorithm framework to apply in the scenario that the tasks have a different number of features.

introduce the feature correlation matrix $C_2 \in \mathbb{R}^{D_2 \times D_2}$ constructed by the first D_2 features of X.

We compare our models (FAS-MTFL and dMTLc) with MTFL. Here, the correlation graphs on the features are sparse with a cutoff of 0.5. Note that the feature correlation graph is calculated using the training data only. We consider the performance measures in terms of the normalized mean squared error (nMSE, Eq. (28)) [25,50] and weighted R-value (wR, Eq. (29)) [51] that evaluate the overall performance of all tasks. The measures are defined as follows:

$$\mathrm{nMSE}(Y, \widehat{Y}) = \frac{\sum_{h=1}^{k} \frac{\left\| Y_h - \widehat{Y}_h \right\|_2^2}{\sigma(Y_h)}}{\sum_{h=1}^{k} n_h} , \tag{28}$$

$$\mathrm{wR}(Y, \widehat{Y}) = \frac{\sum_{h=1}^{k} \mathrm{Corr}(Y_h, \widehat{Y}_h) n_h}{\sum_{h=1}^{k} n_h} , \tag{29}$$

where Y is the ground truth of the target and $\widehat{Y}$ is the prediction value. Especially, for nMSE, the smaller the value, the better the model performance and the larger value of wR indicates the better performance.

Grid search for each method is performed, where the range of each parameter varies from 0.1 to 1000, and the same training and testing data are used for all methods to provide a fair comparison. We randomly split the data into training and testing sets using a ratio of 9:1, i.e., we build models on 90% of the data and evaluate these models on the remaining 10% of the data. In each of the ten trials, a 5-fold nested cross-validation procedure is employed to tune the regularization parameters. The reported results are the best results of each method with the optimal parameters.

5.1.3. Results

The results of the simulation study are presented in Fig. 8. As can be seen, our proposed methods consistently show the better performance over the MTFL. This result is in line with our intuition that, when the task correlation and the feature correlation are considered simultaneously, multi-task learning methods may discover more stable patterns and then perform more accurate prediction results. For FAS-MTFL, we obtain that its performance improves 10.22%, 6.18%, and 10.23% in nMSE for three setups respectively compared to MTFL. This means that our proposed FAS-norm penalty is able to capture the feature correlation in multi-task learning and help to improve the prediction performance. The above results demonstrate that with the adjusted algorithm illustrated in Fig. 7, our proposed multi-task learning framework and FAS-norm are still effective in the scenario where tasks have a different number of features.

5.2. ADNI data study

In this section, the proposed models are extensively evaluated on the cross-sectional and longitudinal datasets from ADNI. We first evaluate the prediction performance of the methods and then analyze the identified biomarkers.

5.2.1. Data

The real datasets used in this work were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database.[1] As such, the investigators within the ADNI contributed to the design and implementation of the ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found.[2]

The goal of ADNI is to validate and standardize biomarkers for AD clinical trials. In ADNI, all subjects received 1.5 T structural MRI. The MRI features used in this paper are based on the imaging data from the ADNI database processed by the UCSF (the University of California at San Francisco) team. They performed cortical reconstruction and volumetric segmentation with the FreeSurfer Software Suite[3] according to the Desikan-Killany atlas [52].

Briefly, this processing includes motion correction and averaging [53] of multiple volumetric T1 weighted images (when more than one is available), removal of non-brain tissue using a hybrid watershed/surface deformation procedure [54], automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles) [55,56] intensity normalization [57], tessellation of the gray matter white matter boundary, automated topology correction [58,59], and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class [60–62].

In total, 71 cortical regions and 44 subcortical regions were generated with typically 4 or 1 feature in each region, and the names of the regions are listed in Table 2. It can be observed that each cortical region contains four features: cortical thickness average (TA), the standard deviation of a thickness (TS), surface area (SA), and cortical volume (CV). Each subcortical region contains one feature: subcortical volume (SV). The surface area (SA) for the hemisphere and the total intracranial volume (ICV) are a bit different from the above two regions. In conclusion, 319 ($=34 \times 2 \times 4 + 1 \times 2 + 1 + 16 \times 2 + 12$) features are involved in the experiments.

The clinic cognitive scores explored in this study are shown in Table 3. Note that the cognitive scores of patients are given by the experts according to the gold standard. The further preprocessing of the data including (1) removing the samples without baseline MRI records, (2) deleting ROIs whose name is "unknown", (3) deleting features whose entries are missed more than 10% (for all patients and all time points), (4) deleting samples without labels, (5) replacing the remaining missing values with average values. Finally, 788 samples at the baseline are obtained as shown in Table 4, where the subjects are categorized into three groups: Normal Control (NC), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). The demographics information of all subjects at the baseline are shown in Table 5, including age, gender, and education. All input data have been normalized by z-scored before applying regression methods.

5.2.2. Settings

In the cross-sectional analysis, the models are applied on multiple cognitive tests and it is assumed that these cognitive tests are correlated. Besides, each task inputs the same data (features and sample numbers). In the longitudinal analysis, modeling approaches are constructed to predict future scores of the specific cognitive test only using the baseline MRI data as the input. Since lack of data, each model inputs the same features but different sample numbers.

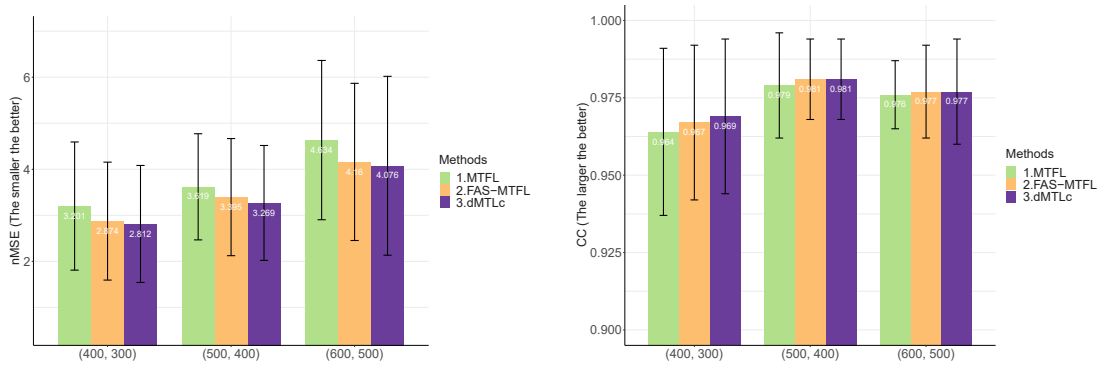
Grid search for each method is performed, and the same training and testing data are used for all methods to provide a fair comparison. Note that the feature correlation graph is calculated using the training data only. For the quantitative performance evaluation, we employ the metrics of correlation coefficient (CC, Eq. (30)) and root mean squared error (rMSE, Eq. (31)) between the predicted clinical scores and the target clinical scores for each regression task. The measures are defined as follows:

$$\mathrm{CC}(y, \widehat{y}) = \frac{cov(y, \widehat{y})}{\sigma(y)\sigma(\widehat{y})} , \tag{30}$$

[1] http://adni.loni.usc.edu/.

[2] http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[3] http://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferMethodsCitation.



(a) nMSE (b) wR

Fig. 8. Comparison of different methods on the synthetic data. For nMSE, the smaller the value, the better the model performance. For wR, the larger the value, the better the model performance.

Table 2
Features from the following 71 cortical regions (the left four columns) and 44 subcortical regions (the right four columns) generated by FreeSurfer.

ID	ROI name	Laterality	Type	ID	ROI name	Laterality	Type
1	Banks superior temporal sulcus	L, R	CV, SA, TA, TS	1	Accumbens area	L, R	SV
2	Caudal anterior cingulate cortex	L, R	CV, SA, TA, TS	2	Amygdala	L, R	SV
3	Caudal middle frontal gyrus	L, R	CV, SA, TA, TS	3	Caudate	L, R	SV
4	Cuneus cortex	L, R	CV, SA, TA, TS	4	Cerebellum cortex	L, R	SV
5	Entorhinal cortex	L, R	CV, SA, TA, TS	5	Cerebellum white matter	L, R	SV
6	Frontal pole	L, R	CV, SA, TA, TS	6	Cerebral cortex	L, R	SV
7	Fusiform gyrus	L, R	CV, SA, TA, TS	7	Cerebral white matter	L, R	SV
8	Inferior parietal cortex	L, R	CV, SA, TA, TS	8	Choroid plexus	L, R	SV
9	Inferior temporal gyrus	L, R	CV, SA, TA, TS	9	Hippocampus	L, R	SV
10	Insula	L, R	CV, SA, TA, TS	10	Inferior lateral ventricle	L, R	SV
11	IsthmusCingulate	L, R	CV, SA, TA, TS	11	Lateral ventricle	L, R	SV
12	Lateral occipital cortex	L, R	CV, SA, TA, TS	12	Pallidum	L, R	SV
13	Lateral orbital frontal cortex	L, R	CV, SA, TA, TS	13	Putamen	L, R	SV
14	Lingual gyrus	L, R	CV, SA, TA, TS	14	Thalamus	L, R	SV
15	Medial orbital frontal cortex	L, R	CV, SA, TA, TS	15	Ventricle diencephalon	L, R	SV
16	Middle temporal gyrus	L, R	CV, SA, TA, TS	16	Vessel	L, R	SV
17	Paracentral lobule	L, R	CV, SA, TA, TS	17	Brain stem	Bilateral	SV
18	Parahippocampal gyrus	L, R	CV, SA, TA, TS	18	Corpus callosum anterior	Bilateral	SV
19	Pars opercularis	L, R	CV, SA, TA, TS	19	Corpus callosum central	Bilateral	SV
20	Pars orbitalis	L, R	CV, SA, TA, TS	20	Corpus callosum middle anterior	Bilateral	SV
21	Pars triangularis	L, R	CV, SA, TA, TS	21	Corpus callosum middle posterior	Bilateral	SV
22	Pericalcarine cortex	L, R	CV, SA, TA, TS	22	Corpus callosum posterior	Bilateral	SV
23	Postcentral gyrus	L, R	CV, SA, TA, TS	23	Cerebrospinal fluid	Bilateral	SV
24	Posterior cingulate cortex	L, R	CV, SA, TA, TS	24	Fourth ventricle	Bilateral	SV
25	Precentral gyrus	L, R	CV, SA, TA, TS	25	Non white matter hypointensities	Bilateral	SV
26	Precuneus cortex	L, R	CV, SA, TA, TS	26	Optic chiasm	Bilateral	SV
27	Rostral anterior cingulate cortex	L, R	CV, SA, TA, TS	27	Third ventricle	Bilateral	SV
28	Rostral middle frontal gyrus	L, R	CV, SA, TA, TS	28	White matter hypointensities	Bilateral	SV
29	Superior frontal gyrus	L, R	CV, SA, TA, TS				
30	Superior parietal cortex	L, R	CV, SA, TA, TS				
31	Superior temporal gyrus	L, R	CV, SA, TA, TS				
32	Supramarginal gyrus	L, R	CV, SA, TA, TS				
33	Temporal pole	L, R	CV, SA, TA, TS				
34	Transverse temporal cortex	L, R	CV, SA, TA, TS				
35	Hemisphere	L, R	SA				
36	Total intracranial volume	Bilateral	ICV				

$$\mathrm{rMSE}(y,\widehat{y}) = \frac{\|y-\widehat{y}\|_2^2}{n}, \tag{31}$$

where y is the ground truth of the target and $\widehat{y}$ is the prediction value. Moreover, nMSE (Eq. (28)) and wR (Eq. (29)) are used to evaluate the overall performance of all tasks, which have been defined in Section 5.1.2. Especially, for rMSE and nMSE, the smaller the value, the better the model performance, and the larger values of CC and wR indicate the better performance.

We randomly split the data into training and testing sets using a ratio of 9:1, i.e., the models are built on 90% of the data and evaluated on the remaining 10% of the data. In each of the ten trials, a 5-fold nested cross-validation procedure is employed to tune the regularization parameters. The range of each parameter varies from 0.1 to 1000. The reported results are the best results of each method with the optimal parameters.

5.2.3. *Performance in the cross-sectional analysis*

In this experiment, we evaluate the effectiveness of our proposed FAS-MTFL and dMTLc methods on the cross-sectional scores prediction, comparing with 11 algorithms including:

(1) single-task learning algorithms: Ridge, Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Lasso;

Table 3
The clinic cognitive scales explored in this study.

Score Name		Description
ADAS		Alzheimers Disease Assessment Scale
MMSE		Mini-Mental State Exam
RAVLT	TOTAL	Total score of the first 5 learning trials
	TOT6	Trial 6 total number of words recalled
	TOTB	Immediately after the fifth learning trial
	T30	30 min delay total number of words recalled
	RECOG	30 min delay recognition
FLU	ANIM	Animal Total score
	VEG	Vegetable Total score
LOGMEM	IMMTOTAL	Immediate recall
	DELTOTAL	Delayed recall
CLOCK	DRAW	Clock drawing
	COPYSCORE	Clock copying
BOSNAM		Total number correct
ANART		ANART total score
DSPAN	For	Digit span forward
	BAC	Digit span backward
DIGIT		Digit symbol substitution

Table 4
Summary of ADNI dataset.

Time point	Category			Total
	NC	MCI	AD	
Baseline	225	390	173	788
Month 6	211	352	155	718
Month 12	198	330	134	662
Month 24	177	254	101	532
Month 36	155	189	1	345

Table 5
Summary of the demographics information for subjects at the baseline.

Category	NC	MCI	AD
Number	225	390	173
Gender (Male/Female)	116/109	252/138	88/85
Age (year)	75.87 ± 5.04	74.75 ± 7.39	75.42 ± 7.25
Education (year)	16.03 ± 2.85	15.67 ± 2.95	14.65 ± 3.17

(2) multi-task learning algorithms that only incorporate the task correlation structures: Multi-task feature learning (MTFL), Multi-task feature learning combined with lasso (SGL-MTFL) [28], Robust Multi-Task Learning (RMTL) [45], Robust Multi-Task Feature Learning (rMTFL) [63] and Trace-norm Multi-Task Learning (Trace) [64];

(3) multi-task learning algorithms that incorporate the feature and task correlation structures: Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) [30].

Two experiments are designed in the cross-sectional analysis: (1) five of the most common cognitive scores are predicted simultaneously (ADAS, MMSE, RAVLT.TOTAL, RAVLT.T30, and RAVLT.RECOG), (2) all of eighteen cognitive scores in Table 3 are predicted simultaneously. Therefore, our experiments could also reveal how the number of cognitive tasks learned at the same time affects the prediction performance of the model. Note that, all methods are evaluated on baseline MRI input and baseline cognitive output, and only one model is trained at a time in single-task learning. The best performance of each column is boldfaced.

The results of the first experiment are reported in Tables 6 and 7. As can be seen, our proposed methods consistently showed the best performance over the baseline methods (Ridge, RF, SVM, XGBoost, Lasso, and MTFL) on overall prediction measures (nMSE and wR). More specifically, the nMSE measurements of FAS-MTFL and dMTLc achieve declines of 7.16% and 7.01% decline compared to MTFL respectively. FAS-MTFL and dMTLc achieve CC gains of 3.91% and 3.52% compared to MTFL respectively. This means that our proposed framework and the FAS-norm penalty could explore the potential feature relationship, which could help to improve predictive performance.

It can also be observed that the performance of FAS-MTFL is similar to that of SGL-MTFL, and SGL-MTFL performs better than dMTLc. First, this may be due to the limited data size of ADNI, which results in the insufficiency of the model training, so it may produce some results that are not significant enough. Second, the learning of each model is relatively easy for these five tasks, and the gap among tasks is small. Therefore, adding task correlation structure information in the regularization is more beneficial and the improvement caused by incorporating feature correlation structure is not significant. Third, the group feature correlation imposed in dMTLc may be too restrictive in the five tasks experimental condition, thus it weakens the sharing of information among tasks and yields suboptimal performance when the task correlation information is more effective.

Experimental results on 18 cognitive scores in Table 3 are shown in Tables 8 and 9. It can be seen that the proposed methods (FAS-MTFL and dMTLc) consistently achieve better prediction performance than the competing methods, which demonstrates the effectiveness of our methods. Besides, the results from Tables 8 and 9 can be compared with the results in Tables 6 and 7 It is observed that the prediction performance of the dMTLc model improves significantly on nMSE. A possible explanation for the observation is that the differences among the cognitive scores are rising along with the number of tasks rising. That is to say, the feature correlation information becomes more important

Table 6
Performance comparison of various methods in terms of rMSE and nMSE on five of the most common cognitive scores. For rMSE and nMSE, the smaller the value, the better the model performance. FAS-MTFL and dMTLc are significantly better than the results marked with * and † respectively. Student's t-test at a level of 0.05 was used.

Method	ADAS	MMSE	RAVLT TOTAL	T30	RECOG	nMSE
Ridge	7.433 ± 0.477	2.783 ± 0.179	11.18 ± 0.788	4.018 ± 0.298	4.283 ± 0.427	5.885 ± 0.626*†
RF	9.643 ± 0.692	3.056 ± 0.164	13.54 ± 1.424	4.678 ± 0.462	5.062 ± 0.375	8.619 ± 0.881*†
SVM	7.835 ± 0.438	2.928 ± 0.258	12.03 ± 0.956	4.347 ± 0.282	4.847 ± 0.435	6.849 ± 0.748*†
XGBoost	7.220 ± 0.742	2.418 ± 0.184	10.56 ± 0.762	3.673 ± 0.282	3.911 ± 0.257	5.114 ± 0.368*†
Lasso	6.718 ± 0.439	2.174 ± 0.093	10.05 ± 0.631	**3.424 ± 0.246**	3.633 ± 0.244	4.485 ± 0.346*†
MTFL	7.045 ± 0.474	2.335 ± 0.216	10.04 ± 0.807	3.535 ± 0.309	3.633 ± 0.187	4.677 ± 0.317*†
SGL-MTFL [28]	**6.674 ± 0.483**	2.186 ± 0.100	9.716 ± 0.675	3.427 ± 0.263	**3.611 ± 0.228**	4.345 ± 0.314
RMTL [45]	7.053 ± 0.443	2.706 ± 0.190	10.86 ± 0.751	3.691 ± 0.271	3.926 ± 0.322	5.314 ± 0.533*†
rMTFL [63]	7.020 ± 0.452	2.569 ± 0.308	10.74 ± 0.735	3.629 ± 0.257	3.935 ± 0.344	5.184 ± 0.498*†
G-SMuRFS [30]	6.725 ± 0.465	**2.170 ± 0.101**	9.741 ± 0.631	3.437 ± 0.227	3.625 ± 0.272	4.373 ± 0.312
Trace [64]	6.835 ± 0.553	3.062 ± 0.285	10.52 ± 0.746	3.605 ± 0.236	3.797 ± 0.275	5.190 ± 0.455*†
FAS-MTFL	6.683 ± 0.487	2.196 ± 0.101	**9.678 ± 0.691**	3.434 ± 0.277	3.614 ± 0.217	**4.342 ± 0.308**
dMTLc	6.711 ± 0.477	2.206 ± 0.100	9.681 ± 0.674	3.437 ± 0.284	3.618 ± 0.222	4.356 ± 0.295

Table 7
Performance comparison of various methods in terms of CC and wR on five of the most common cognitive scores. For CC and wR, the larger the value, the better the model performance. FAS-MTFL and dMTLc are significantly better than the results marked with * and † respectively. Student's t-test at a level of 0.05 was used.

Method	ADAS	MMSE	RAVLT TOTAL	T30	RECOG	wR
Ridge	0.601 ± 0.053	0.421 ± 0.067	0.407 ± 0.124	0.375 ± 0.135	0.268 ± 0.112	0.415 ± 0.079*†
RF	0.455 ± 0.069	0.331 ± 0.045	0.304 ± 0.137	0.317 ± 0.098	0.221 ± 0.053	0.326 ± 0.039*†
SVM	0.571 ± 0.054	0.345 ± 0.086	0.353 ± 0.131	0.342 ± 0.125	0.204 ± 0.103	0.363 ± 0.074*†
XGBoost	0.599 ± 0.064	0.430 ± 0.105	0.412 ± 0.092	0.450 ± 0.069	0.330 ± 0.114	0.444 ± 0.055*†
Lasso	0.660 ± 0.065	0.546 ± 0.068	0.479 ± 0.107	0.513 ± 0.115	0.406 ± 0.111	0.521 ± 0.079*†
MTFL	0.633 ± 0.086	0.518 ± 0.067	0.490 ± 0.129	0.485 ± 0.126	**0.428 ± 0.124**	0.511 ± 0.092*†
SGL-MTFL [28]	**0.664 ± 0.070**	0.545 ± 0.069	0.513 ± 0.099	**0.516 ± 0.119**	0.418 ± 0.128	**0.531 ± 0.082**
RMTL [45]	0.631 ± 0.053	0.429 ± 0.066	0.431 ± 0.116	0.437 ± 0.123	0.337 ± 0.109	0.453 ± 0.079*†
rMTFL [63]	0.637 ± 0.050	0.479 ± 0.054	0.434 ± 0.118	0.448 ± 0.117	0.330 ± 0.109	0.465 ± 0.074*†
G-SMuRFS [30]	0.658 ± 0.065	**0.551 ± 0.060**	0.509 ± 0.099	0.511 ± 0.111	0.412 ± 0.121	0.528 ± 0.079
Trace [64]	0.653 ± 0.050	0.367 ± 0.098	0.450 ± 0.116	0.456 ± 0.115	0.360 ± 0.124	0.457 ± 0.087*†
FAS-MTFL	0.663 ± 0.072	0.543 ± 0.072	**0.518 ± 0.098**	0.515 ± 0.118	0.419 ± 0.129	**0.531 ± 0.082**
dMTLc	0.660 ± 0.069	0.539 ± 0.072	0.517 ± 0.099	0.514 ± 0.117	0.417 ± 0.132	0.529 ± 0.082

Table 8
Performance comparison of various methods in terms of rMSE and nMSE on eighteen cognitive scores. For rMSE and nMSE, the smaller the value, the better the model performance. FAS-MTFL and dMTLc are significantly better than the results marked with * and † respectively. Student's t-test at a level of 0.05 was used.

Method	ADAS	MMSE	RAVLT TOTAL	TOT6	TOTB	T30	RECOG
Ridge	7.433 ± 0.477	2.783 ± 0.179	11.18 ± 0.788	3.859 ± 0.380	1.984 ± 0.117	4.018 ± 0.298	4.283 ± 0.427
RF	9.643 ± 0.692	3.056 ± 0.164	13.54 ± 1.424	4.678 ± 0.318	2.377 ± 0.143	4.678 ± 0.462	5.062 ± 0.375
SVM	7.835 ± 0.438	2.928 ± 0.258	12.03 ± 0.956	4.115 ± 0.343	2.309 ± 0.082	4.347 ± 0.282	4.847 ± 0.435
XGBoost	7.220 ± 0.742	2.418 ± 0.184	10.56 ± 0.762	3.594 ± 0.274	1.805 ± 0.181	3.673 ± 0.281	3.910 ± 0.257
Lasso	6.936 ± 0.670	2.258 ± 0.169	10.43 ± 0.767	3.422 ± 0.303	1.731 ± 0.199	3.517 ± 0.210	3.776 ± 0.281
MTFL	6.881 ± 0.489	2.248 ± 0.105	9.715 ± 0.776	3.339 ± 0.255	**1.651 ± 0.162**	3.471 ± 0.270	**3.608 ± 0.181**
SGL-MTFL [28]	6.689 ± 0.466	**2.191 ± 0.104**	9.815 ± 0.707	3.317 ± 0.281	1.664 ± 0.163	3.434 ± 0.280	3.618 ± 0.229
RMTL [45]	7.048 ± 0.473	2.813 ± 0.390	10.93 ± 0.751	3.594 ± 0.372	1.782 ± 0.140	3.727 ± 0.293	3.929 ± 0.420
rMTFL [63]	6.991 ± 0.443	2.375 ± 0.235	10.79 ± 0.686	3.468 ± 0.330	1.695 ± 0.155	3.602 ± 0.253	3.836 ± 0.401
G-SMuRFS [30]	6.899 ± 0.533	2.258 ± 0.102	**9.673 ± 0.794**	3.324 ± 0.256	1.654 ± 0.158	3.442 ± 0.296	**3.608 ± 0.202**
Trace [64]	6.885 ± 0.551	2.932 ± 0.132	10.55 ± 0.777	3.481 ± 0.298	1.729 ± 0.136	3.619 ± 0.242	3.748 ± 0.278
FAS-MTFL	6.696 ± 0.461	2.210 ± 0.102	9.725 ± 0.743	3.319 ± 0.273	1.674 ± 0.165	3.434 ± 0.283	3.617 ± 0.210
dMTLc	**6.679 ± 0.499**	2.203 ± 0.090	9.694 ± 0.644	**3.316 ± 0.263**	1.680 ± 0.168	**3.427 ± 0.291**	3.616 ± 0.223

Method	FLU ANIM	VEG	LOGMEM IMMTOTAL	DELTOTAL	CLOCK DRAW	COPYSCORE	BOSNAM RECOG
Ridge	6.312 ± 0.603	4.284 ± 0.391	4.673 ± 0.399	5.211 ± 0.542	1.155 ± 0.104	0.779 ± 0.041	4.675 ± 0.423
RF	7.264 ± 0.530	5.252 ± 0.392	6.153 ± 0.640	6.691 ± 0.424	1.374 ± 0.126	0.875 ± 0.130	5.477 ± 0.494
SVM	7.047 ± 0.463	4.864 ± 0.316	5.047 ± 0.411	5.626 ± 0.542	1.314 ± 0.149	0.776 ± 0.076	4.868 ± 0.583
XGBoost	5.641 ± 0.542	4.122 ± 0.308	4.527 ± 0.246	4.957 ± 0.464	1.023 ± 0.141	0.705 ± 0.083	4.313 ± 0.554
Lasso	5.554 ± 0.434	3.755 ± 0.181	4.382 ± 0.424	4.778 ± 0.514	1.022 ± 0.093	0.665 ± 0.079	4.113 ± 0.553
MTFL	5.251 ± 0.492	3.729 ± 0.237	4.142 ± 0.377	4.560 ± 0.509	**0.971 ± 0.110**	**0.648 ± 0.882**	4.044 ± 0.501
SGL-MTFL [28]	5.264 ± 0.505	3.681 ± 0.206	4.162 ± 0.358	4.549 ± 0.494	0.988 ± 0.108	0.658 ± 0.086	**3.945 ± 0.462**
RMTL [45]	5.861 ± 0.605	3.993 ± 0.283	4.442 ± 0.366	4.897 ± 0.507	1.054 ± 0.091	0.775 ± 0.120	4.484 ± 0.386
rMTFL [63]	5.599 ± 0.493	3.846 ± 0.281	4.299 ± 0.307	4.768 ± 0.491	1.004 ± 0.154	0.688 ± 0.156	4.211 ± 0.511
G-SMuRFS [30]	5.245 ± 0.482	3.717 ± 0.232	4.162 ± 0.374	4.565 ± 0.523	0.973 ± 0.108	0.649 ± 0.086	4.044 ± 0.526
Trace [64]	5.532 ± 0.531	3.873 ± 0.281	4.334 ± 0.376	4.747 ± 0.422	1.013 ± 0.118	0.716 ± 0.072	4.382 ± 0.364
FAS-MTFL	5.249 ± 0.505	3.686 ± 0.218	4.143 ± 0.372	4.521 ± 0.523	1.009 ± 0.108	0.673 ± 0.088	3.960 ± 0.473
dMTLc	**5.236 ± 0.468**	**3.679 ± 0.205**	**4.139 ± 0.372**	**4.519 ± 0.540**	1.004 ± 0.095	0.668 ± 0.074	3.953 ± 0.490

Method	ANART	DSPAN For	DSPAN BAC	DIGIT	nMSE
Ridge	11.21 ± 0.731	2.405 ± 0.207	2.571 ± 0.188	12.76 ± 1.305	5.354 ± 0.325*†
RF	13.32 ± 1.488	2.761 ± 0.187	2.999 ± 0.324	16.41 ± 1.060	7.954 ± 0.469*†
SVM	12.12 ± 0.864	2.848 ± 0.222	3.062 ± 0.138	13.64 ± 1.268	6.370 ± 0.355*†
XGBoost	10.32 ± 0.707	2.187 ± 0.135	2.382 ± 0.236	12.17 ± 0.918	4.644 ± 0.187*†
Lasso	10.39 ± 1.233	2.072 ± 0.235	2.192 ± 0.186	12.26 ± 1.524	4.419 ± 0.530*†
MTFL	9.434 ± 0.698	2.004 ± 0.151	**2.117 ± 0.183**	11.58 ± 1.275	3.991 ± 0.229*†
SGL-MTFL [28]	9.500 ± 0.680	2.002 ± 0.151	2.131 ± 0.189	11.43 ± 1.296	3.960 ± 0.223*†
RMTL [45]	10.51 ± 0.696	2.174 ± 0.150	2.266 ± 0.199	12.59 ± 1.219	4.815 ± 0.318*†
rMTFL [63]	10.39 ± 0.730	2.036 ± 0.730	2.167 ± 0.208	12.44 ± 1.169	4.512 ± 0.278*†
G-SMuRFS [30]	9.425 ± 0.694	2.010 ± 0.154	2.123 ± 0.190	11.57 ± 1.297	3.984 ± 0.216*†
Trace [64]	10.01 ± 0.666	2.124 ± 0.126	2.214 ± 0.206	12.00 ± 1.306	4.485 ± 0.250*†
FAS-MTFL	9.451 ± 0.679	2.006 ± 0.156	2.141 ± 0.192	11.38 ± 1.263	3.945 ± 0.221
dMTLc	**9.421 ± 0.680**	**1.999 ± 0.155**	2.133 ± 0.193	**11.28 ± 1.273**	**3.920 ± 0.194**

relatively, while task correlation information becomes less important. Therefore, dMTLc incorporating more complicated feature correlation knowledge performs better.

It is also observed that several multi-task learning methods (RMTL, rMTFL, and Trace) have shown worse prediction performance compared to the single-task learning method (Lasso) in the cross-sectional analysis.

Table 9
Performance comparison of various methods in terms of CC and wR on eighteen cognitive scores. For CC and wR, the larger the value, the better the model performance. FAS-MTFL and dMTLc are significantly better than the results marked with * and † respectively. Student's t-test at a level of 0.05 was used.

Method	ADAS	MMSE	RAVLT TOTAL	TOT6	TOTB	T30	RECOG
Ridge	0.601 ± 0.053	0.421 ± 0.067	0.407 ± 0.124	0.362 ± 0.133	0.141 ± 0.090	0.375 ± 0.135	0.268 ± 0.112
RF	0.455 ± 0.069	0.331 ± 0.045	0.304 ± 0.137	0.253 ± 0.080	0.110 ± 0.098	0.317 ± 0.098	0.221 ± 0.053
SVM	0.571 ± 0.054	0.345 ± 0.086	0.353 ± 0.131	0.354 ± 0.125	0.063 ± 0.107	0.342 ± 0.125	0.203 ± 0.103
XGBoost	0.599 ± 0.064	0.430 ± 0.105	0.412 ± 0.092	0.414 ± 0.078	0.203 ± 0.077	0.450 ± 0.069	0.330 ± 0.114
Lasso	0.638 ± 0.071	0.510 ± 0.057	0.455 ± 0.104	0.466 ± 0.111	0.271 ± 0.124	0.489 ± 0.100	0.375 ± 0.131
MTFL	0.638 ± 0.077	0.541 ± 0.066	0.512 ± 0.107	0.488 ± 0.123	**0.331 ± 0.087**	0.495 ± 0.109	**0.419 ± 0.124**
SGL-MTFL [28]	**0.665 ± 0.065**	**0.548 ± 0.068**	0.504 ± 0.094	0.500 ± 0.122	0.320 ± 0.081	0.512 ± 0.111	0.415 ± 0.126
RMTL [45]	0.629 ± 0.052	0.408 ± 0.067	0.421 ± 0.131	0.414 ± 0.145	0.209 ± 0.104	0.434 ± 0.124	0.330 ± 0.114
rMTFL [63]	0.636 ± 0.051	0.506 ± 0.056	0.429 ± 0.122	0.443 ± 0.122	0.275 ± 0.082	0.455 ± 0.117	0.343 ± 0.115
G-SMuRFS [30]	0.638 ± 0.077	0.542 ± 0.065	**0.522 ± 0.097**	0.497 ± 0.118	0.327 ± 0.080	0.511 ± 0.104	**0.419 ± 0.127**
Trace [64]	0.645 ± 0.054	0.366 ± 0.086	0.443 ± 0.128	0.444 ± 0.135	0.244 ± 0.119	0.454 ± 0.116	0.373 ± 0.122
FAS-MTFL	0.664 ± 0.066	0.542 ± 0.071	0.515 ± 0.093	0.500 ± 0.120	0.324 ± 0.088	0.513 ± 0.110	0.417 ± 0.128
dMTLc	0.664 ± 0.066	0.537 ± 0.067	0.515 ± 0.094	**0.503 ± 0.119**	0.307 ± 0.056	**0.519 ± 0.115**	0.418 ± 0.131

Method	FLU ANIM	VEG	LOGMEM IMMTOTAL	DELTOTAL	CLOCK DRAW	COPYSCORE	BOSNAM RECOG
Ridge	0.201 ± 0.130	0.389 ± 0.128	0.418 ± 0.111	0.433 ± 0.121	0.227 ± 0.107	0.133 ± 0.095	0.363 ± 0.145
RF	0.181 ± 0.118	0.272 ± 0.073	0.215 ± 0.100	0.276 ± 0.109	0.174 ± 0.148	0.114 ± 0.119	0.306 ± 0.087
SVM	0.134 ± 0.118	0.314 ± 0.122	0.385 ± 0.096	0.388 ± 0.130	0.174 ± 0.117	0.135 ± 0.106	0.333 ± 0.135
XGBoost	0.296 ± 0.118	0.357 ± 0.106	0.404 ± 0.079	0.445 ± 0.082	0.335 ± 0.126	0.163 ± 0.096	0.377 ± 0.110
Lasso	0.315 ± 0.097	0.495 ± 0.076	0.473 ± 0.115	0.507 ± 0.123	0.334 ± 0.055	0.068 ± 0.115	0.444 ± 0.101
MTFL	0.395 ± 0.084	0.490 ± 0.091	**0.511 ± 0.084**	0.531 ± 0.094	**0.389 ± 0.085**	0.223 ± 0.097	0.465 ± 0.103
SGL-MTFL [28]	0.384 ± 0.100	0.509 ± 0.080	0.503 ± 0.091	0.535 ± 0.102	0.380 ± 0.076	0.232 ± 0.100	**0.484 ± 0.093**
RMTL [45]	0.265 ± 0.129	0.441 ± 0.107	0.455 ± 0.100	0.475 ± 0.112	0.340 ± 0.104	0.166 ± 0.088	0.386 ± 0.129
rMTFL [63]	0.299 ± 0.126	0.473 ± 0.104	0.474 ± 0.103	0.488 ± 0.120	0.373 ± 0.090	0.230 ± 0.097	0.424 ± 0.138
G-SMuRFS [30]	**0.396 ± 0.073**	0.498 ± 0.086	0.508 ± 0.087	0.534 ± 0.092	0.379 ± 0.081	0.224 ± 0.113	0.458 ± 0.082
Trace [64]	0.318 ± 0.112	0.456 ± 0.089	0.467 ± 0.094	0.496 ± 0.092	0.333 ± 0.111	0.165 ± 0.094	0.378 ± 0.124
FAS-MTFL	0.389 ± 0.096	0.509 ± 0.079	**0.511 ± 0.090**	**0.543 ± 0.098**	0.378 ± 0.078	0.231 ± 0.103	0.481 ± 0.086
dMTLc	0.389 ± 0.090	**0.510 ± 0.081**	**0.511 ± 0.092**	**0.543 ± 0.101**	0.365 ± 0.073	**0.234 ± 0.105**	0.479 ± 0.087

Method	ANART	DSPAN For	DSPAN BAC	DIGIT	wR
Ridge	0.049 ± 0.083	0.011 ± 0.060	0.031 ± 0.118	0.390 ± 0.045	0.290 ± 0.055*†
RF	0.057 ± 0.168	0.025 ± 0.099	−0.01 ± 0.114	0.175 ± 0.118	0.210 ± 0.029*†
SVM	0.038 ± 0.106	−0.03 ± 0.056	−0.04 ± 0.103	0.325 ± 0.080	0.244 ± 0.055*†
XGBoost	0.046 ± 0.103	0.020 ± 0.068	0.013 ± 0.131	0.384 ± 0.091	0.316 ± 0.032*†
Lasso	0.100 ± 0.087	0.026 ± 0.073	0.129 ± 0.094	0.402 ± 0.077	0.361 ± 0.051*†
MTFL	0.160 ± 0.121	0.027 ± 0.075	**0.210 ± 0.129**	0.429 ± 0.114	0.403 ± 0.063*†
SGL-MTFL [28]	0.161 ± 0.100	0.050 ± 0.117	0.180 ± 0.130	0.460 ± 0.064	0.408 ± 0.055*
RMTL [45]	0.075 ± 0.073	0.021 ± 0.066	0.121 ± 0.106	0.401 ± 0.066	0.333 ± 0.064*†
rMTFL [63]	0.085 ± 0.071	0.092 ± 0.092	0.157 ± 0.122	0.402 ± 0.043	0.366 ± 0.056*†
G-SMuRFS [30]	0.161 ± 0.116	0.001 ± 0.062	0.189 ± 0.140	0.432 ± 0.107	0.402 ± 0.061*†
Trace [64]	0.098 ± 0.071	−0.03 ± 0.110	0.138 ± 0.086	0.418 ± 0.049	0.345 ± 0.065*†
FAS-MTFL	0.165 ± 0.107	0.057 ± 0.126	0.183 ± 0.139	0.463 ± 0.068	0.410 ± 0.057
dMTLc	**0.176 ± 0.105**	**0.097 ± 0.118**	0.193 ± 0.130	**0.467 ± 0.065**	**0.413 ± 0.056**

Table 10
Performance comparison of various methods in terms of nMSE and wR on the prediction of ADAS, RAVLT.TOTAL, RAVLT.RECOG and FLU.ANIM scores of the longitudinal formulation. For nMSE, the smaller the value, the better the model performance. For wR, the larger the value, the better the model performance. FAS-MTFL and dMTLl are significantly better than the results marked with * and † respectively. Student's t-test at a level of 0.05 was used.

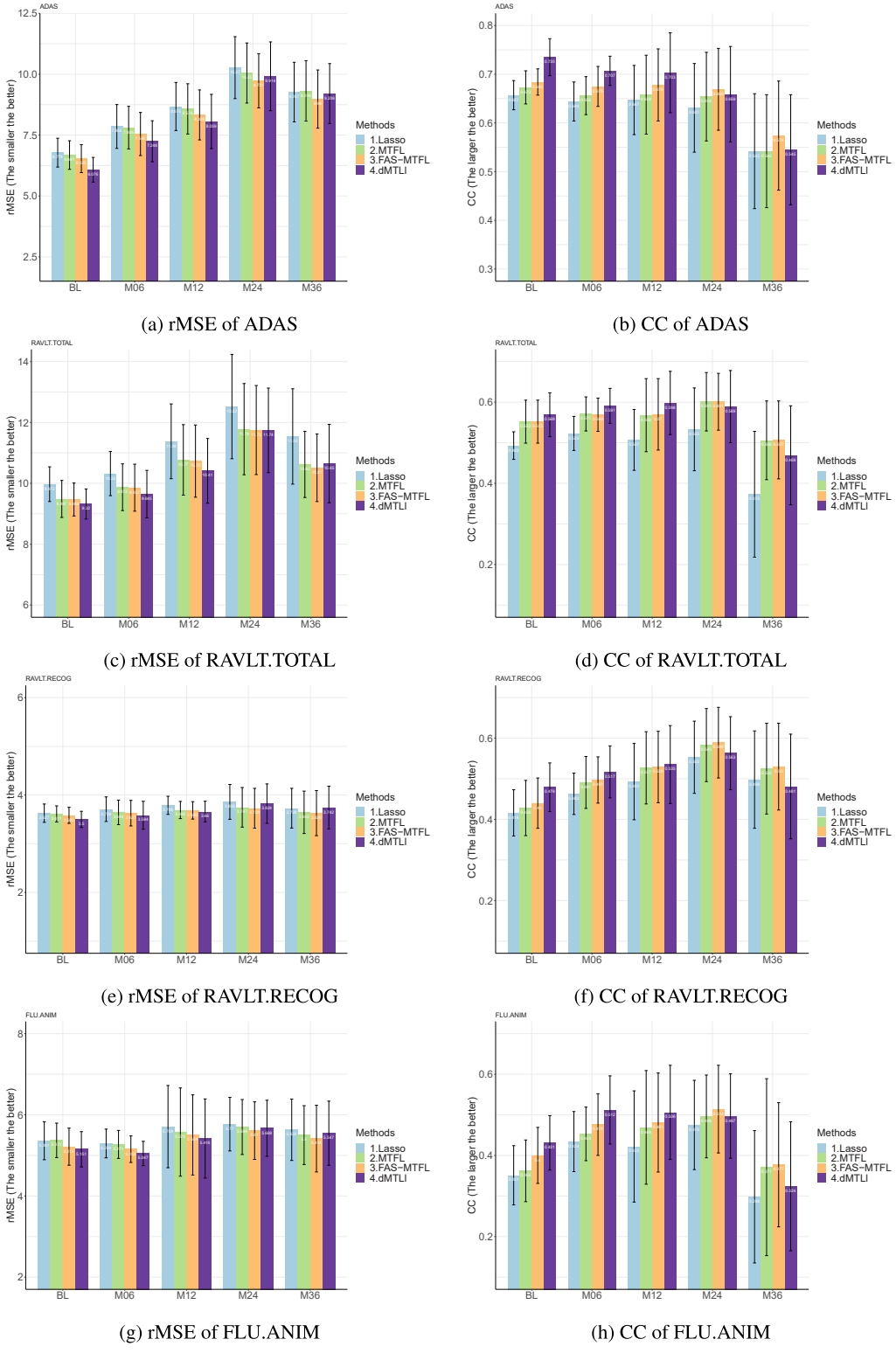
	Lasso	MTFL	FAS-MTFL	dMTLl
Score: ADAS				
nMSE	6.524 ± 0.348*†	6.389 ± 0.339*†	6.005 ± 0.344†	**5.755 ± 0.376**
wR	0.634 ± 0.022*†	0.647 ± 0.026*†	0.665 ± 0.022†	**0.687 ± 0.026**
Score: RAVLT.TOTAL				
nMSE	9.706 ± 0.592*†	8.655 ± 0.555†	8.598 ± 0.511†	**8.378 ± 0.515**
wR	0.496 ± 0.032*†	0.563 ± 0.032†	0.563 ± 0.030	**0.573 ± 0.031**
Score: RAVLT.RECOG				
nMSE	3.281 ± 0.079*†	3.167 ± 0.116*	3.134 ± 0.108	**3.119 ± 0.101**
wR	0.477 ± 0.031*†	0.502 ± 0.035*†	0.509 ± 0.032	**0.515 ± 0.033**
Score: FLU.ANIM				
nMSE	5.082 ± 0.427*†	4.984 ± 0.459*†	4.772 ± 0.375†	**4.716 ± 0.335**
wR	0.402 ± 0.060*†	0.431 ± 0.061*†	0.453 ± 0.053†	**0.466 ± 0.043**

These results demonstrate that the task correlation information added by these multi-task learning methods is not suitable for AD research studies, i.e. the pattern of sharing knowledge among tasks is different for various problem areas, and the prior structural knowledge incorporated in the training phase affects the prediction performance of the model significantly.

5.2.4. *Performance in the longitudinal analysis*

In the longitudinal study, the patients will be followed up for a period of time, and thus the data could be used to build predictive models for multiple time points. To estimate the effectiveness of the proposed models (FAS-MTFL and dMTLl) in the longitudinal analysis, we compare them to the single-task learning method (Lasso) and the multi-task learning method (MTFL) on four of the most common cognitive scores (ADAS, RAVLT.TOTAL, RAVLT.RECOG, and FLU.ANIM). The experiment settings are the same as that of the cross-sectional analysis.

Experimental results of nMSE and wR are reported in Table 10 where the best results are boldfaced. We can derive several observations. First, the multi-task learning models outperform the single-task learning model (Lasso), which justifies the use of longitudinal task correlation information in multi-task learning models, and verifies that the tasks are



(a) rMSE of ADAS (b) CC of ADAS

(c) rMSE of RAVLT.TOTAL (d) CC of RAVLT.TOTAL

(e) rMSE of RAVLT.RECOG (f) CC of RAVLT.RECOG

(g) rMSE of FLU.ANIM (h) CC of FLU.ANIM

Fig. 9. Comparison of different methods on longitudinal prediction in terms of rMSE and CC.

not independent and capturing their relatedness can improve the prediction performance. Second, our proposed FAS-MTFL and dMTLl consistently achieve better prediction performance than MTFL, which demonstrates the effectiveness of our framework and the FAS-norm penalty. That is to say, the combination of the task correlation structure and the feature correlation structure allows more stable and sensitive biomarkers selection for tasks, which could improve the prediction performance. Finally, our dMTLl obtain better prediction results than other models. This demonstrates that our proposed general multi-task learning formulation in Eq. (5) is beneficial to the predicting of disease progression for incorporating complete task and feature correlation information.

Comparative bar charts for all time points between the comparable approaches are shown in Fig. 9. These results reveal several interesting

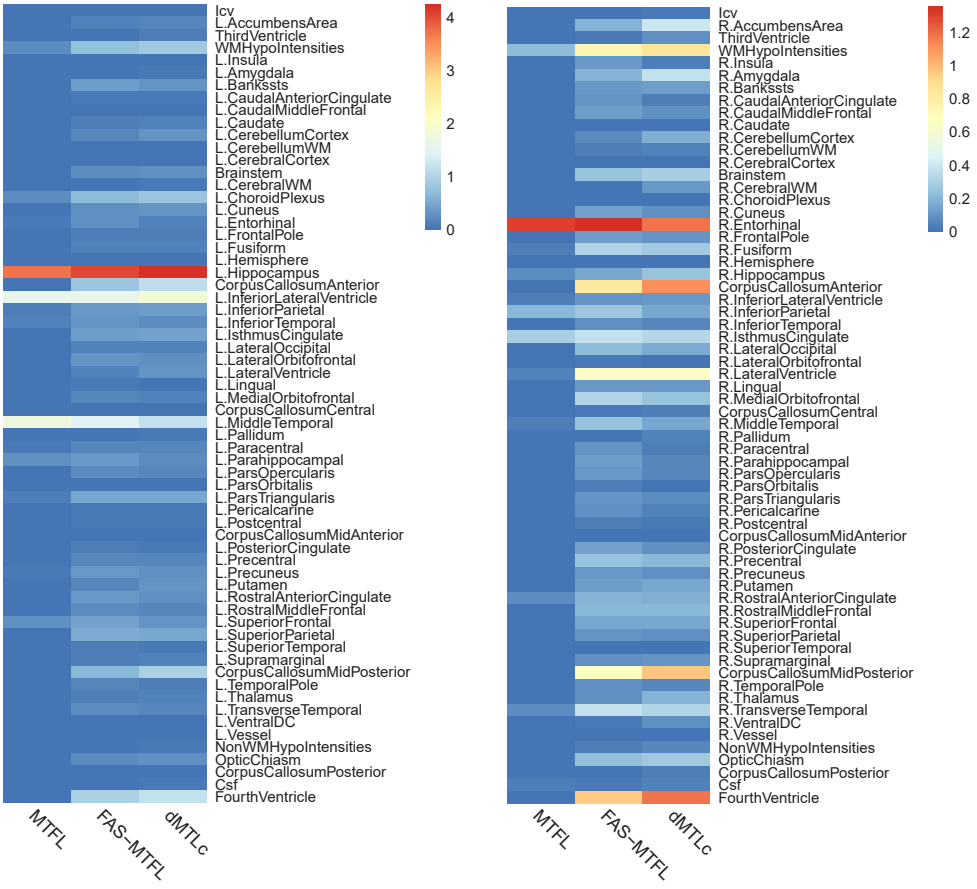
points:

1. The figure shows that the prediction performance of ADAS is better than that of RAVLT.TOTAL. A possible explanation is that the temporal patterns of these cognitive scores are different, which results in the different performances of one model on the different cognitive scores prediction.
2. It can be observed that the performances for predicting earlier time point scores are often better than those for later time point scores, and the performance at the last time point (M36) declines obviously for most models. Some authors have speculated that the lack of predictable biomarkers in later stages is a potential factor [25]. Another possible explanation for this is that the learning of the later prediction models is more difficult for the number of the available samples is reducing.
3. FAS-MTFL consistently outperforms MTFL at all tasks except M06 of the RAVLT.TOTAL, which demonstrates that integrating the implicit feature correlation by the FAS-norm penalty can overcome the limitation of MTFL and improve the prediction performance.
4. Although dMTLl shows the best overall performance at all time points in Table 10, it witnesses sub-optimal performances in predicting later time point scores (M24, M36) in Fig. 9 compared with FAS-MTFL. This inconsistency may be that it is hard for dMTLl to only incorporate the fused lasso penalty to identify task shared biomarkers, for lack of the predictable MRI biomarkers from the input data in later time points. On the other hand, the $\ell_{2,1}$-norm penalty tends to select a set of features over all time points, which helps the FAS-MTFL model to achieve good performance in the later time points. That is to say, learning the early time point scores is beneficial to the learning of the later time point scores.

5.2.5. *Analysis of biomarkers*

The identification of sensitive and stable biomarkers will help to diagnose and prognosis the disease. In this section, we analyze the biomarker patterns in both cross-sectional and longitudinal experiments. The features and ROIs are sorted by the weight calculated as follows. We first average the corresponding parameters in the repeated experiments and then denote the ℓ_2-norm of the feature's parameters as its weight. In order to eliminate the bias arising from the different number of features in ROIs, the ℓ_2-norm of ROIs' parameters are divided by $\sqrt{v_j}$, where v_j is the number of features in ROI j.

Biomarkers in the cross-sectional analysis. Fig. 10 are the heat maps of the weights of all ROIs in each brain hemisphere for MTFL, FAS-MTFL, and dMTLc in the eighteen cognitive score cross-sectional experiments. The larger the value of the weight, the more important its corresponding ROI is in predicting the corresponding cognitive score. First, the value range of weights of ROIs in the left hemisphere is bigger than that in the right hemisphere, which demonstrates that the left hemisphere is more important than the right hemisphere in AD. This observation has been verified in Ref. [65], in which W. Zhang et al. found that the left hemisphere was more severely affected than the right during the early disease stage. Second, it can be observed clearly that the left hippocampus [66,67], the left inferior lateral ventricle [68,69], the left middle temporal [70], and the right entorhinal [11,71] are



(a) The left hemisphere. (b) The right hemisphere.

Fig. 10. Heat maps of all ROIs in each brain hemisphere for each method. The vector of ROIs was estimated from models in the eighteen cognitive score cross-sectional experiments. MTFL: Multi-task feature learning; FAS-MTFL: Feature-aware sparse multi-task feature learning; dMTLc: Dual feature correlation guided multi-task feature learning. (a) The ROIs in the left hemisphere. (b) The ROIs in the right hemisphere.

Table 11
The important features in the eighteen cognitive score cross-sectional experiments.

MTFL		FAS-MTFL		dMTLc	
features name	weight	features name	weight	features name	weight
SV of L.Hippocampus	3.733	SV of L.Hippocampus	4.043	SV of L.Hippocampus	4.253
TA of L.MidTemporal	3.509	TA of L.MidTemporal	2.964	TA of L.MidTemporal	2.210
CV of R.Entorhinal	2.292	CV of R.Entorhinal	2.521	CV of R.Entorhinal	1.890
SV of L.LateralVentricle	1.587	SV of L.LateralVentricle	1.587	SV of L.LateralVentricle	1.868
TA of R.Entorhinal	1.306	TA of R.Entorhinal	1.263	TA of R.Entorhinal	1.433
TA of R.IsthmusCingulate	0.596	SA of L.SuperiorParietal	1.097	SV of FourthVentricle	1.187
TA of L.Parahippocampal	0.548	SV of FourthVentricle	0.961	SV of CCAnterior	1.121
TS of L.SuperiorFrontal	0.513	SA of L.ParsTriangularis	0.935	SV of CCMidPosterior	0.966
TA of R.InferiorParietal	0.427	TS of L.SuperiorFrontal	0.881	SV of WMHypoIntensities	0.865
TA of L.InferiorTemporal	0.268	SV of CCAnterior	0.830	SV of L.ChoroidPlexus	0.823
SV of L.ChoroidPlexus	0.244	TS of L.IsthmusCingulate	0.794	SA of L.SuperiorParietal	0.809
SV of WMHypoIntensities	0.225	TA of L.Precuneus	0.761	SA of L.ParsTriangularis	0.781
TS of L.ParsTriangularis	0.206	SA of L.RostralAnteriorCingulate	0.756	CV of L.MidTemporal	0.741
TA of L.InferiorParietal	0.183	TS of L.ParsTriangularis	0.750	TS of L.IsthmusCingulate	0.731
TS of L.Paracentral	0.166	SV of WMHypoIntensities	0.744	TS of R.Entorhinal	0.683

Table 12
The important ROIs in the eighteen cognitive score cross-sectional experiments.

MTFL		FAS-MTFL		dMTLc	
ROIs name	weight	ROIs name	weight	ROIs name	weight
L.Hippocampus	3.733	L.Hippocampus	4.043	L.Hippocampus	4.253
L.MidTemporal	1.754	L.LateralVentricle	1.587	L.LateralVentricle	1.868
L.LateralVentricle	1.587	L.MidTemporal	1.484	R.Entorhinal	1.192
R.Entorhinal	1.308	R.Entorhinal	1.354	FourthVentricle	1.187
R.IsthmusCingulate	0.298	FourthVentricle	0.961	L.MidTemporal	1.152
L.Parahippocampal	0.274	CCAnterior	0.830	CCAnterior	1.121
L.SuperiorFrontal	0.257	WMHypoIntensities	0.744	CCMidPosterior	0.966
L.ChoroidPlexus	0.244	L.ChoroidPlexus	0.699	WMHypoIntensities	0.865
WMHypoIntensities	0.225	CCMidPosterior	0.672	L.ChoroidPlexus	0.823
R.InferiorParietal	0.214	R.LateralVentricle	0.657	R.LateralVentricle	0.654
L.InferiorTemporal	0.134	L.SuperiorParietal	0.549	L.ParsTriangularis	0.490
L.ParsTriangularis	0.103	L.ParsTriangularis	0.507	L.SuperiorParietal	0.468
L.InferiorParietal	0.093	L.SuperiorFrontal	0.440	L.IsthmusCingulate	0.446
L.Paracentral	0.083	L.IsthmusCingulate	0.413	L.InferiorParietal	0.423
L.Precuneus	0.080	L.Bankssts	0.386	R.AccumbensArea	0.400

important ROIs for all models. These identified brain regions have been shown to be highly related to AD progression in the previous works. Finally, it can be seen that our proposed models (FAS-MTFL and dMTLc) identify more ROIs than MTFL. This indicates that our methods can identify the missing features of MTFL because of incorporating prior feature structure knowledge. The ROIs selected by different methods will be compared in detail in the following. The specific top fifteen features and top fifteen ROIs identified in the eighteen cognitive score cross-sectional experiments are listed in Tables 11 and 12, respectively.

From the identified important features in Table 11, we can derive several interesting observations:

1. The surface area of the left superior parietal, cortical thickness average of left precuneus, and surface area of left rostral anterior cingulate are identified by FAS-MTFL compared with the result of MTFL. Since FAS-MTFL incorporates the FAS-norm penalty to add prior feature structural knowledge, the weights of features that are strongly correlated to the important features ascend. In other words, these features are strongly correlated with other important features, which can be confirmed by the feature correlation matrix C. Speak specifically, the correlation coefficients between the surface area of the left superior parietal and hippocampus, the surface area of the left rostral anterior cingulate and hippocampus are high. We also witness strong correlation coefficients between cortical thickness average of the left precuneus and superior frontal, cortical thickness average of the left precuneus and inferential parietal, cortical thickness average of the left precuneus and middle temporal. Several studies [72,73] have found these features missed by MTFL are highly suggestive and effective for tracking the progression of AD.
2. Similar to FAS-MTFL, dMTLc has selected several features missed by MTFL, such as the subcortical volume of corpus callosum anterior, the subcortical volume of corpus callosum mid posterior, and the surface area of the left superior parietal. The reason is due to the incorporation of the FAS-norm penalty, which is the same as the one of FAS-MTFL. In previous studies [74–76], these regions are reported to be highly associated with the AD.
3. The number of features from the same brain region increases in dMTLc compared to FAS-MTFL and MTFL. For example, dMTLc selects three features from the right entorhinal (cortical volume of right entorhinal, cortical thickness average of right entorhinal, and standard deviation of thickness of right entorhinal) while FAS-MTFL and MTFL select two (cortical volume of right entorhinal and cortical thickness average of right entorhinal). This result verifies that the G_1-norm penalty tends to select or delete all features in every ROI.

Table 12 shows the important ROIs in the eighteen cognitive score cross-sectional experiments. It can be observed that the ranking of several ROIs in the three methods rises in turn, such as inferior lateral ventricle and entorhinal. A possible explanation is that adding feature correlation structure makes closely related important ROIs strengthen each other's importance. In addition, our methods select several ROIs that are not identified in MTFL, such as corpus callosum anterior, corpus callosum mid posterior, lateral ventricle, superior parietal, and isthmus cingulate. This may be because the changes of such ROIs are difficult to

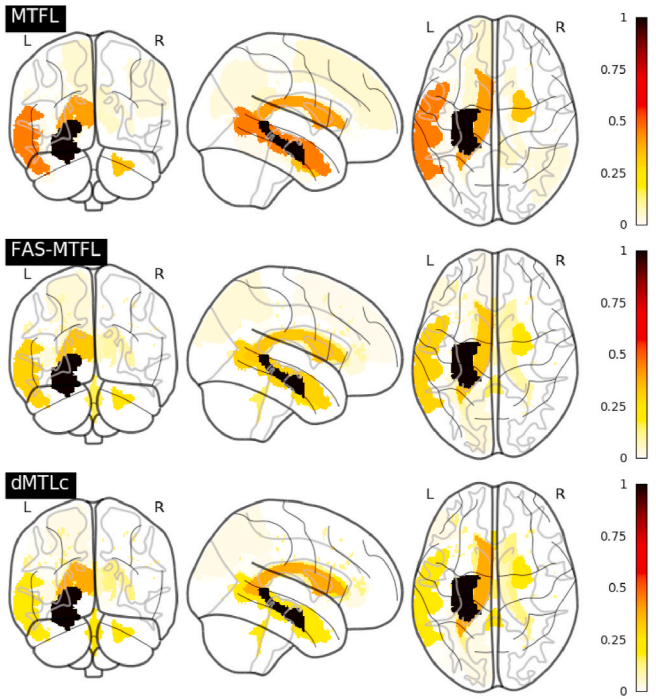


Fig. 11. The visualization of the identified important ROIs in the eighteen cognitive score cross-sectional experiments. The brain regions are segmented based on the Desikan-Killiany atlas. The name of the ROIs is listed in Table 12.

detect for MTFL that only considers the task correlation structure information. For instance, P. J. Wang et al. suggested that corpus callosum changes occur very early in the dementing process, and that these earliest changes may be too subtle for detection by neuropsychological tests [75].

The visualization of the identified important ROIs in Table 12 is shown in Fig. 11. Visually, the identified ROIs are mainly distributed in the left hemisphere of the brain. Some authors have found that the left hemisphere is more severely affected than the right during the early disease stage [65].

Biomarkers in the longitudinal analysis. For ease of exposition, we just plot the ROI magnitudes of ADAS and RAVLT.TOTAL for the longitudinal experiments, which are shown in Fig. 12. From the stable ROIs for ADAS shown in Fig. 12(a), several interesting observations can be derived: (1) some regions are important in all time points, such as the left hippocampus, left middle temporal, and right entorhinal; (2) some regions only work in the later time points, such as the left amygdala, and (3) some regions have strong weights during the first 2 years after baseline screening, such as left inferior temporal and left inferior parietal. These observations are consistent with the results in Ref. [25]. For example, Zhou et al. found that cortical thickness average of left middle temporal, cortical thickness average of left and right entorhinal, and white matter volume of left hippocampus are important biomarkers for all time points.

The longitudinal pattern of RAVLT.TOTAL is shown in Fig. 12(b), which is slightly different from that of ADAS. Specifically, some ROIs are selected in RAVLT.TOTAL only, such as corpus callosum mid posterior, right cerebral white matter, left and right precentral, right precuneus, and left fusiform. The different temporal patterns of biomarkers for these two scores suggest that just restricting several cognitive scores to share a common set of features may cause suboptimal performance. Meanwhile, it can be observed that most ROIs provide significant information in the first year, while few effective ROIs are available at the last time point. A possible explanation for these results may be the lack of adequate predictable MRI biomarkers in later stages [25].

5.2.6. Comparison with the state-of-the-art methods

A lot of works have studied the relationship between imaging markers (such as brain MRI) and cognitive scores using the ADNI dataset. Table 13 compares the result of our ADAS score prediction of the longitudinal analysis with the state-of-the-art works in terms of correlation coefficient (CC) as reported in the respective references. It can be observed that our method performs competitively. Specifically, our method achieves an average correlation coefficient of 0.670. Our method mainly faces the following challenges. First, compared with other methods, all available instances are used in this paper without sample selection. Poor quality data may bring challenges to model training. Second, compared with the methods that only predict one future time point score [27,77], it is more challenging to predict multiple scores in the future. In particular, we only use MRI data as the input compared with [27] which inputs multimodal data. The lack of available features poses a greater challenge to our method. Third, compared with [37,40,78], our method only uses the baseline MRI data to predict cognitive scores in more time points.

Despite the above challenges, our method shows the best performance. On the one hand, compared with the methods that only predict one future time point score [27,77], the power of the correlation among cognitive scores at multiple time points to promote extracting the underlying patterns from data has been well recognized. On the other hand, experiments demonstrate that the proposed dMTLl model performs better than the other methods that predict multiple cognitive scores simultaneously [37,40,78]. Although a previous study in Wang et al. [40] reported a similar correlation coefficient, a number of subjects with multiple time points data are used for training the models. In summary, the results demonstrate that it is beneficial to predict to simultaneously take both the feature correlation and the task correlation information into account.

6. Discussions

This paper proposes a generalized multi-task formulation framework and a novel feature-aware sparsity-inducing norm (FAS-norm) penalty with the view that the feature correlation structure can help multi-task learning to discover more stable biomarkers and achieve better prediction performance. This section discusses the broad applicability of the framework, the prediction performance, and the biomarker identification performance.

6.1. Application of the proposed framework on the other datasets

In this set of experiments, we conduct the multi-task classification on the dataset from the UCI data archive to evaluate the effectiveness of the proposed generalized multi-task formulation framework applied to the other structural data.

A multi-view dataset is used in our experiments, Mfeat.[4] The Mfeat dataset consists of features of handwritten numerals (‘0’-‘9′) extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2000 patterns) have been digitized in binary images. We separate ‘0’-‘4’ (the first dataset) and ‘5’-‘9’ (the second dataset) to form two experiments and choose five views to construct five feature groups, including Fourier coefficients of the character shapes, profile correlations, Karhunen-Love coefficients, pixel and Zernike moments (see Table 14). In each experiment, a classification model is set up for each label. We determine the performance of the classification task by calculating class-specific F_1 scores:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{32}$$

Especially, the larger value of F_1 indicates the better performance.

[4] https://archive.ics.uci.edu/ml/datasets/Multiple+Features.

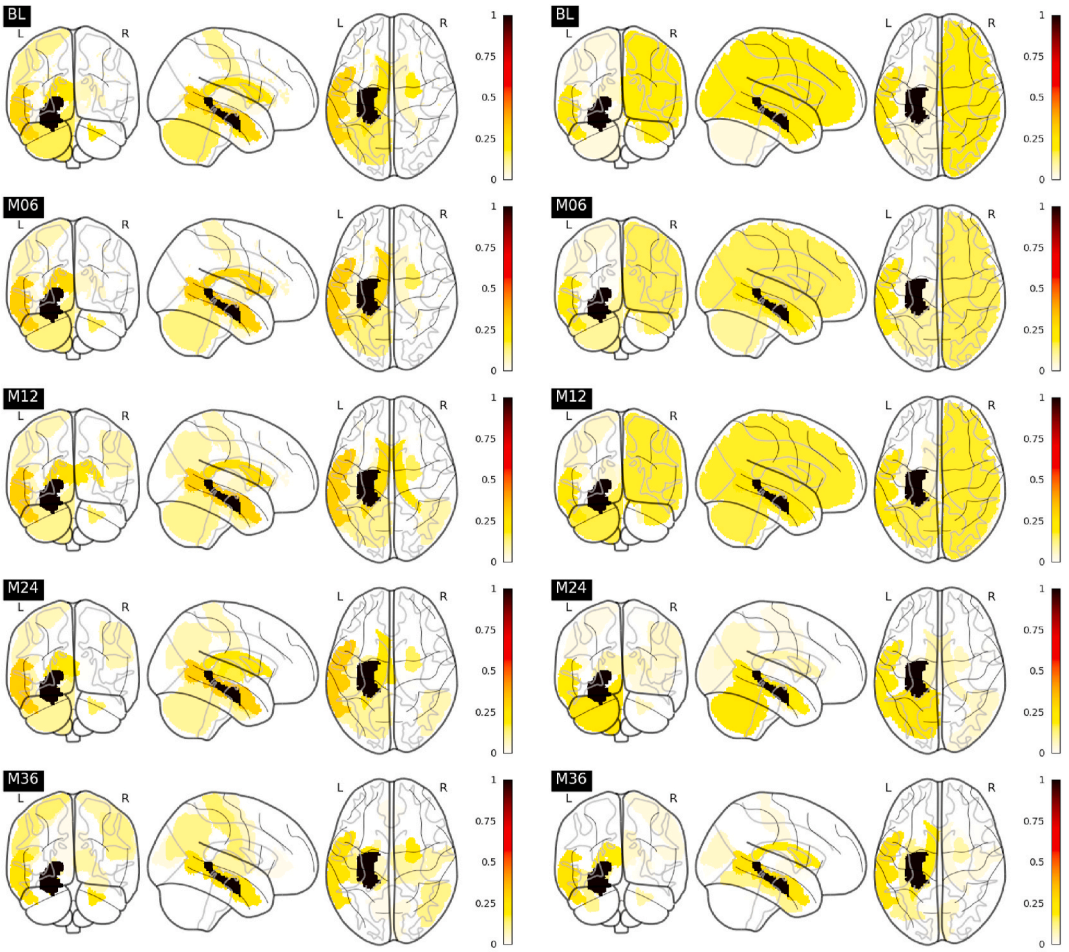


Fig. 12. The visualization of the important ROIs identified by our proposed dMTLl in longitudinal experiments for ADAS and RAVLT.TOTAL. The brain regions are segmented based on the Desikan-Killiany atlas. (a) ADAS, (b) RAVLT.TOTAL.

Table 13
Comparison with the state-of-the-art methods in terms of correlation coefficient (CC) in ADAS score prediction.

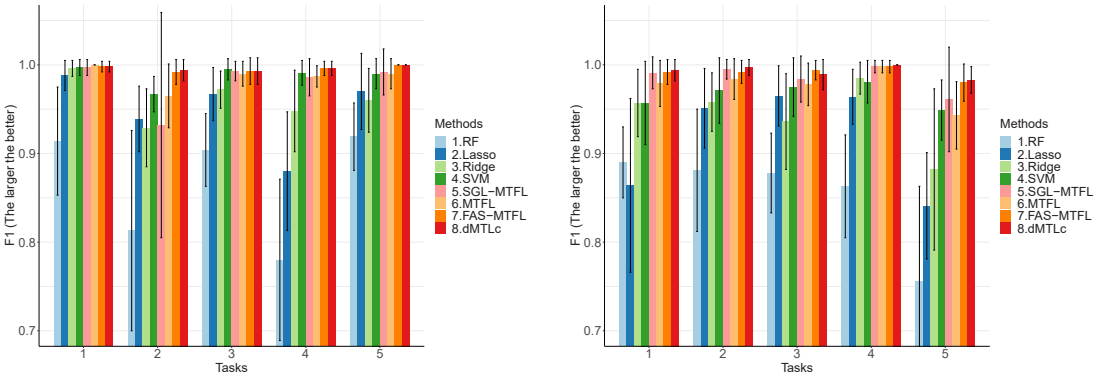
Method	Subjects			Features	Target (ADAS)	CC
	AD	MCI	NC			
Fan et al., 2010 [77]	52	148	64	BL (MRI)	M06	0.522
Zhang et al., 2012 [27]	45	91	50	BL (MRI, PET, CSF)	M24	0.531 ± 0.032
Jie et al., 2017 [37]	91	202	152	BL M06 M12 M24 (MRI)	BL M06 M12 M24	0.639 ± 0.008
Lei et al., 2019 [78]	91	202	152	BL M06 M12 M24 (MRI)	BL M06 M12 M24	0.655
Wang et al., 2019 [40]	91	202	152	BL M06 M12 M24 (MRI)	BL M06 M12 M24	0.664 ± 0.025
dMTLl (ours)	173	390	225	BL (MRI)	BL M06 M12 M24 M36	**0.670 ± 0.075**

Table 14
The constitutions of two datasets in the experiments.

Dataset	Samples	Feature groups	Features
The first dataset: '0'-'4'	1000	1. Fourier coefficients of the character shapes	76
		2. Profile correlations	216
		3. Karhunen-Love coefficients	64
		4. Pixel	240
		5. Zernike moments	47
The second dataset: '5'-'9'	1000	1. Fourier coefficients of the character shapes	76
		2. Profile correlations	216
		3. Karhunen-Love coefficients	64
		4. Pixel	240
		5. Zernike moments	47

We randomly split the data into training and testing sets using a ratio of 9:1, i.e., we build models on 90% of the data and evaluate these models on the remaining 10% of the data. In each of the ten trials, a 5-fold nested cross-validation procedure is employed to tune the regularization parameters. The range of each parameter varies from 0.1 to 1000. The reported results are the best results of each method with the optimal parameters.

Experimental results are shown in Fig. 13. Intuitively, the multi-task learning methods outperform the single-task learning methods. This justifies the motivation of learning correlated multiple tasks simultaneously and verifies that capturing their relatedness can improve learning performance. Meanwhile, the proposed multi-task learning models (FAS-MTFL and dMTLc) outperform MTFL overall, and our methods lead to the best prediction performance in most cases. This result demonstrates the importance of incorporating the feature correlation structure information in the training phase. Thanks to incorporating the feature structure information, the results of the restrictive assumption (only tasks are correlated) have been promoted. To sum up, in the proposed generalized multi-task formulation framework,



(a) The first dataset. (b) The second dataset.

Fig. 13. Performance on Mfeat data from the UCI data archive. There are five tasks in each experiment. For F_1, the larger the value, the better the model performance.

incorporating the feature structure in multi-task learning allows learning more stable and accurate patterns. Despite the fact that our framework is proposed for the AD research problem, it could be applied to other structural data.

6.2. Prediction performance

From Section 5, we can see that our methods consistently outperform MTFL. Specifically, FAS-MTFL achieves the best prediction performance with a 7.16% decrease in nMSE compared with MTFL for the five task cross-sectional experiments. dMTLc achieves the best prediction performance with a 1.78% decrease in nMSE compared with MTFL for the eighteen task cross-sectional experiments. Meanwhile, compared with MTFL, our methods achieve average overall error declines of 7.97%, 1.93%, 1.28%, and 4.82% in the longitudinal analysis for the ADAS, the RAVLT.TOTAL, the RAVLT.RECOG, and the FLU.ANIM respectively.

In addition, there are interesting observations: (1) FAS-MTFL achieves similar performance to SGL-MTFL, and SGL-MTFL achieves better prediction performance compared with dMTLc in the five task cross-sectional experiments; (2) The performances of dMTLc and FAS-MTFL are significantly better than that of SGL-MTFL on nMSE in the eighteen task cross-sectional experiments. These observations may have the following reasons. First, this is partially due to the limited data size of ADNI. The limited sample number hinders the model training, so it may produce some results that are not significant enough. Second, it may be that the gaps among the cognitive scores are rising along with the number of tasks rising. The feature correlation information becomes relatively more important in the eighteen task experiments compared with the five task experiments. Therefore adding feature correlation information into the models will improve the prediction performance. Third, the number of cognitive scores learned together will affect the common information shared among tasks. Adding suitable prior structural knowledge will improve the prediction performance of the model.

6.3. Biomarker identification performance

We incorporate the implicit feature correlation information in multi-task learning to identify the sensitive and stable biomarkers that provide the diagnostic indicators of AD. In Section 1, the cross-regional feature correlation was analyzed in Fig. 2. Further more, the important features and the important ROIs derived from experiments are listed in Tables 11 and 12 respectively. Comparing the calculated correlation in Fig. 2 and the identified biomarkers in Tables 11 and 12, the inconsistent and the consistent result can be seen. On the one hand, it can be observed that the left and right hemispheres are the most important ROIs from Fig. 2 because of the longest arcs of these two ROIs, but they are not selected in the experiments. This is because the left and right hemispheres are correlated with most ROIs in the anatomy but irrelevant to AD prediction. The results demonstrate that our generalized multi-task formulation framework and FAS-norm penalty could incorporate the effective feature correlation information but not redundant information. On the other hand, some ROIs that are only identified by our methods, such as the corpus callosum anterior, corpus callosum mid posterior, and lateral ventricle, are plotted in Fig. 2. This result confirms our initial hypothesis that incorporating feature correlated information will help us identify the stable and sensitive biomarkers which are difficult to be detected only using the task correlation information.

Comparing the important ROIs between the cross-sectional and the longitudinal experiments, we can derive several interesting observations. First, some biomarkers are only identified in the longitudinal experiments, such as parahippocampal and amygdala. This difference may be that the atrophy of the parahippocampal and amygdala is influenced by other longitudinal key biomarkers, therefore they are difficult to be observed in the cross-sectional pattern. Galton et al. proposed that bilateral hippocampus atrophy with involvement of the amygdala bilaterally and the right parahippocampal gyrus [33], which confirms our explanation. The strong correlations between hippocampus with parahippocampal and amygdala can be also observed in the correlation matrix C. Second, some important biomarkers are only identified in the cross-sectional experiments, such as isthmus cingulate. McEvoy et al. pointed out that isthmus cingulate is effective in mild AD and MCI [79], and the imbalanced data in three-year longitudinal monitoring may result in missing isthmus cingulate in the longitudinal analysis. Finally, there are some important ROIs identified in both the cross-sectional and longitudinal experiments, such as the hippocampus, middle temporal, inferior lateral ventricle, and enternal.

To sum up, the identified biomarkers such as the hippocampus [80,81], middle temporal [70,82], lateral ventricle [83,84], and corpus callosum [74,75] are highly suggestive and relevant to the cognitive impairment.

7. Conclusion

This paper has studied multi-task learning methods to predict cognitive outcomes and identify biomarkers in Alzheimer's disease. In order to achieve a better disease prediction outcome, we develop a framework for multi-task learning simultaneously considering the task and feature correlation structures, and a novel FAS-norm penalty that can flexibly integrate the feature correlation information is proposed. To solve the proposed models, we develop an algorithm based on the ADMM. We conduct extensive experiments on both the synthetic datasets and the real-world datasets to verify the effectiveness of our models (FAS-MTFL, dMTLc, and dMTLl), which demonstrate their superior performances compared with the state-of-the-art and baselines.

Specifically, our methods achieve an average overall error decline of 4.28% in the cross-sectional experiments and an average overall error decline of 7.97% in the ADAS longitudinal experiments compared with MTFL. According to the biomarker analysis, our methods could identify the different patterns between the cross-sectional and longitudinal analysis. Important ROIs such as the hippocampus, middle temporal, inferior lateral ventricle, and amygdala are highly suggestive and effective for AD prediction, where the amygdala is specific for the longitudinal analysis. For future work, we plan to investigate other types of correlation calculation (such as inverse covariance matrix). Moreover, we are interested in optimizing the feature correlation during multi-task learning rather than doing a prior calculation.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "Dual feature correlation guided multi-task learning for Alzheimer's Disease prediction".

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62 076 059); the Fundamental Research Funds for the Central Universities (No. N2016001); the NSFC Major International (Regional) Joint Research Project Grant No. 71 620 107 003; the Liaoning Revitalizing Talent Program No. XLYC1802115; the Fundamental Research Funds for State Key Laboratory of Synthetical Automation for Process Industries Grant No. 2013ZCX11; the 111 Incubating Program of Overseas Expert Introduction (BC2018010); the "High-level Overseas Expert" Introduction Program (G20190006026).

References

[1] W.H. Organization, et al., Risk reduction of cognitive decline and dementia: who guidelines, in: Risk Reduction of Cognitive Decline and Dementia, WHO guidelines, 2019.
[2] J. Jia, C. Wei, S. Chen, F. Li, Y. Tang, W. Qin, L. Zhao, H. Jin, H. Xu, F. Wang, et al., The cost of alzheimer's disease in China and re-estimation of costs worldwide, Alzheimer's Dementia 14 (4) (2018) 483–491.
[3] Y. Yang, X. Li, P. Wang, Y. Xia, Q. Ye, Multi-source transfer learning via ensemble approach for initial diagnosis of alzheimer's disease, IEEE J. Transl. Eng. Health Med. 8 (2020) 1–10.
[4] M.S. Albert, M.B. Moss, R. Tanzi, K. Jones, Preclinical prediction of ad using neuropsychological tests, J. Int. Neuropsychol. Soc.: JINS 7 (5) (2001) 631.
[5] W.G. Rosen, R.C. Mohs, K.L. Davis, A new rating scale for alzheimer's disease, Am. J. Psychiatr. 141 (11) (1984) 1356–1364.
[6] M.F. Folstein, S.E. Folstein, P.R. McHugh, mini-mental state: a practical method for grading the cognitive state of patients for the clinician, J. Psychiatr. Res. 12 (3) (1975) 189–198.
[7] M. Schmidt, et al., Rey Auditory Verbal Learning Test: A Handbook, Western Psychological Services Los, Angeles, CA, 1996.
[8] A.L. Chin, S. Negash, S. Xie, S.E. Arnold, R. Hamilton, Quality, and not just quantity, of education accounts for differences in psychometric performance between african americans and white non-hispanics with alzheimer's disease, J. Int. Neuropsychol. Soc. 18 (2) (2012) 277–285.
[9] B. Dickerson, T. Stoub, R. Shah, R. Sperling, R. Killiany, M. Albert, B. Hyman, D. Blacker, L. Detoledo-Morrell, Alzheimer-signature mri biomarker predicts ad dementia in cognitively normal adults, Neurology 76 (16) (2011) 1395–1402.
[10] C. Pettigrew, A. Soldan, Y. Zhu, M.-C. Wang, A. Moghekar, T. Brown, M. Miller, M. Albert, B.R. Team, et al., Cortical thickness in relation to clinical symptom onset in preclinical ad, Neuroimage: Clin. 12 (2016) 116–122.
[11] L. Velayudhan, P. Proitsi, E. Westman, J. Muehlboeck, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, C. Spenger, et al., Entorhinal cortex thickness predicts cognitive decline in alzheimer's disease, J. Alzheim. Dis. 33 (3) (2013) 755–766.
[12] B. Schmand, P. Eikelenboom, W.A. Van Gool, A.D.N. Initiative, Value of neuropsychological tests, neuroimaging, and biomarkers for diagnosing alzheimer's disease in younger and older age cohorts, J. Am. Geriatr. Soc. 59 (9) (2011) 1705–1710.
[13] R.S. Desikan, H.J. Cabral, F. Settecase, C.P. Hess, W.P. Dillon, C.M. Glastonbury, M. W. Weiner, N.J. Schmansky, D.H. Salat, B. Fischl, et al., Automated mri measures predict progression to alzheimer's disease, Neurobiol. Aging 31 (8) (2010) 1364–1374.
[14] I.Á. Illán, J. Górriz, J. Ramírez, D. Salas-Gonzalez, M. López, F. Segovia, P. Padilla, C.G. Puntonet, Projecting independent components of spect images for computer aided diagnosis of alzheimer's disease, Pattern Recogn. Lett. 31 (11) (2010) 1342–1347.
[15] I. Illán, J. Górriz, M. López, J. Ramírez, D. Salas-Gonzalez, F. Segovia, R. Chaves, C. G. Puntonet, Computer aided diagnosis of alzheimer's disease using component based svm, Appl. Soft Comput. 11 (2) (2011) 2376–2382.
[16] A. Ortiz, J.M. Górriz, J. Ramírez, F.J. Martinez-Murcia, A.D.N. Initiative, et al., Automatic roi selection in structural brain mri using som 3d projection, PLoS One 9 (4) (2014), e93851.
[17] Y. Wang, Y. Fan, P. Bhatt, C. Davatzikos, High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables, Neuroimage 50 (4) (2010) 1519–1535.
[18] L. Ferrarini, W.M. Palm, H. Olofsen, R. van der Landen, G.J. Blauw, R. G. Westendorp, E.L. Bollen, H.A. Middelkoop, J.H. Reiber, M.A. van Buchem, et al., Mmse scores correlate with local ventricular enlargement in the spectrum from cognitively normal to alzheimer disease, Neuroimage 39 (4) (2008) 1832–1838.
[19] S. Duchesne, A. Caroli, C. Geroldi, D.L. Collins, G.B. Frisoni, Relating one-year cognitive change in mild cognitive impairment to baseline mri features, Neuroimage 47 (4) (2009) 1363–1370.
[20] J. Zhou, Z. Lu, J. Sun, L. Yuan, F. Wang, J. Ye, Feafiner: biomarker identification from medical data through feature generalization and selection, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1034–1042.
[21] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. B 58 (1) (1996) 267–288.
[22] J. Ye, M. Farnum, E. Yang, R. Verbeeck, V. Lobanov, N. Raghavan, G. Novak, A. DiBernardo, V.A. Narayan, A.D.N. Initiative, et al., Sparse learning and stability selection for predicting mci to ad conversion using baseline adni data, BMC Neurol. 12 (1) (2012) 46.
[23] F. Bunea, Y. She, H. Ombao, A. Gongvatana, K. Devlin, R. Cohen, Penalized least squares regression methods and applications to neuroimaging, Neuroimage 55 (4) (2011) 1519–1527.
[24] J. Zhou, L. Yuan, J. Liu, J. Ye, A multi-task learning formulation for predicting disease progression, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 814–822.
[25] J. Zhou, J. Liu, V.A. Narayan, J. Ye, A.D.N. Initiative, et al., Modeling disease progression via multi-task learning, Neuroimage 78 (2013) 233–248.
[26] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient l2, 1-norm minimization, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 339–348.
[27] 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.
[28] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A.J. Saykin, L. Shen, Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 557–562.
[29] J. Ye, J. Liu, Sparse methods for biomedical data, ACM Sigkdd Explor. Newsl. 14 (1) (2012) 4–15.
[30] 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 l2, 1 norm, Neurobiol. Aging 36 (2015) S185–S193.
[31] X. Liu, P. Cao, D. Zhao, O. Zaiane, et al., Group guided sparse group lasso multi-task learning for cognitive performance prediction of alzheimer's disease, in: International Conference on Brain Informatics, Springer, 2017, pp. 202–212.
[32] X. Liu, A.R. Goncalves, P. Cao, D. Zhao, A. Banerjee, A.D.N. Initiative, et al., Modeling alzheimer's disease cognitive scores using multi-task sparse group lasso, Comput. Med. Imag. Graph. 66 (2018) 100–114.
[33] C.J. Galton, K. Patterson, K. Graham, M.A. Lambon-Ralph, G. Williams, N. Antoun, B. Sahakian, J. Hodges, Differing patterns of temporal atrophy in alzheimer's disease and semantic dementia, Neurology 57 (2) (2001) 216–225.
[34] X. Chen, X. Shi, X. Xu, Z. Wang, R. Mills, C. Lee, J. Xu, A two-graph guided multi-task lasso approach for eqtl mapping, in: Artificial Intelligence and Statistics, PMLR, 2012, pp. 208–217.
[35] 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: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2012, pp. 940–947.
[36] D. Zhang, J. Liu, D. Shen, Temporally-constrained group sparse learning for longitudinal data analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2012, pp. 264–271.
[37] B. Jie, M. Liu, J. Liu, D. Zhang, D. Shen, Temporally constrained group sparse learning for longitudinal data analysis in alzheimer's disease, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 64 (1) (2016) 238–249.
[38] X. Liu, P. Cao, A.R. Gonçalves, D. Zhao, A. Banerjee, Modeling alzheimer's disease progression with fused laplacian sparse group lasso, ACM Trans. Knowl. Discov. Data 12 (6) (2018) 1–35.
[39] 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: Advances in Neural Information Processing Systems, 2012, pp. 1277–1285.

[40] M. Wang, D. Zhang, D. Shen, M. Liu, Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data, Med. Image Anal. 53 (2019) 111–122.
[41] 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. Imag. 33 (7) (2014) 1475–1487.
[42] L. Brand, K. Nichols, H. Wang, L. Shen, H. Huang, Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction, IEEE Trans. Med. Imag. 39 (6) (2019) 1845–1855.
[43] L. Sun, R. Patel, J. Liu, K. Chen, T. Wu, J. Li, E. Reiman, J. Ye, Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 1335–1343, https://doi.org/10.1145/1557019.1557162.
[44] J. Zhou, J. Liu, V.A. Narayan, J. Ye, Modeling disease progression via fused sparse group lasso, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1095–1103.
[45] J. Chen, J. Zhou, J. Ye, Integrating low-rank and group-sparse structures for robust multi-task learning, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 42–50.
[46] X. Liu, P. Cao, J. Wang, J. Kong, D. Zhao, Fused group lasso regularized multi-task feature learning and its application to the cognitive performance prediction of alzheimer's disease, Neuroinformatics 17 (2) (2019) 271–294.
[47] H. Wang, F. Nie, H. Huang, Multi-view clustering and feature learning via structured sparsity, in: International Conference on Machine Learning, 2013, pp. 352–360.
[48] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imag. Sci. 2 (1) (2009) 183–202.
[49] L. Yuan, J. Liu, J. Ye, Efficient methods for overlapping group lasso, IEEE Trans. Pattern Anal. Mach. Intell. 35 (9) (2013) 2104–2116.
[50] A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, Mach. Learn. 73 (3) (2008) 243–272.
[51] C.M. Stonnington, C. Chu, S. Klöppel, C.R. Jack Jr., J. Ashburner, R.S. Frackowiak, A.D.N. Initiative, et al., Predicting clinical scores from magnetic resonance scans in alzheimer's disease, Neuroimage 51 (4) (2010) 1405–1413.
[52] 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) 968–980..
[53] M. Reuter, H.D. Rosas, B. Fischl, Highly accurate inverse consistent registration: a robust approach, Neuroimage 53 (4) (2010) 1181–1196.
[54] F. Ségonne, A.M. Dale, E. Busa, M. Glessner, D. Salat, H.K. Hahn, B. Fischl, A hybrid approach to the skull stripping problem in mri, Neuroimage 22 (3) (2004) 1060–1075.
[55] B. Fischl, D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, et al., Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain, Neuron 33 (3) (2002) 341–355.
[56] B. Fischl, D.H. Salat, A.J. Van Der Kouwe, N. Makris, F. Ségonne, B.T. Quinn, A. M. Dale, Sequence-independent segmentation of magnetic resonance images, Neuroimage 23 (2004) S69–S84.
[57] J.G. Sled, A.P. Zijdenbos, A.C. Evans, A nonparametric method for automatic correction of intensity nonuniformity in mri data, IEEE Trans. Med. Imag. 17 (1) (1998) 87–97.
[58] B. Fischl, A. Liu, A.M. Dale, Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex, IEEE Trans. Med. Imag. 20 (1) (2001) 70–80.
[59] F. Ségonne, J. Pacheco, B. Fischl, Geometrically accurate topology-correction of cortical surfaces using nonseparating loops, IEEE Trans. Med. Imag. 26 (4) (2007) 518–529.
[60] A.M. Dale, B. Fischl, M.I. Sereno, Cortical surface-based analysis: I. segmentation and surface reconstruction, Neuroimage 9 (2) (1999) 179–194.
[61] A.M. Dale, M.I. Sereno, Improved localizadon of cortical activity by combining eeg and meg with mri cortical surface reconstruction: a linear approach, J. Cognit. Neurosci. 5 (2) (1993) 162–176.
[62] B. Fischl, A.M. Dale, Measuring the thickness of the human cerebral cortex from magnetic resonance images, Proc. Natl. Acad. Sci. Unit. States Am. 97 (20) (2000) 11050–11055.
[63] P. Gong, J. Ye, C. Zhang, Robust multi-task feature learning, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 895–903.
[64] S. Ji, J. Ye, An accelerated gradient method for trace norm minimization, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 457–464.
[65] W. Zhang, J. Shi, C. Stonnington, R.J. Bauer, B.A. Gutman, K. Chen, P. M. Thompson, E.M. Reiman, R.J. Caselli, Y. Wang, Morphometric analysis of hippocampus and lateral ventricle reveals regional difference between cognitively stable and declining persons, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 14–18.
[66] X. Zhu, H.-I. Suk, S.-W. Lee, D. Shen, Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 63 (3) (2015) 607–618.
[67] H. Braak, E. Braak, On areas of transition between entorhinal allocortex and temporal isocortex in the human brain. normal morphology and lamina-specific pathology in alzheimer's disease, Acta Neuropathol. 68 (4) (1985) 325–332.
[68] M. Kayalvizhi, G. Kavitha, C. Sujatha, Analysis of ventricle regions in alzheimer's brain mr images using level set based methods, Int. J. Biomed. Eng. Technol. 12 (3) (2013) 300–319.
[69] K. Anandh, C. Sujatha, S. Ramakrishnan, A method to differentiate mild cognitive impairment and alzheimer in mr images using eigen value descriptors, J. Med. Syst. 40 (1) (2016) 25.
[70] V.E. Sturm, J.S. Yokoyama, W.W. Seeley, J.H. Kramer, B.L. Miller, K.P. Rankin, Heightened emotional contagion in mild cognitive impairment and alzheimer's disease is associated with temporal lobe degeneration, Proc. Natl. Acad. Sci. Unit. States Am. 110 (24) (2013) 9944–9949.
[71] G.W. Van Hoesen, B.T. Hyman, A.R. Damasio, Entorhinal cortex pathology in alzheimer's disease, Hippocampus 1 (1) (1991) 1–8.
[72] P. Prawiroharjo, K.-i. Yamashita, K. Yamashita, O. Togao, A. Hiwatashi, R. Yamasaki, J.-i. Kira, Disconnection of the right superior parietal lobule from the precuneus is associated with memory impairment in oldest-old alzheimer's disease patients, Heliyon 6 (7) (2020), e04516.
[73] G. Koch, S. Bonnì, M.C. Pellicciari, E.P. Casula, M. Mancini, R. Esposito, V. Ponzo, S. Picazio, F. Di Lorenzo, L. Serra, et al., Transcranial magnetic stimulation of the precuneus enhances memory and neural activity in prodromal alzheimer's disease, Neuroimage 169 (2018) 302–311.
[74] A.H. Bachman, S.H. Lee, J.J. Sidtis, B.A. Ardekani, Corpus callosum shape and size changes in early alzheimer's disease: a longitudinal mri study using the oasis brain database, J. Alzheim. Dis. 39 (1) (2014) 71–78.
[75] P.J. Wang, A.J. Saykin, L.A. Flashman, H.A. Wishart, L.A. Rabin, R.B. Santulli, T. L. McHugh, J.W. MacDonald, A.C. Mamourian, Regionally specific atrophy of the corpus callosum in ad, mci and cognitive complaints, Neurobiol. Aging 27 (11) (2006) 1613–1617.
[76] D.M. Wolpert, S.J. Goodbody, M. Husain, Maintaining internal representations: the role of the human superior parietal lobe, Nat. Neurosci. 1 (6) (1998) 529–533.
[77] Y. Fan, D. Kaufer, D. Shen, Joint estimation of multiple clinical variables of neurological diseases from imaging patterns, in: 2010 IEEE International Symposium on Biomedical Imaging: from Nano to Macro, IEEE, 2010, pp. 852–855.
[78] B. Lei, W. Hou, W. Zou, X. Li, C. Zhang, T. Wang, Longitudinal score prediction for alzheimer's disease based on ensemble correntropy and spatial–temporal constraint, Brain Imag. Behav. 13 (1) (2019) 126–137.
[79] L.K. McEvoy, C. Fennema-Notestine, J.C. Roddey, D.J. Hagler Jr., D. Holland, D. S. Karow, C.J. Pung, J.B. Brewer, A.M. Dale, Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment, Radiology 251 (1) (2009) 195–205.
[80] R. Killiany, B. Hyman, T. Gomez-Isla, M. Moss, R. Kikinis, F. Jolesz, R. Tanzi, K. Jones, M. Albert, Mri measures of entorhinal cortex vs hippocampus in preclinical ad, Neurology 58 (8) (2002) 1188–1196.
[81] D. Devanand, G. Pradhaban, X. Liu, A. Khandji, S. De Santi, S. Segal, H. Rusinek, G. Pelton, L. Honig, R. Mayeux, et al., Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of alzheimer disease, Neurology 68 (11) (2007) 828–836.
[82] L. Mah, M.A. Binns, D.C. Steffens, A.D.N. Initiative, et al., Anxiety symptoms in amnestic mild cognitive impairment are associated with medial temporal atrophy and predict conversion to alzheimer disease, Am. J. Geriatr. Psychiatr. 23 (5) (2015) 466–476.
[83] K. Anandh, C. Sujatha, S. Ramakrishnan, Segmentation of ventricles in alzheimer mr images using anisotropic diffusion filtering and level set method, Biomed. Sci. Instrum. 50 (2014) 307.
[84] T. Ertekin, N. Acer, E. Köseoğlu, G. Zararsız, A. Sönmez, K. Gümüş, E. Kurtoğlu, Total intracranial and lateral ventricle volumes measurement in alzheimer's disease: a methodological study, J. Clin. Neurosci. 34 (2016) 133–139.