Neuroinformatics (2019) 17:271–294
https://doi.org/10.1007/s12021-018-9398-5

ORIGINAL ARTICLE

Fused Group Lasso Regularized Multi-Task Feature Learning and Its Application to the Cognitive Performance Prediction of Alzheimer's Disease

Xiaoli Liu[1,2] · **Peng Cao**[1] · **Jianzhong Wang**[3] · **Jun Kong**[3,4] · **Dazhe Zhao**[1,2]

Published online: 4 October 2018

Abstract

Alzheimer's disease (AD) is characterized by gradual neurodegeneration and loss of brain function, especially for memory during early stages. Regression analysis has been widely applied to AD research to relate clinical and biomarker data such as predicting cognitive outcomes from MRI measures. Recently, multi-task based feature learning (MTFL) methods with sparsity-inducing $\ell_{2,1}$-norm have been widely studied to select a discriminative feature subset from MRI features by incorporating inherent correlations among multiple clinical cognitive measures. However, existing MTFL assumes the correlation among all tasks is uniform, and the task relatedness is modeled by encouraging a common subset of features via sparsity-inducing regularizations that neglect the inherent structure of tasks and MRI features. To address this issue, we proposed a fused group lasso regularization to model the underlying structures, involving 1) a graph structure within tasks and 2) a group structure among the image features. To this end, we present a multi-task feature learning framework with a mixed norm of fused group lasso and $\ell_{2,1}$-norm to model these more flexible structures. For optimization, we employed the alternating direction method of multipliers (ADMM) to efficiently solve the proposed non-smooth formulation. We evaluated the performance of the proposed method using the Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets. The experimental results demonstrate that incorporating the two prior structures with fused group lasso norm into the multi-task feature learning can improve prediction performance over several competing methods, with estimated correlations of cognitive functions and identification of cognition-relevant imaging markers that are clinically and biologically meaningful.

Keywords Alzheimer's disease · Multi-task learning · Sparse group lasso · Fused lasso

Introduction

Alzheimer's disease (AD) is the most common cause of dementia, which mainly affects memory function, and progress ultimately culminate in a state of dementia where all cognitive functions are affected. The disease poses a serious challenge to the aging society. The worldwide prevalence of AD is predicted to quadruple from 46.8 million in 2016 (Alzheimer's Association and et al. 2016) to 131.5 million by 2050 according to ADI's World Alzheimer Report (Batsch and Mittelman 2015). Dementia also has a huge economic impact. Today, the total estimated worldwide cost of dementia of US is $818 billion, and it will achieve a trillion dollar disease by 2018.

Predicting cognitive performance of subjects from neuroimaging measures and identifying relevant imaging biomarkers are important focuses of the study of Alzheimer's disease. Magnetic resonance imaging (MRI) allows the direct observation of brain changes such as cerebral atrophy or ventricular expansion (Castellani et al. 2010). Previous work showed that brain atrophy detected by MRI is correlated with neuropsychological deficits (Frisoni et al. 2010). The relationships between commonly used cognitive measures and structural changes detected by MRI have been

✉ Peng Cao
caopeng@mail.neu.edu.cn

Xiaoli Liu
neuxiaoliliu@gmail.com

[1] Computer Science and Engineering, Northeastern University, Shenyang, China

[2] Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang, China

[3] College of Information Science and Technology, Northeast Normal University, Changchun, China

[4] Key Laboratory of Applied Statistics of MOE, Changchun, China

previously studied using regression models. Many analyses have demonstrated a relationship between baseline MRI features and cognitive measures. The most commonly used cognitive measures are the Alzheimer's Disease Assessment Scale cognitive total score (ADAS), the Mini Mental State Exam score (MMSE), and the Rey Auditory Verbal Learning Test (RAVLT). ADAS is the gold standard in AD drug trials for cognitive function assessment, and is the most popular cognitive testing instrument to measure the severity of the most important symptoms of AD. MMSE measures cognitive impairment, including orientation to time and place, attention and calculation, immediate and delayed recall of words, language and visuo-constructional functions. RAVLT measures episodic memory and is used for the diagnosis of memory disturbances, including eight recall trials and a recognition test.

Early studies focused on traditional regression models to predict cognitive outcomes one at a time. To achieve more appropriate predictive models of performance and identify relevant imaging biomarkers, many previous works formulated the prediction of multiple cognitive outcomes as a multi-task learning problem and developed regularized multi-task learning methods to model disease cognitive outcomes (Wan et al. 2012, 2014; Zhou et al. 2013; Wang et al. 2012). Multi-task learning (MTL) (Caruana 1998) describes a learning paradigm that seeks to improve the generalization performance of a learning task with the help of other related tasks. The fundamental hypothesis of the MTL methods is to assume that if tasks are related then learning of one task can benefit from the learning of other tasks. Learning multiple related tasks simultaneously has been theoretically and empirically shown to significantly improve performance. The key of MTL is how to exploit correlation among tasks via an appropriate shared representation. Two popular shared representations to model task relatedness are model parameter sharing (Argyriou et al. 2008; Jebara 2011) and feature representation sharing (Evgeniou and learning 2004; Yu et al. 2005; Xue et al. 2007). There are inherent correlations among different cognitive scores. Therefore, the prediction of different types of cognitive scores can be modeled as an MTL formulation, and the tasks are related in the sense that they all share a small set of features, which is multi-task feature learning (MTFL) problem. To solve MTFL problems, regularization has been introduced to produce better performance than traditional solution using single-task learning. The most commonly used regularization is $\ell_{2,1}$-norm (Liu et al. 2009), which is employed to extract features that impact all or most clinical scores, since the assumption is that a given imaging marker can affect multiple cognitive scores and only a subset of brain regions (region-of-interest, ROI) are relevant. (Wang et al. 2011) and (Zhang and Yeung 2012a) employed multi-task feature learning strategies to select biomarkers that could predict multiple clinical scores. Specifically, (Wang et al. 2011) employed an ℓ_1-norm regularizer to impose sparsity among all elements and proposed the use of a combined $\ell_{2,1}$-norm and ℓ_1-norm regularizations to select features. (Zhang et al. 2012) proposed a multi-task learning with $\ell_{2,1}$-norm to select a common subset of relevant features for multiple variables from each modality.

A major limitation of existing MTFL methods is that complex relationships among imaging markers and among cognitive outcomes are often ignored. The correlating of multiple prediction models assumes that all tasks shared the same feature subset. This is not a realistic assumption, since it treats all cognitive outcomes (response) and MRI features (predictors) equally and neglects underlying correlations between the cognitive tasks and structure within MRI features. Specifically, 1) for the cognitive outcomes, each assessment typically yields multiple evaluation scores from a set of relevant cognitive tasks, and thus these scores are inherently correlated. An example would be the scores of TOTAL and TOT6 in the RAVLT. Different assessments can evaluate different cognitive functions, resulting in low correlation and preferring different brain regions. For example, the tasks in TRAILS aim to test a combination of visual, motor, and executive functions, while the set of RAVLT aims testing of verbal learning memory. It is reasonable to assume that correlations among tasks are not equal, and some tasks may be more closely related than others in assessment tests of cognitive outcomes. 2) On the other hand, for MRI data, many MRI features are interrelated and together reveal brain cognitive functions (Yan et al. 2015). In our data, multiple shape measures (volume, area, and thickness) from the same region provide a comprehensively quantitative evaluation of cortical atrophy, and tend to be selected together as joint predictors. Our previous study proposed a model in which prior knowledge guided multi-task feature learning model. Using group information to enforce intra-group similarity has been demonstrated to be an effective approach (Liu et al. 2017). Overall, it is important to explore and utilize such interrelated structures and select important and structurally correlated features together.

To address these model limitations, we designed a novel multi-task feature learning that models a common representation with respect to MRI features across tasks as well as the local task structure with respect to brain regions. Specifically, we designed novel mixed structured sparsity norms, called fused group lasso, to capture the underlying structures at the level of tasks and features. This regularizer is based on the natural assumption that if some tasks are correlated, they should have a small similar weight vector and similar selected brain regions. It penalizes differences between prediction models of highly correlated tasks, and encourages similarity in the selected features

of highly correlated tasks. To discover such dependent structures among the cognitive outcomes, we employed the Pearson correlation coefficient to uncover the interrelations among cognitive measures and estimated the correlation matrix of all tasks. In our work, all the cognitive measures (20 in total) in the ADNI dataset were used to exploit the relationship. To the best of our knowledge, our approach is the first work to analyze all cognitive measures in the ADNI dataset and their relationships. With estimated task correlation, we employ the idea of fused lasso to capture the dependence of response variables. At the same time, taking into account the group structure among predictors, prior group information is incorporated into the fused lasso norm, promoting intra-group similarity with group sparsity. By incorporated fused group lasso into the MTFL model, we can better understand the underlying associations of prediction tasks of cognitive measures, allowing more stable identification of cognition-relevant imaging markers. The resulting formulation is challenging to solve due to the use of non-smooth penalties including the fused group lasso and the $\ell_{2,1}$-norm. An effective ADMM algorithm is proposed to tackle the complex non-smoothness.

Through empirical evaluation and comparison with different baseline methods and recently developed MTL methods using data from ADNI, we illustrate that the proposed FGL-MTFL method outperforms other methods. Improvements are statistically significant for most scores (tasks). The results demonstrate that incorporation of the fused group lasso into the traditional MTFL formulation improves predictive performance relative to traditional machine learning methods. We discuss the most prominent ROIs and task correlations identified by FGL-MTFL. We found that the results corroborate previous studies in neuroscience. Our previous works also formulate prediction tasks with a multi-task learning scheme. The algorithm SMKMTL (Sparse multi-kernel based multi-task learning) in (Cao et al. 2017) exploits a nonlinear prediction model based on multi-kernel learning. However, it assumes that correlations among tasks are equal and the features are independent. Although the GSGL-MTL (Group-guided Sparse Group Lasso regularized multi-task learning) algorithm (Liu et al. 2017) exploits the group structure of features by incorporating a priori group information, it does not consider the complex relationships among cognitive outcomes.

The rest of the paper is organized as follows. In "Preliminary Methodology", we provide a description of the preliminary methodology: multi-task learning (MTL), $\ell_{2,1}$-norm, group lasso norm, and fused lasso norm. A detailed mathematical formulation and optimization of FGL-MTFL is provided in "Fused Group Lasso Regularized Multi-Task Feature Learning, FGL-MTFL". In "Experimental Results and Discussions", we present the experimental results and evaluate the performance of FGL-MTFL using data from the ADNI-1 dataset. The conclusions are presented in "Conclusion".

Preliminary Methodology

Multi-Task Learning

Consider a multi-task learning (MTL) setting with k tasks. Let p be the number of covariates, shared across all the tasks, and let n be the number of samples. Let $X \in \mathbb{R}^{n\times p}$ denote the matrix of covariates, $Y \in \mathbb{R}^{n\times k}$ be the matrix of responses with each row corresponding to a sample, and $\Theta \in \mathbb{R}^{p\times k}$ denote the parameter matrix, with column $\theta_{\cdot m} \in \mathbb{R}^p$ corresponding to task m, $m = 1, \ldots, k$, and row $\theta_{j\cdot} \in \mathbb{R}^k$ corresponding to feature j, $j = 1, \ldots, p$. The MTL problem can be constructed by estimating the parameters based on suitable regularized loss function. In order effectively to associate imaging markers and cognitive measures, the MTL model minimizes the following objective:

$$\min_{\Theta\in\mathbb{R}^{p\times k}} \; L(Y, X, \Theta) + \lambda R(\Theta) \,, \tag{1}$$

where $L(\cdot)$ denotes the loss function and $R(\cdot)$ is the regularizer. In the current context, we assume the loss to be the square loss, i.e.,

$$L(Y, X, \Theta) = \|Y - X\Theta\|_F^2 = \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i\Theta\|_2^2 \,, \tag{2}$$

where $\mathbf{y}_i \in \mathbb{R}^{1\times k}$, $\mathbf{x}_i \in \mathbb{R}^{1\times p}$ are the i-th rows of Y, X, respectively corresponding to the multi-task response and covariates for the i-th sample. We note that the MTL framework can be easily extended to other loss functions. Clearly, different choices of the penalty $R(\Theta)$ may present quite different multi-task methods. Using some prior knowledge, we then add penalty $R(\Theta)$ to encode relatedness among tasks.

$\ell_{2,1}$-norm

One appealing property of the $\ell_{2,1}$-norm regularization is that it encourages multiple predictors from different tasks to share similar parameter sparsity patterns. The MTFL model via the $\ell_{2,1}$-norm regularization considers

$$\|\Theta\|_{2,1} = \sum_{j=1}^{p} \|\theta_{j\cdot}\|_2 \,, \tag{3}$$

and is suitable for the simultaneous prioritization of sparsity over features for all tasks.

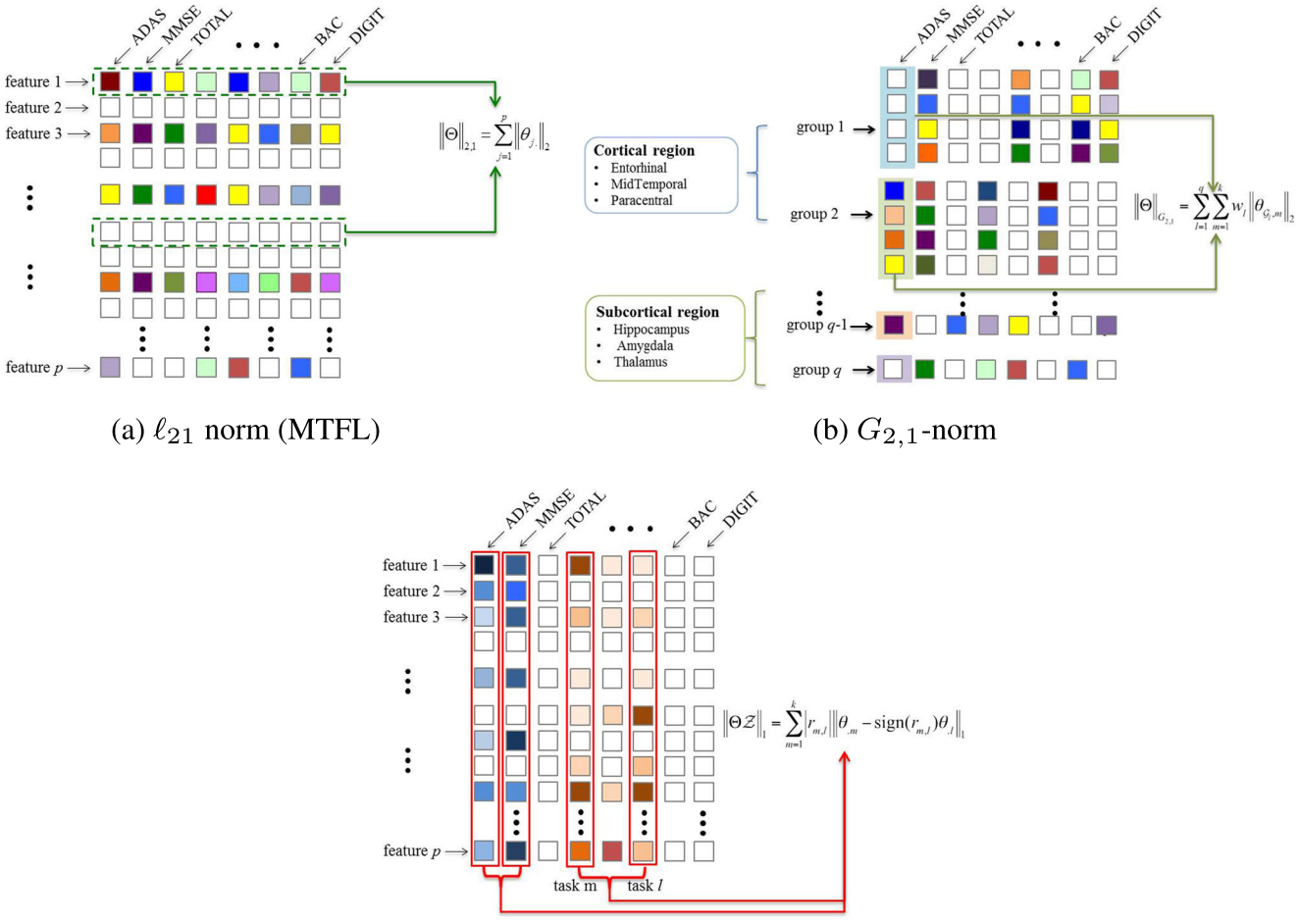
The key point of Eq. 3 is the use of ℓ_2-norm for $\theta_{j\cdot}$, which forces grouping of the weights corresponding to the j-th feature across multiple tasks and tends to select features

based on the joint strength of k tasks jointly (See Fig. 1a). There is a correlation among multiple cognitive tests. A relevant imaging predictor typically may have more or less influence on all these scores, and it is possible that only a subset of brain regions are relevant to each assessment. By employing MTFL, the correlation among different tasks can be incorporated into the model to build a more appropriate predictive model and identify a subset of features. The rows of Θ are equally treated in MTFL, which implies that the underlying structures among predictors are ignored.

$G_{2,1}$-norm

Despite the above achievements, few regression models take into account the covariance structure among predictors. To achieve a certain function, brain imaging measures are often correlated with each other. For MRI data, groups correspond to specific regions-of-interest (ROIs) in the brain, e.g., entorhinal and hippocampus. Individual features are the specific properties of those regions, e.g., cortical volume and thickness. In this study, for each region (group), multiple features were extracted to measure the atrophy information of each ROI involving cortical thickness, surface area, and volume from gray matter and white matter. Multiple shape measures from the same region provide a comprehensively quantitative evaluation of cortical atrophy, and tend to be selected together as joint predictors.

We assume the p covariates to be divided into g disjoint groups $\mathcal{G}_\ell, \ell = 1, \ldots, g$, with each group including ν_ℓ covariates respectively. In the context of AD, each group corresponds to a region-of-interest (ROI) in the brain, and the covariates in each group correspond to specific features of that region. For AD, the number of features in each group, ν_ℓ, is 1 or 4, and the number of groups g can be in



(a) ℓ_{21} norm (MTFL)

(b) $G_{2,1}$-norm

(c) Fused lasso

Fig. 1 The illustration of three different regularizations. Each column of Θ is corresponding to a single task and each row represents a feature dimension. The MRI features in each region belong to a group. We assume the p features to be divided into g disjoint groups $\mathcal{G}_\ell, \ell = 1, \ldots, g$, with each group having ν_ℓ features respectively. For each element in Θ, white color means zero-valued elements and color indicates non-zero values

the hundreds. We then introduce a $G_{2,1}$-norm according to the relationship between brain regions (ROIs) and cognitive tasks, and encourage a task-specific subset of ROIs (See Fig. 1b). The $G_{2,1}$-norm $\|\Theta\|_{G_{2,1}}$ is defined as:

$$\|\Theta\|_{G_{2,1}} = \sum_{\ell=1}^{g} \sum_{m=1}^{k} w_\ell \|\theta_{\mathcal{G}_\ell,m}\|_2 \,. \tag{4}$$

where $w_\ell = \sqrt{v_\ell}$ is the weight for each group and $\theta_{\mathcal{G}_\ell,m} \in \mathbb{R}^{v_\ell}$ is the coefficient vector for group $\mathcal{G}_\ell$ and task m.

Fused Lasso

Fused lasso is one of these variants, where pairwise differences between variables are penalized using the ℓ_1 norm, which results in successive variables being similar. The Fused lasso norm is defined as:

$$\|\mathcal{H}\Theta^{\mathrm{T}}\|_1 = \sum_{m=1}^{k-1} |\theta_{\cdot m} - \theta_{\cdot m+1}| \,, \tag{5}$$

where $\mathcal{H}$ is a $(k-1) \times k$ sparse matrix with $\mathcal{H}_{m,m} = 1$, and $\mathcal{H}_{m,m+1} = -1$.

It encourages $\theta_{\cdot m}$ and $\theta_{\cdot m+1}$ to take the same value by shrinking the difference between them toward zero. This approach has been employed to incorporate temporal smoothness to model disease progression. In longitudinal model, it is assumed that the difference of the cognitive scores between two successive time points is relatively small.

Fused Group Lasso Regularized Multi-Task Feature Learning, FGL-MTFL

Formulation

There are two limitations of traditional MTFL. When regularized by the $\ell_{2,1}$-norm in Eq. 3, sparsity is achieved by treating each task equally, which ignores the underlying structures among predictors. On the other hand, for some highly correlated features, the traditional MTFL tends to identify one and ignore the others, which was inadequate for yielding a biologically meaningful interpretation. Our previous work exploited relationships of features in the learning procedure, which motivated us to consider the underlying structure of these relationships. To address the limitations of MTFL with $\ell_{2,1}$-norm, we consider the structure of task and features instead of assuming that all tasks have similar weights and commonly selected features. More specifically, we propose a new regularization for multi-task feature learning, called the fused group lasso. The underlying idea is that if tasks are highly correlated according to the task interrelations, the tasks are supposed to share common brain regions, but tasks with low correlation are more likely to have different brain regions.

To model the task-feature relationships, we first construct a graph $\mathbb{G}$ to model the task correlation, as shown in Fig. 2 (left). Let $\mathbb{G} = (V, E)$ be an undirected graph where V represents the set of vertices and E is the set of edges. Each task is treated as a graph node, and each edge (pairwise link) $e_{m,l}(m,l) \in E$ in $\mathbb{G}$ corresponds to an edge from the m-th task to the l-th task, and $|r_{m,l}|$ encodes the strength of the relationship between the m-th task and the l-th task. In

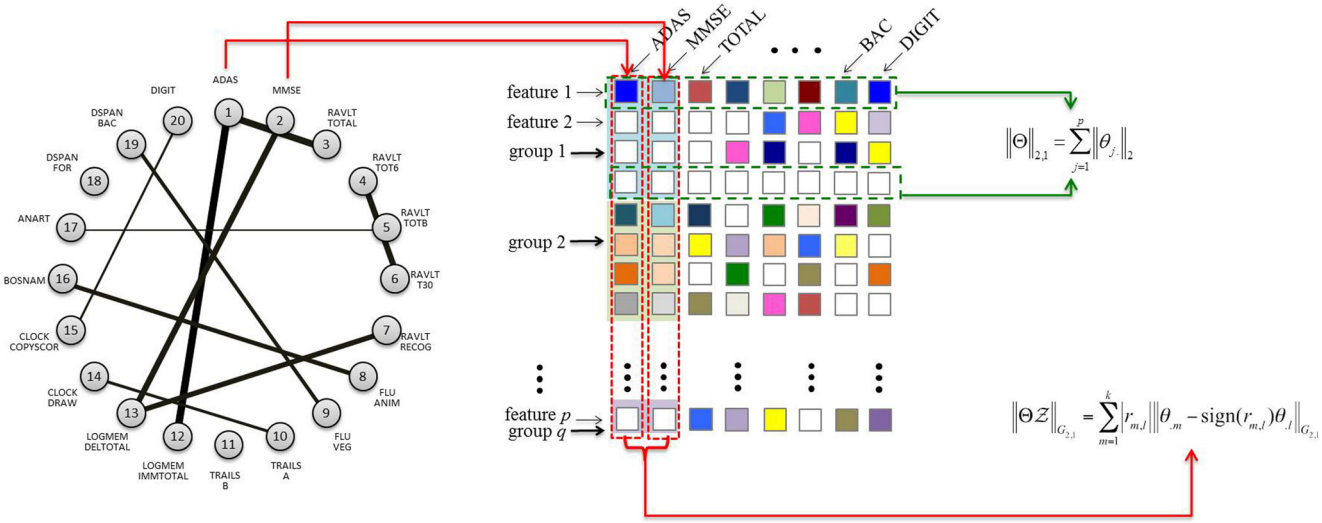


Fig. 2 The illustration of the FGL-MTFL method. The method involves two steps: 1) estimation of task correlation and construct an undirected graph (Left); and 2) joint learning of all regression models in a fused group lasso regularized multi-task feature learning (FGL-MTFL) formulation based on the estimated correlation matrix (Right)

this model, we adopt a simple and commonly-used approach to infer $\mathbb{G}$ from data. In this approach, we first compute pairwise Pearson correlation coefficients for each pair of tasks, and then connect two nodes with an edge only if their correlation coefficient is above a given threshold τ.

Brain imaging measures are often correlated with each other, so incorporation of the covariance of MRI features can improve the performance of traditional MTL methods. Thus, to identify biologically meaningful markers, we utilize prior knowledge of interrelated structure to group related features together to guide the learning process. The benefit of this strategy was also revealed in our previous work (Liu et al. 2017) , in which we proposed a group lasso multi-task learning algorithm that explicitly models the correlation structure within features, and achieved good performance in the prediction of cognitive scores from imaging measures.

Once the graph of task correlation $\mathbb{G}$ is constructed, we then integrate the group structure of features into the fused lasso norm and propose a new fused group lasso regularized multi-task feature learning (FGL-MTFL) model as follows.

$$\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_{2,1} + \lambda_2 \sum_{(m,l)\in E} |r_{m,l}| \|\theta_{.m} - \text{sign}(r_{m,l})\theta_{.l}\|_{G_{2,1}} , \tag{6}$$

where $|r_{m,l}|$ indicates the strength weight of the correlation between two tasks connected by an edge, λ_1 and λ_2 represent regularization parameters that determine the amount of two separate penalties. In the new fused group lasso norm, if the two tasks m and l are highly correlated, the difference between the two corresponding regression coefficients $\theta_{.m}$ and $\theta_{.l}$ will be penalized more than differences for other pairs of tasks with weaker correlation. The regularization tends to flatten the values of regression coefficients for each feature across multiple highly correlated tasks, so that the strength of the influence of each task becomes more similar across those tasks.

The FGL-MTFL model includes two regularization processes: (1) all tasks are regularized by the $\ell_{2,1}$-norm regularizer, which captures global relationships by encouraging multiple predictors across tasks to share similar parameter sparsity patterns. This ensures that a small subset of features will be selected for all regression models from a global perspective. (2) Two local structures (graph structures within tasks and group structures within features) are considered from a local perspective by the proposed fused group lasso (FGL) regularizer, which combines two structured sparsity regularizers (fused lasso and group lasso) to capture the specific task-ROI specific structures. This encourages the matching of related tasks to similar selected brain regions.

This new model not only preserves the strength of $\ell_{2,1}$-norm to require similarity across multiple scores from a cognitive test, but also considers the complex graph structure among responses and the interrelated group structure of imaging predictors. To the best of our knowledge, no sparsity-based algorithm has been described that includes both global and local task correlation for the prediction of cognitive outcomes in AD.

We used a symmetric correlation matrix $\mathcal{D} \in \mathbb{R}^{k\times k}$ to describe the correlation in the fusion penalty, which is defined as:

$$\mathcal{D}_{m,l} = \begin{cases} -r_{m,l} & (m,l) \in E, m \neq l \\ \sum_{\substack{m=1 \\ m\neq l}}^{k} |r_{m,l}| & (m,l) \in E, m = l \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

where $m, l = 1, 2, \ldots, k$. We normalize the matrix by κ, which is the number of edges in E, defined as $\mathcal{Z} = (1/\kappa)\mathcal{D}$. Then the formulation can be written into the following matrix form:

$$\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_{2,1} + \lambda_2\|\Theta\mathcal{Z}\|_{G_{2,1}} , \tag{8}$$

The framework of the proposed method is illustrated in Fig. 2.

Efficient Optimization for FGL-MTFL

ADMM

In recent years, ADMM has become popular since it is often easy to parallelize distributed convex problems. In ADMM, the solutions to small local subproblems are coordinated to identify the globally optimal solution (Boyd et al. 2011).

$$\min_{x,z} f(x) + g(z)$$
$$s.t.\ Ax + Bz = c.$$

The variant augmented Largrangian of ADMM method is formulated as follows:

$$L_\rho(x,z,u) = f(x) + g(z) + u^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2$$

where f and g are convex functions, and variables $A \in \mathbb{R}^{p\times n}, x \in \mathbb{R}^n, B \in \mathbb{R}^{p\times m}, z \in \mathbb{R}^m, c \in \mathbb{R}^p$. u is a scaled dual augmented Lagrangian multiplier, and ρ is a non-negative penalty parameter. In each iteration of ADMM, this problem is solved by alternating minimization $L_\rho(x,z,u)$

over $x, z,$ and u. At the $(k+1)$-th iteration, the update of ADMM is carried out by:

$$\begin{aligned} x^{k+1} &:= \arg\min_x L_\rho\left(x, z^k, u^k\right) \\ z^{k+1} &:= \arg\min_z L_\rho\left(x^{k+1}, z, u^k\right) \\ u^{k+1} &:= u^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right) \end{aligned}$$

Efficient Optimization for FGL-MTFL

We developed an efficient ADMM-based algorithm to solve the objective function in Eq. 8, which is equivalent to the following constrained optimization problem:

$$\begin{aligned} &\min_{\Theta, Q, S} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|Q\|_{2,1} + \lambda_2\|S\|_{G_{2,1}} \\ &\text{s.t.} \Theta - Q = 0,\ \Theta\mathcal{Z} - S = 0\,, \end{aligned} \tag{9}$$

where Q, S are slack variables. Then Eq. 9 can be solved by ADMM. The augmented Lagrangian is

$$\begin{aligned} \mathcal{L}_\rho(\Theta, Q, S, U, V) = &\frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|Q\|_{2,1} \\ &+\lambda_2\|S\|_{G_{2,1}} + \langle U, \Theta - Q\rangle + \frac{\rho}{2}\|\Theta \\ &-Q\|^2 + \langle V, \Theta\mathcal{Z} - S\rangle \\ &+\frac{\rho}{2}\|\Theta\mathcal{Z} - S\|^2\,, \end{aligned} \tag{10}$$

where U and V are augmented Lagrangian multipliers.

Update Θ: From the augmented Lagrangian in Eq. 10, the update of Θ at $(t+1)$-th iteration is carried out by:

$$\begin{aligned} \Theta^{(t+1)} = &\arg\min_\Theta \frac{1}{2}\|Y - X\Theta\|_F^2 + \left\langle U^{(t)}, \Theta\right. \\ &\left. - Q^{(t)}\right\rangle + \frac{\rho}{2}\left\|\Theta - Q^{(t)}\right\|^2 \\ &+ \left\langle V^{(t)}, \Theta\mathcal{Z} - S^{(t)}\right\rangle + \frac{\rho}{2}\left\|\Theta\mathcal{Z} - S^{(t)}\right\|^2 \end{aligned} \tag{11}$$

which is the closed form and can be derived by setting Eq. 11 to zero.

$$\begin{aligned} 0 = &-X^{\mathrm{T}}(Y - X\Theta) + U^{(t)} + \rho\left(\Theta - Q^{(t)}\right) + V^{(t)}\mathcal{Z} \\ &+ \rho\left(\Theta\mathcal{Z} - S^{(t)}\right)\mathcal{Z} \end{aligned} \tag{12}$$

Note that $\mathcal{Z}$ is a symmetric matrix. We define $\Phi = \mathcal{Z}\mathcal{Z}$, where Φ is also a symmetric matrix with $\Phi_{m,l}$ denoting the value of weight (m, l). With this linearization, the value for Θ can be updated in parallel by the individual $\theta_{\cdot m}$. Thus in the $(t+1)$-th iteration, $\theta_{\cdot m}^{(t+1)}$ can be updated efficiently using Cholesky factorization.

$$\begin{aligned} 0 = &-X^{\mathrm{T}}\left(\mathbf{y}_m - X\theta_{\cdot m}\right) + u_{\cdot m}^{(t)} + \rho\left(\theta_{\cdot m} - q_{\cdot m}^{(t)}\right) \\ &+\left(v_{\cdot m}^{(t)} - \sum_{\substack{m=1 \\ m\neq l}}^{k} \mathcal{Z}_{m,l} v_{\cdot l}^{(t)}\right) \\ &+\rho\left(\Phi_{m,m}\theta_{\cdot m} - \sum_{\substack{m=1 \\ m\neq l}}^{k} \Phi_{m,l}\theta_{\cdot l}\right) \\ &-\rho\left(s_{\cdot m}^{(t)} - \sum_{\substack{m=1 \\ m\neq l}}^{k} \mathcal{Z}_{m,l} s_{\cdot l}^{(t)}\right) \end{aligned} \tag{13}$$

The above optimization problem is quadratic. The optimal solution is given by $\theta_{\cdot m}^{(t+1)} = F_m^{-1} b_m^{(t)}$, where

$$\begin{aligned} F_m &= X^{\mathrm{T}}X + \rho(1 + \Phi_{m,m})I \\ b_m^{(t)} &= X^{\mathrm{T}}\mathbf{y}_m - u_{\cdot m}^{(t)} - \left(v_{\cdot m}^{(t)} - \sum_{\substack{m=1 \\ m\neq l}}^{k} \mathcal{Z}_{m,l} v_{\cdot l}^{(t)}\right) + \rho q_{\cdot m}^{(t)} \\ &\quad + \rho\left(s_{\cdot m}^{(t)} - \sum_{\substack{m=1 \\ m\neq l}}^{k} \mathcal{Z}_{m,l} s_{\cdot l}^{(t)}\right) + \rho\sum_{\substack{m=1 \\ m\neq l}}^{k} \Phi_{m,l}\theta_{\cdot l}\,. \end{aligned} \tag{14}$$

The computation of $\theta_{\cdot m}^{(t+1)}$ involves solving a linear system, which is the most time-consuming part of the whole algorithm. To efficiently compute $\theta_{\cdot m}^{(t+1)}$ efficiently, we can compute the Cholesky factorization of F at the beginning of the algorithm:

$$F_m = A_m^{\mathrm{T}} A_m\,. \tag{15}$$

Note that F is a constant and positive definite matrix. Using Cholesky factorization, we only need to solve the following two linear systems for each iteration:

$$A_m^{\mathrm{T}}\hat{\theta}_{\cdot m} = b^{(t)},\ A\theta_{\cdot m} = \hat{\theta}_{\cdot m}\,. \tag{16}$$

Since A_m is an upper triangular matrix, it is very efficient to solve these two linear systems.

Update Q: The update for Q effectively needs to solve the following problem

$$\begin{aligned} Q^{(t+1)} = &\arg\min_Q \frac{\rho}{2}\left\|Q - \Theta^{(t+1)}\right\|^2 + \lambda_1\|Q\|_{2,1} \\ &-\left\langle U^{(t)}, Q\right\rangle\,, \end{aligned} \tag{17}$$

which is equivalent to the following problem:

$$Q^{(t+1)} = \arg\min_Q \left\{\frac{\lambda_1}{\rho}\|Q\|_{2,1} + \frac{1}{2}\|Q - O^{(t+1)}\|^2\right\}\,. \tag{18}$$

where $O^{(t+1)} = \Theta^{(t+1)} + \frac{1}{\rho}U^{(t)}$. It is clear that Eq. 18 can be decoupled into

$$q_{i.}^{(t+1)} = \arg\min_{q_{i.}} \phi(q_{i.}) = \arg\min_{q_{i.}} \left\{ \frac{1}{2}\|q_{i.} - o_{i.}^{(t+1)}\|^2 + \frac{\lambda_1}{\rho}\|q_{i.}\| \right\} \quad (19)$$

where $q_{i.}$ and $o_{i.}$ are the i-th row of $Q^{(t+1)}$ and $O^{(t+1)}$, respectively. Since $\phi(q_{i.})$ is strictly convex, we conclude that $q_{i.}^{(t+1)}$ is its unique minimizer. Then we introduce the following lemma (Liu et al. 2009) to solve the Eq. 19.

Lemma 1 *For any $\lambda_1 \geq 0$, we have*

$$q_{i.} = \frac{\max\left\{\|o_{i.}\|_2 - \frac{\lambda_1}{\rho}, 0\right\}}{\|o_{i.}\|_2} o_{i.}\,, \quad (20)$$

Update S: The update for S effectively needs to solve the following problem

$$S^{(t+1)} = \arg\min_{S} \frac{\rho}{2}\left\|S - \Theta^{(t+1)}\mathcal{Z}\right\|^2 + \lambda_2\|S\|_{G_{2,1}} - \left\langle V^{(t)}, S\right\rangle, \quad (21)$$

which is effectively equivalent to computation of the proximal operator for $G_{2,1}$-norm. In particular, the problem can be written as

$$S^{(t+1)} = \arg\min_{S} \left\{ \frac{\lambda_2}{\rho}\|S\|_{G_{2,1}} + \frac{1}{2}\|S - \Pi^{(t+1)}\|^2 \right\}. \quad (22)$$

where $\Pi^{(t+1)} = \Theta^{(t+1)}\mathcal{Z} + \frac{1}{\rho}V^{(t)}$. Since the groups $\mathcal{G}_\ell$ used in our work are disjointed, the Eq. 22 can be decoupled into

$$s_{\mathcal{G}_l m}^{(t+1)} = \arg\min_{s_{\mathcal{G}_l m}} \phi(s_{\mathcal{G}_l m}) = \arg\min_{s_{\mathcal{G}_l m}} \left\{ \frac{1}{2}\|s_{\mathcal{G}_l m} - \pi_{\mathcal{G}_l m}^{(t+1)}\|^2 + \frac{\lambda_2}{\rho}\|s_{\mathcal{G}_l m}\| \right\} \quad (23)$$

where $s_{\mathcal{G}_\ell m}, \pi_{\mathcal{G}_\ell m}$ are rows in group $\mathcal{G}_\ell$ for task m of $S^{(t+1)}$ and $\Pi^{(t+1)}$, respectively. Then we introduce the following lemma (Yuan et al. 2013).

Lemma 2 *For any $\lambda_2 \geq 0$, we have*

$$s_{\mathcal{G}_\ell m} = \frac{\max\left\{\|\pi_{\mathcal{G}_\ell m}\|_2 - \frac{\lambda_2 v_\ell}{\rho}, 0\right\}}{\|\pi_{\mathcal{G}_\ell m}\|_2} \pi_{\mathcal{G}_\ell m}\,, \quad (24)$$

Dual Update for U and V: Following standard ADMM dual update, the update for the dual variables according to our setting is as follows:

$$U^{(t+1)} = U^{(t)} + \rho(\Theta^{(t+1)} - Q^{(t+1)}) \quad (25a)$$

$$V^{(t+1)} = V^{(t)} + \rho(\Theta^{(t+1)}\mathcal{Z} - S^{(t+1)})\,. \quad (25b)$$

The dual updates can be performed in an element-wise parallel manner. Algorithm 1 summarizes the whole algorithm. The MATLAB codes of the proposed algorithm are available at: https://bitbucket.org/XIAOLILIU/fgl-mtfl.

Algorithm 1 ADMM optimization of FGL-MTFL

Require: X, Y, λ_1, λ_2, ρ, E.
Ensure: Θ.
1: Initialization: $\Theta^{(0)} \leftarrow 0$, $Q^{(0)} \leftarrow 0$, $S^{(0)} \leftarrow 0$, $U^{(0)} \leftarrow 0$, $V^{(0)} \leftarrow 0$.
2: Compute the Cholesky factorization of F.
3: **repeat**
4: Update $\Theta^{(t+1)}$ according to Eq. (11).
5: Update $Q^{(t+1)}$ according to Eq. (17).
6: Update $S^{(t+1)}$ according to Eq. (21).
7: Update $U^{(s+1)}$, $V^{(s+1)}$ according to Eq. (25).
8: **until** Convergence.

Convergence

The convergence of the Algorithm 1 is shown in the following theorem

Theorem 3 *Suppose there exists at least one solution Θ^* of* Eq. 8. *Assume $L(\Theta)$ is convex, $\lambda_1 > 0$, $\lambda_2 > 0$. Then the following property for FGL-MTFL iteration in Algorithm 1 holds:*

$$\lim_{t\to\infty} L(\Theta^{(t)}) + \lambda_1\left\|\Theta^{(t)}\right\|_{2,1} + \lambda_2\left\|\Theta^{(t)}\mathcal{Z}\right\|_{G_{2,1}} = L(\Theta^*) + \lambda_1\left\|\Theta^*\right\|_{2,1} + \lambda_2\left\|\Theta^*\mathcal{Z}\right\|_{G_{2,1}} \quad (26)$$

Furthermore,

$$\lim_{t\to\infty}\left\|\Theta^{(t)} - \Theta^*\right\| = 0 \quad (27)$$

whenever Eq. 8 has a unique solution.

Note that the condition that allowed convergence in Theorem 3 is quite easy to satisfy. λ_1, λ_2 are regularization parameters and should always be larger than zero. The detailed proof is discussed in (Cai et al. 2009). Unlike Cai et al., we do not require $L(\Theta)$ to be differentiable, and explicitly treat the non-differentiability of $L(\Theta)$ by using its subgradient vector $\partial L(\Theta)$, similar to the strategy used by (Ye and Xie 2011).

Experimental Results and Discussions

In this section, we present experimental results to demonstrate the effectiveness of the proposed FGL-MTFL to characterize AD progression using a dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Weiner et al. 2010).

Experimental Setup

MR images and data used in this work were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu) (Weiner et al. 2010). The primary goal of ADNI is to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. Approaches to characterize AD progression will help researchers and clinicians develop new treatments and monitor their effectiveness. Further, a better understanding of disease progression will increase the safety and efficacy of drug development and potentially decrease the time and cost of clinical trails. In ADNI, all participants received 1.5 Tesla (T) structural MRI. The MRI features used in our experiments are based on imaging data from the ADNI database processed by a team from UCSF (University of California at San Francisco), who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) according to the atlas generated in (Desikan et al. 2006). The FreeSurfer software was employed to automatically label the cortical and subcortical tissue classes for the structural MRI scan of each subject and to extract thickness measures of the cortical regions of interests (ROIs) and volume measures of cortical and subcortical regions.

Briefly, this processing includes motion correction and the averaging (Reuter et al. 2010) of multiple volumetric T1-weighted images (when more than one is available), removal of non-brain tissue using a hybrid watershed/surface deformation procedure (Segonne et al. 2004), automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, and ventricles) (Fischl et al. 2002, 2004), intensity normalization (Sled et al. 1998), tessellation of the gray matter white matter boundary, automated topology correction (Fischl et al. 2001; Ségonne et al. 2007), 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 (Dale et al. 1999; Dale and Sereno 1993).

In total, 48 cortical regions and 44 subcortical regions were generated, , with typically 1 or 4 features in each group. The names of the cortical and subcortical regions are listed in Tables 1 and 2. For each cortical region, the cortical thickness average (TA), standard deviation of thickness (TS), surface area (SA), and cortical volume (CV) were calculated as features. For each subcortical region, the subcortical volume was calculated as a feature. The separate SA values for the left and right hemisphere and the total intracranial volume (ICV) were also included.

Table 1 Cortical features from the following 71 ($= 35 \times 2 + 1$) cortical regions generated by FreeSurfer

ID	ROI name	Laterality	Type
1	Banks superior temporal sulcus	L, R	CV, SA, TA, TS
2	Caudal anterior cingulate cortex	L, R	CV, SA, TA, TS
3	Caudal middle frontal gyrus	L, R	CV, SA, TA, TS
4	Cuneus cortex	L, R	CV, SA, TA, TS
5	Entorhinal cortex	L, R	CV, SA, TA, TS
6	Frontal pole	L, R	CV, SA, TA, TS
7	Fusiform gyrus	L, R	CV, SA, TA, TS
8	Inferior parietal cortex	L, R	CV, SA, TA, TS
9	Inferior temporal gyrus	L, R	CV, SA, TA, TS
10	Insula	L, R	CV, SA, TA, TS
11	IsthmusCingulate	L, R	CV, SA, TA, TS
12	Lateral occipital cortex	L, R	CV, SA, TA, TS
13	Lateral orbital frontal cortex	L, R	CV, SA, TA, TS
14	Lingual gyrus	L, R	CV, SA, TA, TS
15	Medial orbital frontal cortex	L, R	CV, SA, TA, TS
16	Middle temporal gyrus	L, R	CV, SA, TA, TS
17	Paracentral lobule	L, R	CV, SA, TA, TS
18	Parahippocampal gyrus	L, R	CV, SA, TA, TS
19	Pars opercularis	L, R	CV, SA, TA, TS
20	Pars orbitalis	L, R	CV, SA, TA, TS
21	Pars triangularis	L, R	CV, SA, TA, TS
22	Pericalcarine cortex	L, R	CV, SA, TA, TS
23	Postcentral gyrus	L, R	CV, SA, TA, TS
24	Posterior cingulate cortex	L, R	CV, SA, TA, TS
25	Precentral gyrus	L, R	CV, SA, TA, TS
26	Precuneus cortex	L, R	CV, SA, TA, TS
27	Rostral anterior cingulate cortex	L, R	CV, SA, TA, TS
28	Rostral middle frontal gyrus	L, R	CV, SA, TA, TS
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	CV

275 ($= 34 \times 2 \times 4 + 1 \times 2 \times 1 + 1$) cortical features calculated were analyzed in this study. Laterality indicates different feature types calculated for L (left hemisphere), R (right hemisphere) or Bilateral (whole hemisphere)

This yielded a total of $p = 319$ MRI features extracted from cortical/subcortical ROIs in each hemisphere (Tables 1 and 2). Details of the analysis procedure are available at http://adni.loni.ucla.edu/research/mri-post-processing/.

The ADNI project is a longitudinal study, with data collected repeatedly over a 6-month or 1-year interval. The

Table 2 Subcortical features from the following 44 ($= 16 \times 2 + 12$) subcortical regions generated by FreeSurfer

number	ROI	Laterality	Type
1	Accumbens area	L, R	SV
2	Amygdala	L, R	SV
3	Caudate	L, R	SV
4	Cerebellum cortex	L, R	SV
5	Cerebellum white matter	L, R	SV
6	Cerebral cortex	L, R	SV
7	Cerebral white matter	L, R	SV
8	Choroid plexus	L, R	SV
9	Hippocampus	L, R	SV
10	Inferior lateral ventricle	L, R	SV
11	Lateral ventricle	L, R	SV
12	Pallidum	L, R	SV
13	Putamen	L, R	SV
14	Thalamus	L, R	SV
15	Ventricle diencephalon	L, R	SV
16	Vessel	L, R	SV
17	Brain stem	Bilateral	SV
18	Corpus callosum anterior	Bilateral	SV
19	Corpus callosum central	Bilateral	SV
20	Corpus callosum middle anterior	Bilateral	SV
21	Corpus callosum middle posterior	Bilateral	SV
22	Corpus callosum posterior	Bilateral	SV
23	Cerebrospinal fluid	Bilateral	SV
24	Fourth ventricle	Bilateral	SV
25	Non white matter hypointensities	Bilateral	SV
26	Optic chiasm	Bilateral	SV
27	Third ventricle	Bilateral	SV
28	White matter hypointensities	Bilateral	SV

44 subcortical features calculated were analyzed in this study. Laterality indicates different feature types calculated for L (left hemisphere), R (right hemisphere) or Bilateral (whole hemisphere)

scheduled screening date for subjects becomes baseline after approval and the time point for the follow-up visits is denoted by the duration starting from the baseline. In our current work, we investigated the prediction performance of our method to infer cognitive outcomes based on a number of neuropsychological assessments at the time of the initial baseline. In this work, we further performed the following preprocessing steps:

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

This yields a total of $n = 788$ subjects, who were then categorized into three baseline diagnostic groups: Cognitively Normal (CN, $n_1 = 225$), Mild Cognitive Impairment (MCI, $n_2 = 390$), and Alzheimer's Disease (AD, $n_3 = 173$). Table 3 lists the demographics information of all subjects, including age, gender, and education. We used 10-fold cross validation to evaluate our model and conducted the comparison. In each of ten trials, a 5-fold nested cross validation procedure was employed to tune the regularization parameters. The data were z-scored before applying the regression methods. The range of each parameter varied from 10^{-1} to 10^{3}. The average (avg) and standard deviation (std) of performance measures were calculated and shown as avg $\pm$ std for each experiment. For each run, all methods received exactly the same train and test set. The reported results are the best results of each method with the optimal parameters. For predictive modeling, all the cognitive assessments (a total of 20 tasks) in Table 4 were examined. To the best of our knowledge, no previous works have used all the cognitive scores to train and evaluate their MTL models.

For the quantitative performance evaluation, we employed the metrics of Correlation Coefficient (CC) and Root Mean Squared Error (rMSE) between the predicted clinical scores and the target clinical scores for each regression task. To evaluate the overall performance for all tasks, the normalized mean squared error (nMSE) (Argyriou et al. 2008; Zhou et al. 2013) and weighted R-value (wR) (Stonnington et al. 2010) were used. The rMSE, CC, nMSE and wR are defined as follows:

$$\mathrm{rMSE}\left(\mathrm{y}, \hat{\mathrm{y}}\right) = \frac{\left\|\mathrm{y} - \hat{\mathrm{y}}\right\|_2^2}{n} \tag{28}$$

$$\mathrm{Corr}\left(\mathrm{y}, \hat{\mathrm{y}}\right) = \frac{\mathrm{cov}(\mathrm{y}, \hat{\mathrm{y}})}{\sigma(\mathrm{y})\sigma(\hat{\mathrm{y}})} \tag{29}$$

where y is the ground truth of the target at a single task and $\hat{\mathrm{y}}$ is the corresponding prediction according to a prediction model, cov is the covariance, and σ is the standard deviation.

$$\mathrm{nMSE}\left(Y, \hat{Y}\right) = \frac{\sum_{h=1}^{k} \frac{\|Y_h - \hat{Y}_h\|_2^2}{\sigma(Y_h)}}{\sum_{h=1}^{k} n_h} \tag{30}$$

Table 3 Summary of ADNI dataset and subject information

Category	CN	MCI	AD
Number	225	390	173
Gender (M/F)	116/109	252/138	88/85
Age (y, ag $\pm$ sd)	75.87 $\pm$ 5.04	74.75 $\pm$ 7.39	75.42 $\pm$ 7.25
Education (y, ag $\pm$ sd)	16.03 $\pm$ 2.85	15.67 $\pm$ 2.95	14.65 $\pm$ 3.17

M, male; F, female; y, years; ag, average; sd, standard deviation

Table 4 Description of the cognitive scores considered in the experiments

Num	Score name		Description
1	ADAS		Alzheimer's disease assessment scale
2	MMSE		Mini-mental state exam
3	RAVLT	TOTAL	Total score of the first 5 learning trials
4		TOT6	Trial 6 total number of words recalled
5		TOTB	Immediately after the fifth learning trial
6		T30	30 minute delay total number of words recalled
7		RECOG	30 minute delay recognition
8	FLU	ANIM	Animal total score
9		VEG	Vegetable total score
10	TRAILS	A	Trail making test A score
11		B	Trail making test B score
12	LOGMEM	IMMTOTAL	Immediate recall
13		DELTOTAL	Delayed recall
14	CLOCK	DRAW	Clock drawing
15		COPYSCORE	Clock copying
16	BOSNAM		Total number correct
17	ANART		ANART total score
18	DSPAN	For	Digit span forward
19		BAC	Digit span backward
20	DIGIT		Digit symbol substitution

$$\mathrm{wR}\left(Y, \hat{Y}\right) = \frac{\sum_{h=1}^{k} \mathrm{Corr}(Y_h, \hat{Y}_h) n_h}{\sum_{h=1}^{k} n_h} \tag{31}$$

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

Smaller values of nMSE and rMSE and larger values of CC and wR indicate better regression performance. We report the mean and standard deviation based on 10 experimental iterations on different splits of data for all comparable experiments. A Student's t-test at a significance level of 0.05 is performed to determine whether the performances difference are significant.

Comparison with Baseline Comparable Methods

In this section, we conduct empirical evaluation for the proposed methods by comparison with two single-task learning methods: Lasso and Ridge. Both methods were applied independently to each task, and compared to representative multi-task learning methods:

1. Multi-task feature learning (MTFL): $\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda\|\Theta\|_{2,1}$.
2. Multi-task feature learning combined with lasso (SGL-MTFL): $\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_{2,1} + \lambda_2\|\Theta\|_1$.
3. Fused lasso regularized multi-task learning (FL-MTL): $\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda\|\Theta\mathcal{Z}\|_1$.
4. Fused group lasso regularized multi-task learning (FGL-MTL): $\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda\|\Theta\mathcal{Z}\|_{G_{2,1}}$.
5. Fused lasso regularized multi-task feature learning (FL-MTFL): $\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_{2,1} + \lambda_2\|\Theta\mathcal{Z}\|_1$.
6. Group lasso regularized multi-task feature learning (GL-MTFL): $\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_{2,1} + \lambda_2\|\Theta\|_{G_{2,1}}$.
7. Fused lasso regularized SGL-MTL (FSGL-MTL):$\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_1 + \lambda_2\|\Theta\mathcal{Z}\|_{G_{2,1}}$.
8. Fused group lasso regularized MTFL (FGL-MTFL): $\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_{2,1} + \lambda_2\|\Theta\mathcal{Z}\|_{G_{2,1}}$.

The average and standard deviation of performance measures were calculated by 10-fold cross validation on different splits of data, and are shown in Tables 5 and 6. It is worth noting that the same training and testing data were used across experiments for all methods for fair comparison. Note that all pairwise links between tasks were incorporated into both the fused lasso and fused group lasso regularized methods, and the graph was created using only the training data.

From the experimental results in Tables 5 and 6, we observe the following:

1. FL-MTFL and FGL-MTFL both demonstrated an improved performance over the other baseline methods in terms of nMSE and wR, while FGL-MTFL performed the best among all competing methods. The

Table 5 Performance comparison of various methods in terms of CC (the first 20 columns) and wR (the last column) on 10 cross validation cognitive prediction tasks

Method	ADAS	MMSE	RAVLT					FLU		TRAILS	
			TOTAL	TOT6	TOTB	T30	RECOG	ANIM	VEG	A	B
Ridge	0.598±0.053	0.417±0.073	0.407±0.120	0.352±0.131	0.145±0.093	0.367±0.132	0.259±0.107	0.200±0.135	0.390±0.132	0.316±0.111	0.363±0.125
Lasso	0.659±0.066	0.547±0.065	0.512±0.108	0.501±0.117	0.312±0.083	0.514±0.116	0.407±0.112	0.353±0.095	0.504±0.089	0.382±0.099	0.477±0.094
MTFL	0.667±0.065	0.536±0.065	0.515±0.121	0.491±0.132	0.297±0.094	0.509±0.123	0.405±0.123	0.377±0.091	0.505±0.089	0.413±0.087	0.476±0.098
FL-MTL	0.641±0.074	0.523±0.095	0.491±0.112	0.472±0.102	0.287±0.085	0.473±0.096	0.409±0.085	0.355±0.107	0.485±0.076	0.352±0.085	0.385±0.159
FGL-MTL	0.662±0.060	0.557±0.065	0.479±0.109	0.501±0.116	0.286±0.080	0.516±0.113	0.409±0.115	0.370±0.106	0.512±0.077	0.347±0.101	0.351±0.119
SGL-MTFL	0.667±0.063	0.543±0.061	0.517±0.120	0.495±0.128	0.308±0.089	0.513±0.120	0.408±0.123	0.380±0.090	0.506±0.089	0.414±0.086	0.477±0.097
GL-MTFL	0.668±0.063	0.546±0.062	0.517±0.119	0.496±0.129	0.308±0.087	0.514±0.120	0.410±0.123	0.380±0.090	0.508±0.088	0.414±0.086	0.477±0.097
FL-MTFL	0.665±0.067	0.548±0.074	0.522±0.118	0.498±0.121	0.324±0.084	0.513±0.115	0.420±0.127	0.396±0.086	0.510±0.084	0.414±0.086	0.479±0.095
FSGL-MTL	0.623±0.071	0.503±0.083	0.519±0.091	0.437±0.101	0.306±0.065	0.430±0.092	0.352±0.122	0.387±0.072	0.483±0.080	0.379±0.104	0.476±0.100
FGL-MTFL	0.668±0.068	0.549±0.068	0.522±0.114	0.499±0.121	0.324±0.069	0.514±0.110	0.429±0.121	0.399±0.081	0.505±0.087	0.419±0.086	0.481±0.095

Method	LOGMEM		CLOCK		BOSNAM	ANART	DSPAN		DIGIT	wR
	IMMTOTAL	DELTOTAL	DRAW	COPYSCOR			FOR	BAC		
Ridge	0.417±0.108	0.426±0.121	0.222±0.100	0.126±0.091	0.359±0.149	0.054±0.071	0.002±0.055	0.041±0.114	0.389±0.049	0.293±0.046*
Lasso	0.493±0.093	0.524±0.107	0.374±0.078	0.209±0.050	0.472±0.104	0.129±0.078	0.025±0.083	0.167±0.132	0.415±0.099	0.399±0.049*
MTFL	0.507±0.085	0.519±0.099	0.380±0.097	0.199±0.114	0.468±0.126	0.104±0.080	0.089±0.106	0.153±0.081	0.478±0.064	0.404±0.055*
FL-MTL	0.470±0.097	0.514±0.108	0.379±0.093	0.221±0.102	0.442±0.085	0.094±0.104	0.083±0.102	0.177±0.108	0.414±0.083	0.383±0.046*
FGL-MTL	0.499±0.093	0.531±0.106	0.378±0.087	0.212±0.098	0.479±0.096	0.139±0.079	0.110±0.105	0.205±0.108	0.450±0.049	0.400±0.047*
SGL-MTFL	0.508±0.086	0.520±0.098	0.394±0.092	0.210±0.096	0.471±0.118	0.110±0.081	0.099±0.104	0.167±0.087	0.477±0.063	0.409±0.054*
GL-MTFL	0.509±0.084	0.521±0.098	0.395±0.087	0.220±0.102	0.471±0.118	0.109±0.081	0.089±0.103	0.169±0.085	0.477±0.063	0.410±0.054*
FL-MTFL	0.510±0.084	0.526±0.095	0.408±0.083	0.251±0.092	0.471±0.096	0.133±0.087	0.093±0.098	0.208±0.113	0.478±0.069	0.418±0.054
FSGL-MTL	0.452±0.073	0.492±0.087	0.398±0.074	0.2412±0.091	0.440±0.070	0.119±0.113	0.109±0.113	0.209±0.126	0.435±0.085	0.390±0.045*
FGL-MTFL	0.514±0.084	0.533±0.094	0.408±0.083	0.252±0.092	0.473±0.089	0.137±0.082	0.096±0.090	0.218±0.113	0.484±0.070	**0.421±0.052**

Superscript symbols ∗ indicates that FGL-MTFL significantly outperformed that method on that score. Student's t-test at a level of 0.05 was used

The bold value indicates the best performance

Table 6 Performance comparison of various methods in terms of rmse (the first 20 columns) and nMSE (the last column) on 10 cross validation cognitive prediction tasks

Method	ADAS	MMSE	RAVLT					FLU		TRAILS	
			TOTAL	TOT6	TOTB	T30	RECOG	ANIM	VEG	A	B
Ridge	7.445±0.370	2.568±0.146	11.16±0.735	3.909±0.366	1.982±0.125	4.061±0.287	4.311±0.430	6.307±0.552	4.276±0.386	26.19±3.764	80.01±8.103
Lasso	6.727±0.424	2.215±0.088	9.909±0.907	3.305±0.266	1.666±0.166	3.421±0.244	3.634±0.247	5.464±0.486	3.690±0.192	23.29±4.297	69.78±5.152
MTFL	6.662±0.411	2.191±0.106	9.656±0.695	3.324±0.260	1.671±0.149	3.441±0.233	3.626±0.272	5.266±0.448	3.676±0.180	23.01±3.492	69.88±5.281
FL-MTL	7.012±0.471	2.295±0.226	10.02±0.830	3.429±0.435	1.796±0.199	3.598±0.536	3.684±0.319	5.362±0.617	3.772±0.239	24.77±4.299	78.18±10.52
FGL-MTL	6.691±0.458	2.157±0.097	10.04±0.593	3.309±0.271	1.676±0.153	3.420±0.235	3.627±0.265	5.280±0.471	3.672±0.182	24.85±3.600	81.07±7.425
SGL-MTFL	6.653±0.427	2.191±0.098	9.647±0.755	3.315±0.271	1.664±0.150	3.432±0.246	3.627±0.246	5.260±0.499	3.676±0.208	23.00±3.568	69.82±5.184
GL-MTFL	6.650±0.429	2.186±0.099	9.644±0.754	3.315±0.270	1.664±0.150	3.431±0.248	3.623±0.247	5.261±0.497	3.673±0.207	22.99±3.566	69.82±5.177
FL-MTFL	6.680±0.445	2.204±0.096	9.6223±0.784	3.324±0.273	1.663±0.169	3.445±0.286	3.617±0.192	5.220±0.503	3.677±0.222	22.97±3.660	69.67±5.007
FSGL-MTL	7.183±0.730	2.523±0.124	9.811±0.859	3.623±0.481	2.009±0.180	3.791±0.537	3.912±0.222	5.365±0.635	3.929±0.383	23.27±4.204	69.75±5.084
FGL-MTFL	6.678±0.474	2.214±0.081	9.631±0.788	3.335±0.276	1.668±0.167	3.456±0.305	3.610±0.186	5.221±0.493	3.697±0.223	22.90±3.666	69.53±5.011

Method	LOGMEM		CLOCK		BOSNAM	ANART	DSPAN		DIGIT	nMSE
	IMMTOTAL	DELTOTAL	DRAW	COPYSCOR			FOR	BAC		
Ridge	4.696±0.366	5.277±0.509	1.153±0.108	0.775±0.060	4.563±0.540	11.24±0.756	2.570±0.262	2.559±0.193	12.76±1.273	10.36±1.089*
Lasso	4.198±0.327	4.586±0.471	0.990±0.099	0.657±0.082	3.992±0.457	9.813±0.651	2.100±0.285	2.134±0.191	11.65±1.318	7.898±0.586*
MTFL	4.145±0.303	4.589±0.436	0.967±0.120	0.647±0.103	3.952±0.478	9.611±0.722	2.010±0.123	2.160±0.178	11.22±1.225	7.762±0.633*
FL-MTL	4.319±0.496	4.672±0.689	1.174±0.263	0.942±0.323	4.073±0.494	9.806±0.850	2.131±0.252	2.225±0.179	11.76±1.391	9.102±1.134*
FGL-MTL	4.170±0.334	4.560±0.494	0.973±0.112	0.656±0.081	3.938±0.471	9.746±0.640	1.989±0.136	2.116±0.199	11.58±1.287	9.114±0.923*
SGL-MTFL	4.143±0.327	4.585±0.456	0.960±0.117	0.644±0.093	3.965±0.425	9.584±0.743	1.998±0.134	2.143±0.186	11.23±1.261	7.742±0.617*
GL-MTFL	4.140±0.326	4.584±0.456	0.960±0.114	0.643±0.092	3.965±0.424	9.585±0.743	2.001±0.133	2.142±0.186	11.23±1.261	7.740±0.616*
FL-MTFL	4.147±0.354	4.582±0.491	0.974±0.101	0.654±0.076	3.978±0.459	9.491±0.728	1.998±0.145	2.121±0.187	11.21±1.263	7.711±0.585*
FSGL-MTL	4.467±0.525	4.859±0.719	1.485±0.072	1.307±0.049	4.208±0.411	9.553±0.721	2.289±0.208	2.394±0.244	11.56±1.320	8.329±0.593*
FGL-MTFL	4.139±0.366	4.566±0.505	0.973±0.102	0.654±0.076	3.984±0.485	9.471±0.708	1.997±0.149	2.119±0.191	11.17±1.241	**7.689±0.565**

Superscript symbols ∗ indicates that FGL-MTFL significantly outperformed that method on that score. Student's t-test at a level of 0.05 was used

The bold value indicates the best performance

t-test results show that the FGL-MTFL significantly outperforms all the other methods in terms of nMSE and all the other methods except FL-MTFL in terms of CC. Thus, simultaneously exploiting the structure among the tasks and features resulted in significantly better prediction performance.

2. The norm of both the fused lasso and fused group lasso can improve the performance of the traditional MTFL, which demonstrates that considering the local structure within the tasks can improve the overall prediction performance. A fundamental step is to estimate the true relationships among tasks, thus promoting the most appropriate information sharing among related tasks while avoiding the use of information from unrelated tasks.
3. We investigated the effect of group penalty in our model. The traditional MTFL consideres only the sparsity of regression coefficients, thus failing to capture the group structure of features in the data.

From the Tables 5 and 6, we can find that both FL-MTFL and FGL-MTFL are better than traditional MTFL. This is because the assumption in MTFL is a strict constraint and may affect the flexibility of the multi-task model as described in "$\ell_{2,1}$-norm". The extraction of multiple features to measure the atrophy of each imaging biomarker can further improve prediction performance by capturing inherent feature structures.

GL-MTFL is similar to FGL-MTFL but the regularization term of $G_{2,1}$-norm ignore the structure of tasks. This approach can flexibly take into account the complex relationships among imaging markers in a fused lasso regularization rather than relying on a simple grouping scheme. Moreover, compared with GL-MTFL, our FGL-MTFL can flexibly consider the complex relationships among outcomes in a group format rather than within a simple group lasso scheme. Overall, FGL-MTFL achieved better performances than

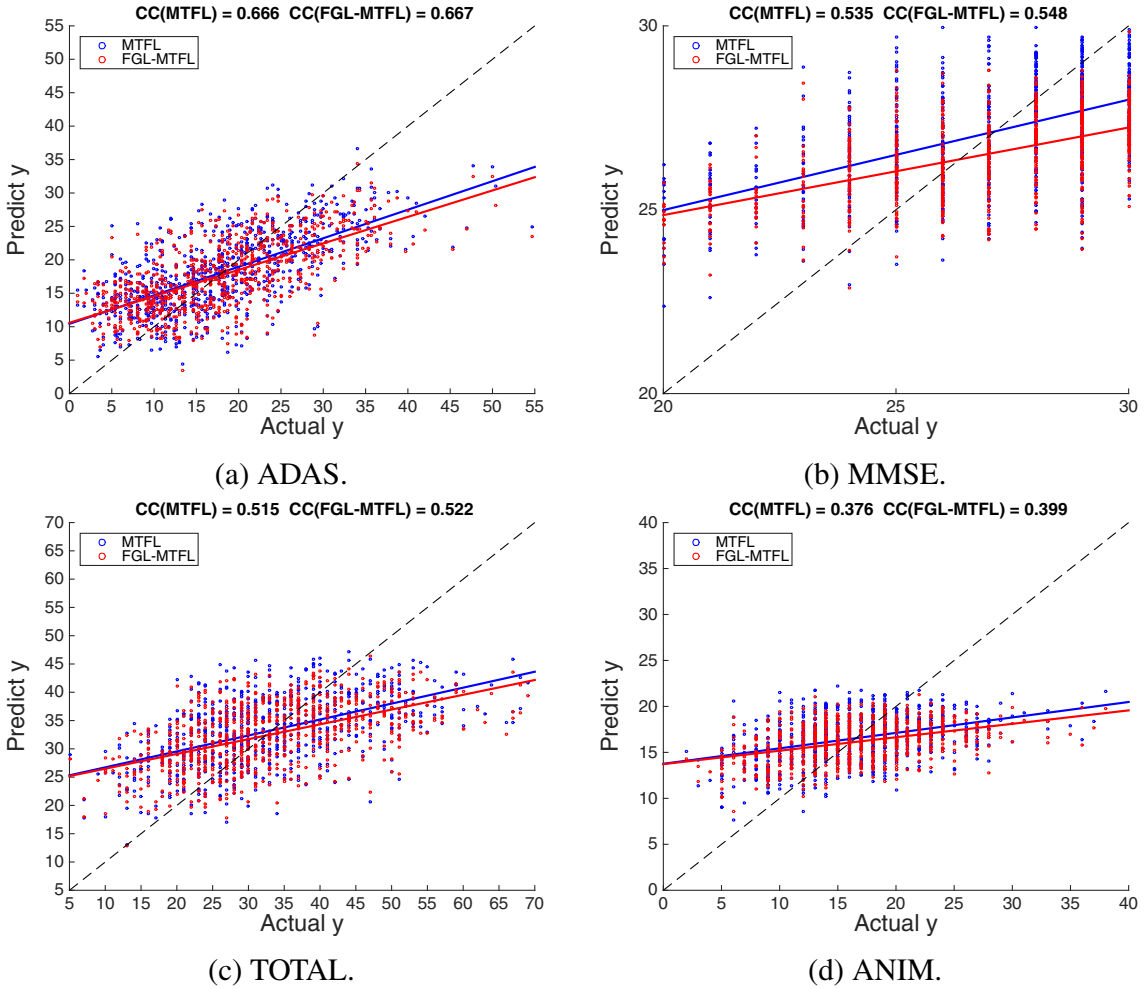


Fig. 3 Scatter plots of actual versus predicted values of cognitive scores on each fold testing data using MTFL and FBGL-MTL methods based on MRI features

FL-MTFL, demonstrating the benefit of employing group structural information among the features.

4. All compared multi-task learning methods improve predictive performance over the independent regression algorithm (Ridge, Lasso, GOSCAR, and ncFGS). This justifies the motivation to learn multiple tasks simultaneously.

Additionally, we show the scatter plots for the predicted values versus the actual values for ADAS, MMSE, TOTAL, and ANIM for the testing data in the cross-validation in Fig. 3.

To investigate the influence of edge weight on performance, we compared our proposed FGL-MTFL with an unweighted FGL-MTFL ($\lambda_1\|\Theta\|_{2,1} + \lambda_2\sum_{m=1}^{k}\|\theta_{.m} - \text{sign}(r_{m,l})\theta_{.l}\|_{G_{2,1}}$), which only considers the graph topology of tasks and treats all links equally. The result is shown in Table 7, and shows obvious improvement with the weighted FGL-MTFL compared to the unweighted one.

Comparison with the State-of-the-Art MTL Methods

To illustrate how well our FGL-MTFL works by means of modeling the correlation among the tasks, we comprehensively compared our proposed methods with several popular state-of-the-art related methods. The representative multi-task learning algorithms used for comparison include the following models:

1. Robust multi-Task Feature Learning (RMTL) (Chen et al. 2011):
   RMTL ($\min_\Theta L(X, Y, \Theta) + \lambda_1\|P\|_* + \lambda_2\|S\|_{2,1}$, subject to $\Theta = P + S$), which assumes that the model Θ can be decomposed into two components: a shared feature structure P that captures task relatedness and a group-sparse structure S that can detect outliers.
2. Clustered multi-Task Learning (CMTL) (Zhou et al. 2011):
   CMTL ($\min_{\Theta, M: M^{\mathrm{T}}M=I_c} L(X, Y, \Theta) + \lambda_1(\text{Tr}(\Theta^{\mathrm{T}}\Theta) - \text{Tr}(M^{\mathrm{T}}\Theta^{\mathrm{T}}\Theta M)) + \lambda_2\text{Tr}(\Theta^{\mathrm{T}}\Theta)$, where $M \in \mathbb{R}^{c\times k}$ is an orthogonal cluster indicator matrix, and the tasks are clustered into $c < k$ clusters) incorporates a regularization term to induce clustering between tasks and then shares information only to tasks belonging to the same cluster. In CMTL, the number of clusters is set to 11 because the 20 tasks belong to 11 sets of cognitive functions.
3. Trace-Norm Regularized multi-Task Learning (Trace) (Ji and Ye 2009): Assumes that all models share a common low-dimensional subspace ($\min_\Theta L(X, Y, \Theta) + \lambda\|\Theta\|_*$).
4. Sparse regularized multi-task learning formulation (SRMTL) (Zhou):
   SRMTL ($\min_\Theta L(X, Y, \Theta) + \lambda_1\|\Theta\mathcal{Z}\|_F^2 + \lambda_2\|\Theta\|_1$, where $\mathcal{Z} \in \mathbb{R}^{k\times k}$) contains two regularization

Table 7 The influence of weighting scheme of fused group lasso norm in FGL-MTFL in terms of CC and wR

Method	ADAS	MMSE	RAVLT					FLU		TRAILS	
			TOTAL	TOT6	TOTB	T30	RECOG	ANIM	VEG	A	B
unweighed	0.669±0.062	0.542±0.055	0.517±0.116	0.4923±0.128	0.295±0.095	0.510±0.120	0.399±0.120	0.381±0.090	0.507±0.091	0.416±0.085	0.478±0.098
weighted	0.668±0.068	0.549±0.068	0.522±0.114	0.499±0.121	0.324±0.069	0.514±0.110	0.429±0.121	0.399±0.081	0.505±0.087	0.419±0.086	0.481±0.095
Method	LOGMEM		CLOCK		BOSNAM	ANART	DSPAN		DIGIT	wR	
	IMMTOTAL	DELTOTAL	DRAW	COPYSCOR			FOR	BAC			
unweighed	0.508±0.086	0.519±0.099	0.341±0.077	0.090±0.132	0.471±0.113	0.107±0.085	0.086±0.114	0.141±0.098	0.479±0.065	0.397±0.052*	
weighted	0.514±0.084	0.533±0.094	0.408±0.083	0.252±0.092	0.473±0.089	0.137±0.082	0.096±0.090	0.218±0.113	0.484±0.070	**0.421±0.052**	

The first 20 columns are CC and the last one is wR
The bold value indicates the best performance

Table 8 Performance comparison of various methods in terms of CC (the first 20 columns) and wR (the last column) on 10-fold cross validation for all the cognitive prediction tasks

Method	ADAS	MMSE	RAVLT					FLU		TRAILS	
			TOTAL	TOT6	TOTB	T30	RECOG	ANIM	VEG	A	B
Robust	0.604±0.056	0.394±0.085	0.423±0.132	0.420±0.143	0.248±0.121	0.427±0.119	0.339±0.122	0.272±0.130	0.444±0.102	0.309±0.114	0.319±0.121
CMTL	0.646±0.060	0.429±0.070	0.494±0.090	0.474±0.106	0.263±0.084	0.480±0.102	0.381±0.094	0.360±0.088	0.494±0.086	0.419±0.092	0.479±0.088
Trace	0.559±0.070	0.255±0.121	0.413±0.125	0.428±0.124	0.227±0.142	0.445±0.111	0.356±0.135	0.265±0.155	0.419±0.111	0.282±0.104	0.313±0.094
SRMTL	0.664±0.062	0.542±0.064	0.495±0.095	0.497±0.117	0.323±0.081	0.514±0.112	0.401±0.118	0.373±0.101	0.502±0.089	0.376±0.095	0.425±0.122
p-MSSL	0.666±0.071	0.541±0.074	0.520±0.099	0.493±0.113	0.312±0.066	0.510±0.107	0.420±0.128	0.394±0.085	0.497±0.080	0.378±0.087	0.398±0.114
G-SMuRFS	0.670±0.062	0.535±0.060	0.506±0.114	0.482±0.124	0.284±0.100	0.492±0.116	0.391±0.113	0.366±0.095	0.495±0.092	0.409±0.085	0.471±0.101
MTRL	0.636±0.056	0.543±0.071	0.504±0.081	0.476±0.104	0.313±0.070	0.478±0.095	0.387±0.113	0.395±0.082	0.506±0.081	0.417±0.097	0.477±0.091
FGL-MTFL	0.668±0.068	0.549±0.068	0.522±0.114	0.499±0.121	0.324±0.069	0.514±0.110	0.429±0.121	0.399±0.081	0.505±0.087	0.419±0.086	0.481±0.095

Method	LOGMEM		CLOCK		BOSNAM	ANART	DSPAN		DIGIT	wR
	IMMTOTAL	DELTOTAL	DRAW	COPYSCOR			FOR	BAC		
Robust	0.448±0.109	0.470±0.106	0.342±0.099	0.133±0.112	0.373±0.130	0.050±0.083	0.007±0.120	0.130±0.089	0.392±0.049	0.327±0.059*
CMTL	0.497±0.086	0.515±0.096	0.370±0.077	0.203±0.114	0.440±0.079	0.134±0.080	0.044±0.130	0.188±0.116	0.468±0.070	0.389±0.046*
Trace	0.423±0.121	0.470±0.119	0.294±0.123	0.065±0.136	0.236±0.141	0.045±0.082	0.041±0.137	0.149±0.069	0.323±0.094	0.301±0.070*
SRMTL	0.504±0.091	0.525±0.102	0.362±0.077	0.084±0.120	0.482±0.095	0.117±0.082	0.092±0.092	0.178±0.125	0.463±0.051	0.396±0.046*
p-MSSL	0.512±0.087	0.538±0.098	0.384±0.063	0.236±0.110	0.483±0.085	0.150±0.078	0.096±0.113	0.197±0.128	0.469±0.074	0.410±0.045*
G-SMuRFS	0.502±0.091	0.514±0.105	0.366±0.100	0.197±0.110	0.475±0.129	0.091±0.078	0.070±0.102	0.132±0.104	0.477±0.059	0.396±0.054*
MTRL	0.493±0.075	0.507±0.089	0.402±0.076	0.262±0.084	0.470±0.087	0.143±0.089	0.108± 0.102	0.223±0.128	0.469±0.0741	0.410±0.045*
FGL-MTFL	0.514±0.084	0.533±0.094	0.408±0.083	0.252±0.092	0.473±0.089	0.137±0.082	0.096±0.090	0.218±0.113	0.484±0.070	**0.421±0.052**

Superscript symbols ∗ indicates that FGL-MTFL significantly outperformed that method on that score. Student's t-test at a level of 0.05 was used

The bold value indicates the best performance

Table 9 Performance comparison of various methods in terms of rmse (the first 20 columns) and nMSE (the last column) on 10 cross validation cognitive prediction tasks

Method	ADAS	MMSE	RAVLT					FLU		TRAILS	
			TOTAL	TOT6	TOTB	T30	RECOG	ANIM	VEG	A	B
Robust	7.339±0.549	2.811±0.131	10.82±0.815	3.587±0.333	1.729±0.139	3.742±0.238	3.887±0.364	5.7623±0.491	3.948±0.307	26.52±3.501	85.30±7.559
CMTL	15.88±1.291	21.52±0.443	27.94±1.645	4.944±0.654	3.520±0.172	4.562±0.623	8.824±0.537	14.06±1.105	9.731±0.777	43.51±4.236	126.2±11.40
Trace	7.969±0.925	4.310±1.464	11.50±1.000	3.629±0.325	1.805±0.136	3.746±0.220	3.988±0.562	6.086±0.724	4.179±0.593	28.41±2.181	88.38±5.778
SRMTL	6.925±0.464	2.405±0.316	10.40±0.814	4.067±0.878	3.028±1.786	4.233±0.940	4.116±0.737	5.432±0.495	4.243±0.869	24.07±3.794	75.16±8.209
p-MSSL	6.698±0.521	2.330±0.117	9.675±0.803	3.400±0.383	1.802±0.177	3.514±0.397	3.693±0.211	5.267±0.577	3.780±0.303	23.49±3.806	75.87±7.699
G-SMuRFS	6.641±0.422	2.209±0.107	9.741±0.708	3.348±0.283	1.684±0.150	3.483±0.240	3.666±0.262	5.307±0.482	3.719±0.196	23.07±3.581	70.23±5.622
MTRL	7.010±0.567	2.197±0.101	9.825±0.749	3.386±0.282	1.659±0.155	3.540±0.323	3.682±0.232	5.232±0.436	3.708±0.181	22.95±4.051	69.57±4.612
FGL-MTFL	6.678±0.474	2.214±0.081	9.631±0.788	3.335±0.276	1.668±0.167	3.456±0.305	3.610±0.186	5.221±0.493	3.697±0.223	22.90±3.666	69.53±5.011

Method	LOGMEM		CLOCK		BOSNAM	ANART	DSPAN		DIGIT	nMSE
	IMMTOTAL	DELTOTAL	DRAW	COPYSCOR			FOR	BAC		
Robust	4.436±0.374	4.909±0.431	1.002±0.125	0.699±0.098	4.488±0.355	10.74±0.650	2.146±0.143	2.228±0.182	12.55±1.275	10.44±1.070*
CMTL	7.890±0.807	6.559±0.976	3.465±0.099	3.771±0.093	20.76±0.510	13.83±1.009	6.899±0.230	5.386±0.251	31.85±1.789	47.57±2.176*
Trace	4.649±0.487	4.994±0.645	1.134±0.255	0.939±0.306	5.595±1.363	10.97±0.858	2.313±0.360	2.331±0.227	13.80±1.025	11.88±1.516*
SRMTL	4.896±1.048	5.197±0.629	2.698±2.063	3.356±3.077	4.040±0.519	10.08±0.852	3.635±2.039	3.130±1.288	12.15±1.571	11.68±4.228*
p-MSSL	4.199±0.440	4.586±0.598	1.200±0.058	0.953±0.049	4.022±0.390	9.481±0.719	2.114±0.184	2.231±0.222	11.30±1.162	8.518±0.771*
G-SMuRFS	4.168±0.310	4.615±0.454	0.976±0.116	0.653±0.088	3.960±0.408	9.681±0.731	2.025±0.127	2.179±0.184	11.26±1.246	7.844±0.668*
MTRL	4.223±0.308	4.685±0.468	0.959±0.119	0.632±0.099	3.966±0.540	9.443±0.611	1.985±0.136	2.106±0.187	11.278±1.315	7.761±0.045
FGL-MTFL	4.139±0.366	4.566±0.505	0.973±0.102	0.654±0.076	3.984±0.485	9.471±0.708	1.997±0.149	2.119±0.191	11.17±1.241	**7.689±0.565**

Superscript symbols ∗ indicates that FGL-MTFL significantly outperformed that method on that score. Student's t-test at a level of 0.05 was used

The bold value indicates the best performance

processes: (1) all tasks are regularized by their mean value, and therefore knowledge obtained from one task can be utilized by other tasks via the mean value; (2) sparsity is enforced in the learning with ℓ_1 norm.

5. Multi-task Sparse Structure Learning from tasks parameters (p-MSSL) (Goncalves et al. 2014):
   p-MSSL $(\min_{\Theta,\Omega\succ 0} \; L(X, Y, \Theta) - \frac{k}{2}\log|\Omega| + \mathrm{Tr}(\Theta\Omega\Theta^T) + \lambda_1\|\Omega\|_1 + \lambda_2\|\Theta\|_1$, where $\Omega \in \mathbb{R}^{k\times k}$ is a matrix that captures the task relationship structure.
6. Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) (Yan et al. 2015):
   G-SMuRFS $(\min_{\Theta} \; L(X, Y, \Theta) + \lambda_1\|\Theta\|_{2,1} + \lambda_2 \sum_{l=1}^{q} w_l \sqrt{\sum_{j\in\mathcal{G}_l} \|\theta_{j.}\|_2})$ takes into account coupled features and group sparsity across tasks. Parameters for these three methods are set following the same approach as that used in the baseline comparable methods. Tables 8 and 9 show the results of the comparable MTL methods in terms of rMSE, CC, nMSE and wR.
7. Multi-task relationship learning (MTRL) (Zhang and Yeung 2012b):
   MTRL $(\min_{W,b,\Omega} \; \sum_{I=1}^{m} \frac{1}{n_i} \sum_{j=1} n_i (y_j^i - w_i^T x_j^i - b_i)^2 + \frac{\lambda_1}{2}\mathrm{tr}(WW^T) + \frac{\lambda_2}{2}\mathrm{tr}(W\Omega^{-1}W^T), \;\; \text{s.t. } \Omega \succeq 0, \mathrm{tr}(\Omega) = 1)$ can simultaneously model positive task correlation, describe negative task correlation, and identify outlier tasks based on the same underlying principle.

From the experimental results in Tables 8 and 9, we observe that the proposed method significantly outperforms the other recently developed algorithms. Specifically, the t-test results indicate that FGL-MTFL significantly outperforms all other methods significantly in terms of CC and all but FL-MTRL in terms of nMSE. Compared with the other multi-task learning that utilize different assumptions, G-SMuRFS and our proposed methods, both multi-task feature learning methods with sparsity-inducing norms, have an advantage. Since not all the brain regions are associated with AD, many of the features are irrelevant and redundant. Thus, sparse-based MTL methods involving FGL-MTFL and G-SMuRFS are more appropriate to predict cognitive measures with better performance than the non-sparse based MTL methods. Unlike G-SMuRFS, the group regularization $G_{2,1}$-norm in FGL-MTFL decouples the group sparse regularization across tasks to provide more flexibility.

Identification of Correlation among Cognitive Outcomes

A fundamental component of our FGL-MTFL is to estimate the relationship structure among tasks, thus promoting appropriate information sharing among related tasks. In this section, we investigate and evaluate the estimated task relationships. Figure 4 shows the normalized estimated correlation matrix $\mathcal{Z}$.

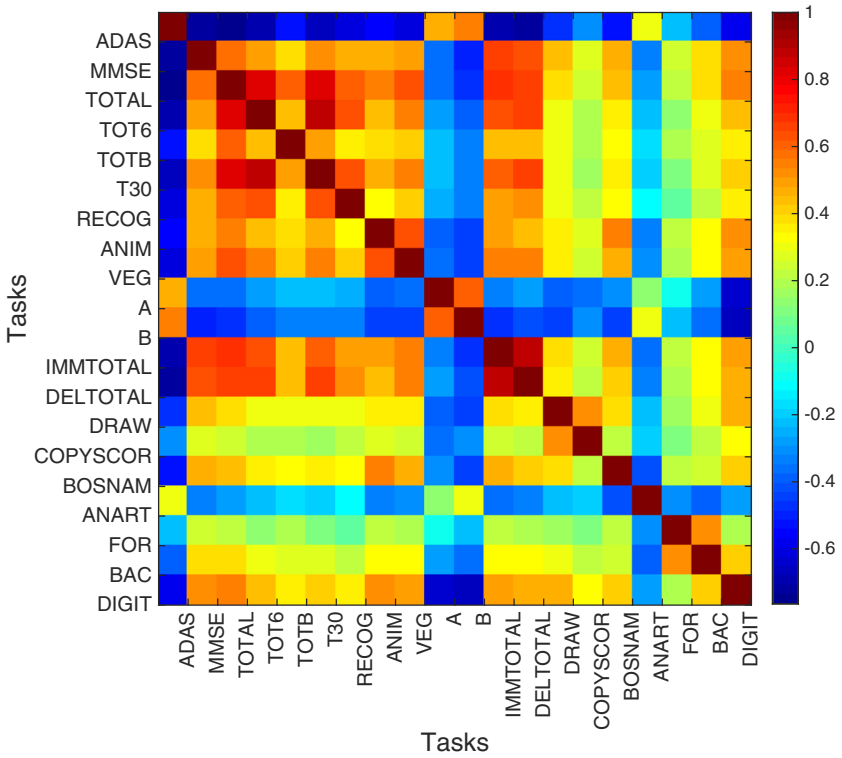


Fig. 4 Correlation matrix

Using the previous results obtained by our FGL-MTFL, the constructed graph $\mathbb{G}$ includes all pairwise links for each pair of tasks. The correlation of these links may be weak or not actually correlated due to bad estimation of the correlation matrix. To clearly analyze the influence of the estimated task relationships to the prediction performance of cognitive outcomes, we constructed graphs $\mathbb{G}$ that vary the value of the threshold τ based on the estimated correlation matrix $\mathcal{Z}$. The range of the τ value is 0.1, 0.3, 0.5, and 0.7. The graphs of these four examples are shown in Fig. 5. In this experiment, rather than including all pairwise links, a specific threshold value was applied to the correlation matrix $\mathcal{Z}$, and the performance with different threshold values is presented in Table 10. In this case, by thresholding the estimated correlation matrix with a higher τ, we can construct a sparse undirected graph to represent only the most reliable correlation. From the data shown in Fig. 5, we can find that with an increase in the threshold value, the graph of tasks becomes less dense. When $\tau = 0.1$, only one link is removed, and when τ increases to 0.3 and 0.5, the numbers of remaining links are only 133 and 48, respectively. When $\tau = 0.7$, there are only eight highly correlated cognitive scores. We found these remaining cognitive outcomes (including ADAS, MMSE, and RAVLT) are the most commonly used in the multi-task learning to predict cognitive outcomes (Yan et al. 2013,

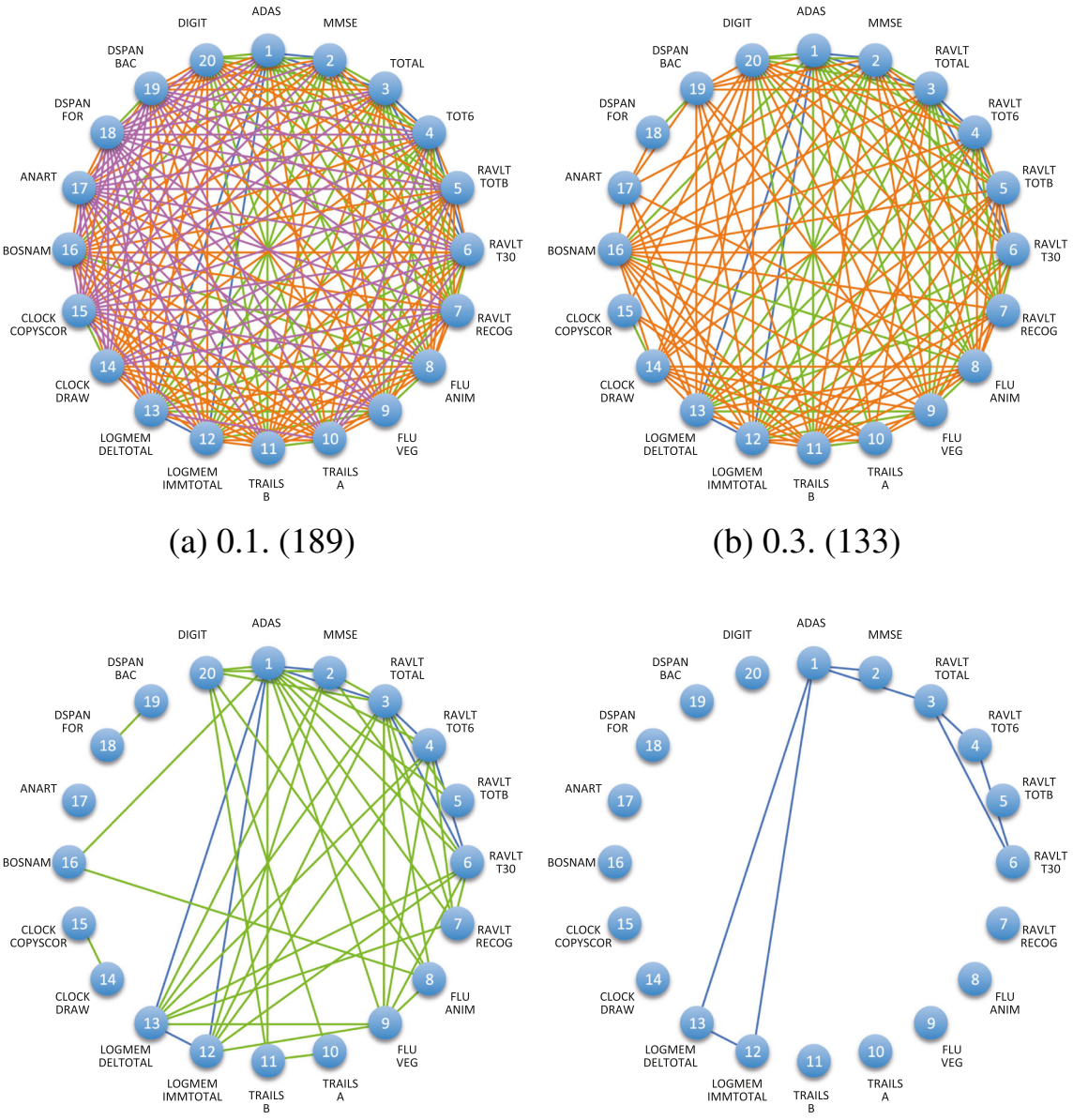


Fig. 5 The correlation graphs with four different threshold values

Table 10 The influence of threshold to the prediction performance for the fused lasso or fused group lasso regularized model

Method	τ	ADAS	MMSE	RAVLT					FLU		TRAILS	
				TOTAL	TOT6	TOTB	T30	RECOG	ANIM	VEG	A	B
FL-MTL	0.7	0.665±0.068	0.553±0.069	0.498±0.101	0.506±0.120	0.286±0.060	0.517±0.117	0.411±0.128	0.382±0.102	0.509±0.081	0.363±0.089	0.368±0.122
	0.5	0.662±0.066	0.550±0.070	0.488±0.103	0.507±0.119	0.308±0.068	0.516±0.116	0.410±0.119	0.375±0.104	0.509±0.085	0.355±0.093	0.358±0.121
	0.3	0.657±0.064	0.557±0.070	0.474±0.108	0.501±0.117	0.287±0.082	0.511±0.114	0.407±0.107	0.366±0.106	0.505±0.086	0.348±0.097	0.348±0.120
	0.1	0.641±0.074	0.522±0.095	0.491±0.112	0.472±0.102	0.287±0.085	0.472±0.096	0.409±0.085	0.356±0.107	0.485±0.076	0.352±0.085	0.386±0.159
	–	0.641±0.074	0.523±0.095	0.491±0.112	0.472±0.102	0.287±0.085	0.473±0.096	0.409±0.085	0.355±0.107	0.485±0.076	0.352±0.085	0.385±0.159
FGL-MTL	0.7	0.659±0.066	0.541±0.068	0.504±0.097	0.503±0.122	0.286±0.053	0.521±0.116	0.409±0.137	0.390±0.086	0.508±0.081	0.367±0.093	0.385±0.119
	0.5	0.660±0.064	0.538±0.067	0.496±0.099	0.506±0.120	0.307±0.064	0.521±0.117	0.412±0.130	0.388±0.098	0.513±0.081	0.361±0.096	0.372±0.119
	0.3	0.663±0.062	0.553±0.067	0.487±0.105	0.506±0.117	0.291±0.077	0.520±0.116	0.411±0.120	0.380±0.102	0.514±0.078	0.353±0.099	0.359±0.119
	0.1	0.662±0.060	0.556±0.065	0.479±0.109	0.501±0.116	0.287±0.080	0.517±0.113	0.410±0.116	0.371±0.106	0.512±0.077	0.347±0.101	0.351±0.119
	–	0.662±0.060	0.557±0.065	0.479±0.109	0.501±0.116	0.286±0.080	0.516±0.113	0.409±0.115	0.370±0.106	0.512±0.077	0.347±0.101	0.351±0.119
FGL-MTFL	0.7	0.667±0.062	0.549±0.061	0.517±0.118	0.500±0.127	0.320±0.081	0.516±0.117	0.421±0.126	0.380±0.083	0.505±0.082	0.414±0.085	0.479±0.098
	0.5	0.665±0.063	0.547±0.057	0.523±0.113	0.496±0.120	0.317±0.078	0.513±0.114	0.413±0.133	0.388±0.081	0.499±0.089	0.416±0.092	0.481±0.095
	0.3	0.667±0.069	0.549±0.069	0.522±0.113	0.498±0.121	0.327±0.069	0.512±0.110	0.427±0.123	0.400±0.080	0.503±0.087	0.419±0.087	0.482±0.095
	0.1	0.668±0.068	0.549±0.068	0.522±0.114	0.499±0.121	0.324±0.069	0.514±0.110	0.429±0.121	0.399±0.081	0.505±0.087	0.419±0.086	0.481±0.095
	–	0.668±0.068	0.549±0.068	0.522±0.114	0.499±0.121	0.324±0.069	0.514±0.110	0.429±0.121	0.399±0.081	0.505±0.087	0.419±0.086	0.481±0.095

Method	τ	LOGMEM		CLOCK		BOSNAM	ANART	DSPAN		DIGIT	wR
		IMMTOTAL	DELTOTAL	DRAW	COPYSCOR			For	BAC		
FL-MTL	0.7	0.510±0.092	0.534±0.104	0.369±0.075	0.226±0.099	0.478±0.088	0.156±0.093	0.108±0.120	0.187±0.137	0.456±0.065	0.404±0.048*
	0.5	0.503±0.093	0.529±0.105	0.378±0.073	0.226±0.096	0.480±0.096	0.142±0.087	0.117±0.116	0.193±0.134	0.448±0.059	0.403±0.047*
	0.3	0.496±0.095	0.524±0.108	0.372±0.089	0.218±0.100	0.470±0.103	0.119±0.078	0.121±0.112	0.191±0.102	0.441±0.056	0.396±0.048*
	0.1	0.470±0.097	0.514±0.108	0.379±0.093	0.221±0.102	0.442±0.084	0.094±0.104	0.083±0.102	0.177±0.108	0.414±0.083	0.383±0.046*
	–	0.470±0.097	0.514±0.108	0.379±0.093	0.221±0.102	0.442±0.085	0.094±0.104	0.083±0.102	0.177±0.108	0.414±0.083	0.383±0.046*
FGL-MTL	0.7	0.511±0.094	0.541±0.101	0.366±0.075	0.225±0.101	0.477±0.084	0.191±0.103	0.104±0.119	0.185±0.138	0.461±0.059	0.407±0.048*
	0.5	0.506±0.093	0.539±0.101	0.374±0.074	0.226±0.097	0.483±0.089	0.179±0.097	0.117±0.119	0.197±0.134	0.456±0.055	0.408±0.047*
	0.3	0.502±0.093	0.535±0.104	0.374±0.087	0.216±0.099	0.479±0.094	0.157±0.088	0.129±0.111	0.202±0.105	0.454±0.051	0.404±0.048*
	0.1	0.499±0.093	0.531±0.106	0.378±0.087	0.212±0.099	0.479±0.096	0.140±0.080	0.110±0.104	0.205±0.107	0.450±0.049	0.400±0.047*
	–	0.499±0.093	0.531±0.106	0.378±0.087	0.212±0.098	0.479±0.096	0.139±0.079	0.110±0.105	0.205±0.108	0.450±0.049	0.400±0.047*
FGL-MTFL	0.7	0.515±0.090	0.528±0.097	0.390±0.079	0.225±0.106	0.469±0.108	0.118±0.084	0.087±0.095	0.163±0.095	0.478±0.063	0.412±0.052*
	0.5	0.517±0.086	0.531±0.097	0.379±0.083	0.226±0.112	0.466±0.090	0.130±0.090	0.081±0.111	0.169±0.117	0.480±0.066	0.412±0.049*
	0.3	0.515±0.083	0.534±0.094	0.408±0.085	0.254±0.097	0.472±0.085	0.139±0.086	0.087±0.102	0.216±0.112	0.484±0.071	**0.421±0.053**
	0.1	0.514±0.084	0.533±0.094	0.408±0.083	0.252±0.092	0.473±0.089	0.137±0.083	0.096±0.089	0.218±0.113	0.484±0.070	**0.421±0.052**
	–	0.514±0.084	0.533±0.094	0.408±0.083	0.252±0.092	0.473±0.089	0.137±0.082	0.096±0.090	0.218±0.113	0.484±0.070	**0.421±0.052**

The first 20 columns are CC and the last one is wR. Superscript symbols ∗ indicates that FGL-MTFL significantly outperformed that method on that score. Student's t-test at a level of 0.05 was used
The bold value indicates the best performance

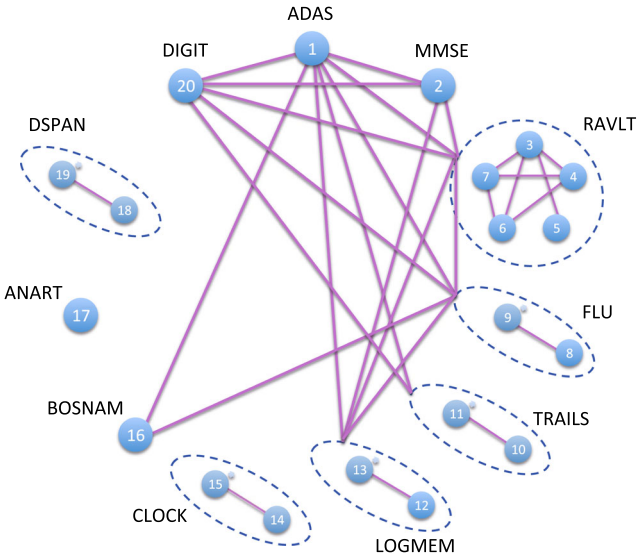


Fig. 6 The correlation graph when $\tau = 0.5$

2015; Wan et al. 2012; Li et al. 2012). Moreover, as shown in the Fig. 5d, ADAS is the most important of the cognitive outcome tests, with the most edges to other tasks, especially when the threshold is large, i.e., 13 links remains when $\tau = 0.5$ and 4 links remains when $\tau = 0.7$.

Additionally, we evaluated the performance of FL-MTL, FGL-MTL and FGL-MTFL methods in response to changing the thresholds in terms of CC and wR, as presented in Table 10. The symbol "–" indicates that no threshold was used. The performance of both FL-MTL and FGL-MTL increased with increased threshold value. However, FGL-MTFL achieved better results as the threshold value decreased. For values of τ less than 0.3, FGL-MTFL obtains the best result and is stable, which indicates that many edges do not contribute to performance in Fig. 5a.

In order to investigate the correlation of multiple cognitive tests, we constructed a graph to show the estimated correlation of inter-cognitive assessment tests, such as ADAS and MMSE, and intra-cognitive assessment tests, such as TOTAL and TOT6 in RAVLT. From Fig. 6 and Table 11, we can observe strong correlation of all intra cognitive assessment tests. Correlation analysis of inter-cognitive assessment tests shows that ADAS is an important test with strong correlations with other tests.

Table 12 Top 10 selected ROIs by MTFL, GL-MTFL and FGL-MTFL

Numbers	ROIs		
	MTFL	GL-MTFL	FGL-MTFL
1	L.MidTemporal	L.MidTemporal	L.MidTemporal
2	R.LatVent	R.LatVent	R.LatVent
3	R.InfParietal	R.InfParietal	R.InfParietal
4	R.Entorhinal	R.Entorhinal	WMHypoInt
5	R.PostCing	R.PostCing	R.Entorhinal
6	L.SupFrontal	WMHypoInt	L.SupFrontal
7	WMHypoInt	L.SupFrontal	L.Hippocampus
8	L.Hippocampus	L.Hippocampus	R.PostCing
9	R.BanksSTS	L.SupParietal	L.InfParietal
10	OpticChiasm	R.BanksSTS	L.IsthmCing

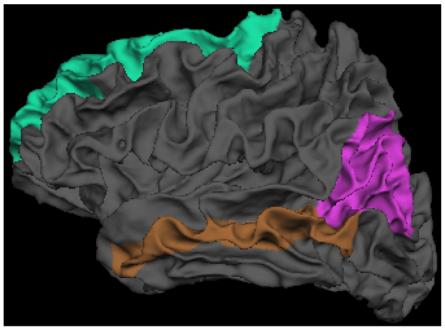
Identification of MRI Biomarkers

Key goals of studies of Alzheimer's disease are better cognitive score prediction and identification of which brain areas are more affected by the disease to help diagnose early stages of the disease and determine how it spreads. One focus of this work was the identification of MRI biomarkers. Our FGL-MTFL is a group sparse model that can identify a compact set of relevant neuroimaging biomarkers at the region level due to the group lasso on the features, allowing better interpretability of the brain region. The top ten selected MRI brain regions are shown in Table 12, as determined by calculating the overall weights for all cognitive tasks.

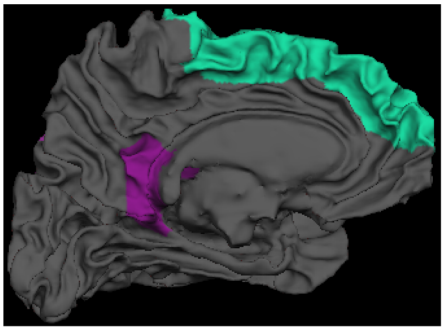
Some important brain regions are identified by our FGL-MTFL (see Fig. 7), such as Middle Temporal (Yan et al. 2015; Xu et al. 2016; Visser et al. 2002; Zhu et al. 2016), Hippocampus (Zhu et al. 2016) and Entorhinal (Yan et al. 2015), regions that are highly relevant to cognitive impairment. These findings are in accordance with current understanding of the pathological pathway of AD, and reports that these identified brain regions are highly related to clinical functions. For example, the hippocampus is

Table 11 The edge number of nodes (tasks) when $\tau = 0.5$

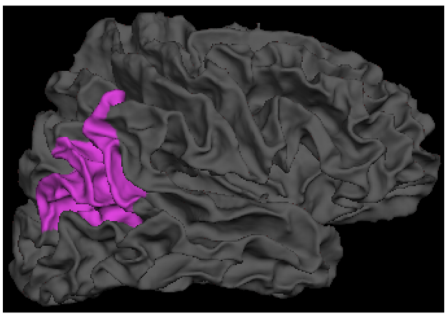
Tasks	ADAS	MMSE	RAVLT	FLU	TRAILS	LOGMEM	CLOCK	BOSNAM	ANART	DSPAN	DIGIT
points in set	1	1	5	2	2	2	2	1	1	2	1
Number of edges											
inside	–	–	7	1	1	1	1	–	–	1	–
outside	13	5	12	4	2	0	0	0	0	0	0
sum	13	5	19	5	3	1	1	0	0	1	0



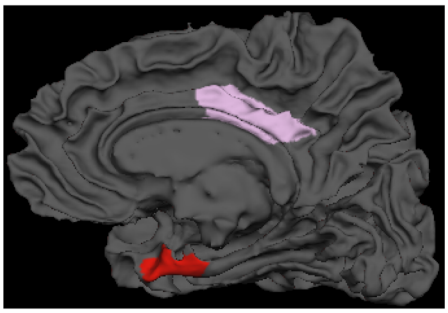
(a) Left-Hemisphere (outside): *Left InfParietal, Left MidTemporal, Left SupFrontal*.



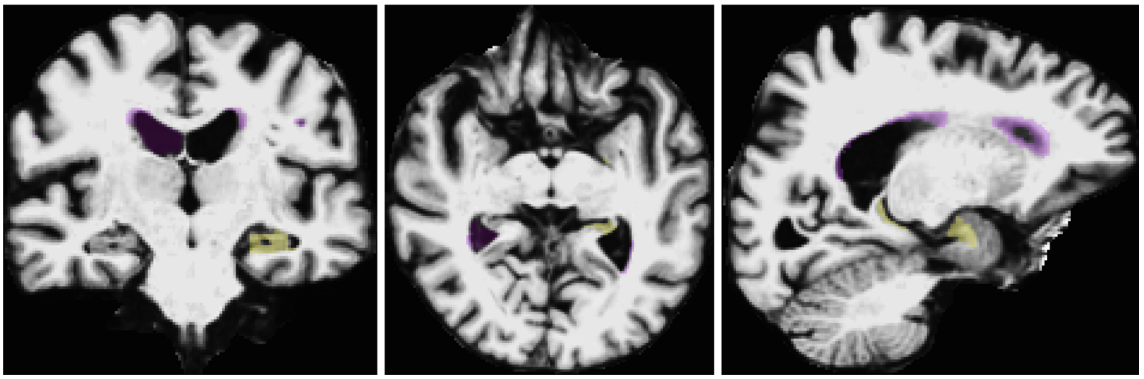
(b) Left-Hemisphere (inside): *Left SupFrontal, Left IsthmCing*.



(c) Right-Hemisphere (outside): *Right InfParietal*.



(d) Right-Hemisphere (inside): *Right PostCing, Right Entorhinal*.



(e) Subcortical: *Left Hippocampus, Right LatVent, WMHypoInt*.

Fig. 7 [Best viewed in color] Plots show the top 10 ROI's selected by FGL-MTFL. These were the most relevant areas for predicting all cognitive scores jointly

located in the temporal lobe of the brain and participates in memory and spatial navigation. The Entorhinal cortex is the first area of the brain to be affected in Alzheimer's disease, and is typically subjected to the most heavy damage with the progression of Alzheimer's disease (Hoesen et al. 1991).

Conclusion

Many clinical/cognitive measures have been designed to evaluate patient cognitive status for use as criteria for clinical diagnosis of probable AD. In this paper, we propose a multi-task learning framework for predictive modeling of cognitive measures based on MRI data from ADNI. The existing MTL approach neglects the relationships between outcomes and between features. We consider the multi-task learning problem assuming unequal correlation of the tasks and effects of different correlated tasks on different brain regions. Based on the intuitive motivation that tasks should be related to a group of features, we exploited the global task-common structure as well as task-ROI specific structure, and present a novel fused group lasso regularized

multi-task learning method (FGL-MTFL). Experiments and comparisons of this model with baseline methods illustrate that FGL-MTFL offers consistently better performance than several currently applied multi-task learning algorithms. In the current work, only a priori group information is incorporated into the multi-task predictive model, but there is no ability to automatically learn the feature groups. In future work, we will investigate other structures in features, such as graph structure, which can provide additional insights to understand and interpret data. Our current work is based on linear methods, but kernel methods can model the cognitive scores as nonlinear functions of neuroimaging measures. Recently, many kernel-based classification or regression methods with faster optimization speed or stronger generalization performance have been investigated by theoretically and experimentally. Our future work will focus on kernel-based multi-task learning to better capture the complex but more flexible relationship between cognitive scores and the neuroimaging measures.

Information Sharing Statement

Source code is available at: https://bitbucket.org/XIAOLILIU/fgl-mtfl. Both the source code and documentation are available on request. Contact: neuxiaoliliu@gmail.com and caopeng@cse.neu.edu.cn.

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

References

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

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

Batsch, N.L., & Mittelman, M.S. (2015). World Alzheimer Report 2012. Overcoming the stigma of dementia. Alzheimer's Disease International (ADI), p. 5.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundation and Trends in Machine Learning*, *3*(1), 1–122.

Cai, J.-F., Osher, S., Shen, Z. (2009). Split bregman methods and frame based image restoration. *Multiscale modeling & simulation*, *8*(2), 337–369.

Cao, P., Liu, X., Yang, J., Zhao, D., Zaiane, O. (2017). Sparse multi-kernel based multi-task learning for joint prediction of clinical scores and biomarker identification in alzheimer's disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 195–202*.

Caruana, R. (1998). Multitask learning. In *Learning to learn. Springer, pp. 95–133*.

Castellani, R.J., Rolston, R.K., Smith, M.A. (2010). Alzheimer disease. *Disease-a-month: DM*, *56*(9), 484.

Chen, J., Zhou, J., Ye, J. (2011). Integrating low-rank and group-sparse structures for robust multi-task learning.

Dale, A.M., Fischl, B., Sereno, M.I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*, 179–194.

Dale, A.M., & Sereno, M.I. (1993). Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. *Journal of Cognitive Neuroscience*, *5*(2), 162–176.

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

Evgeniou, T., & learning, M.P. (2004). Regularized multi–task. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 109–117*.

Fischl, B., Liu, A., Dale, A.M. (2001). Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging*, *20*, 70–80.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*, 341–355.

Fischl, B., Salat, D.H., van der Kouwe, A.J., Makris, N., Segonne, F., Quinn, B.T., Dale, A.M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, *23*, S69–S84.

Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, *6*(2), 67–77.

Goncalves, A., Das, P., Chatterjee, S., Sivakumar, V., Zuben, F.J.V., Banerjee, A. (2014). Multi-task sparse structure learning. In *In CIKM, pp. 451–460*.

Jebara, T. (2011). Multitask sparsity via maximum entropy discrimination. *Journal of Machine Learning Research*, *12*(Jan), 75–110.

Ji, S., & Ye, J. (2009). An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning. ACM, pp. 457–464*.

Liu, J., Ji, S., Ye, J. (2009). Multi-task feature learning via $\ell_{2,1}$-norm minimization. In *Proceedings of the 25th conference on uncertainty in artificial intelligence. AUAI Press, pp. 339–348*.

Liu, X., Cao, P., Zhao, D., Zaiane, O., et al. (2017). Group guided sparse group lasso multi-task learning for cognitive performance prediction of alzheimer's disease. In *International Conference on Brain Informatics. Springer, pp. 202–212*.

Liu, X., Goncalves, A.R., Cao, P., Zhao, D., Banerjee, A., et al. (2017). Modeling Alzheimer's disease cognitive scores using multi-task sparse group lasso. *Computerized Medical Imaging and Graphics*, *66*, 100–114.

Reuter, M., Rosas, H.D., Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, *53*(4), 1181–1196.

Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, *22*, 1060–1075.

Ségonne, F., Pacheco, J., Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Transactions on Medical Imaging*, *26*(4), 518–529.

Li, S., Saykin, A.J., Risacher, S.L., Kim, S., Fang, S., Rao, B.D., Li, T., Yan, J., Zhang, Z., Wan, J. (2012). Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer.

Sled, J.G., Zijdenbos, A.P., Evans, A.C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, *17*, 87–97.

Stonnington, C.M., Chu, C., Klöppel, S., Jack, C.R., Ashburner, J., Frackowiak, R.S.J. (2010). Alzheimer disease neuroimaging initiative predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *NeuroImage*, *51*(4), 1405–1413.

Hoesen, G.W.v., Hyman, B.T., Damasio, A.R. (1991). Entorhinal cortex pathology in Alzheimer's disease. *Hippocampus*, *1*(1), 1–8.

Visser, P.J., Verhey, F.R.J., Hofman, P.A.M., Scheltens, P., Jolles, J. (2002). Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. *Journal of Neurology Neurosurgery & Psychiatry*, *72*(4), 491–497.

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

Wan, J., Zhang, Z., Yan, J., Li, T., Rao, B.D., Fang, S., Kim, S., Risacher, S.L., Saykin, A.J., Li, S. (2012). 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), pp. 940–947*.

Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Li, S. (2011). ADNI Sparse Multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *International Conference on Computer Vision, pp. 6–13*.

Wang, H., Nie, F., Huang, H., Yan, J., Kim, S., Risacher, S., Saykin, A., Li, S. (2012). High-order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction. In *Advances in Neural Information Processing Systems, pp. 1277–1285*.

Weiner, M.W., Aisen, P.S., Jack, C.R. Jr., Jagust, W.J., Trojanowski, J.Q., Shaw, L., Saykin, A.J., Morris, J.C., Cairns, N., Beckett, L.A., Toga, A., Green, R., Walter, S., Soares, H., Snyder, P., Siemers, E., Potter, W., Cole, P.E., Schmidt, M. (2010). The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia*, *6*, 202–211.

Xu, L., Wu, X., Li, R., Chen, K., Long, Z., Zhang, J., Guo, X., Yao, L. (2016). Prediction of progressive mild cognitive impairment by multi-modal neuroimaging biomarkers. *Journal of Alzheimer's Disease*, *51*(4), 1045–1056.

Xue, Y., Liao, X., Carin, L., Krishnapuram, B. (2007). Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, *8*(Jan), 35–63.

Yan, J., Huang, H., Risacher, S.L., Kim, S., Inlow, M., Moore, J.H., Saykin, A.J., Shen, L. (2013). Network-guided sparse learning for predicting cognitive outcomes from MRI measures. In *International Workshop on Multimodal Brain Image Analysis. Springer, pp. 202–210*.

Yan, J., Li, T., Wang, H., Huang, H., Wan, J., Nho, K., Kim, S., Risacher, S.L., Saykin, A.J., Shen, L., et al. (2015). Cortical surface biomarkers for predicting cognitive outcomes using group $\ell_{2,1}$ norm. *Neurobiology of aging*, *36*, S185–S193.

Ye, G.-B., & Xie, X. (2011). Split bregman method for large scale fused lasso. *Computational Statistics & Data Analysis*, *55*(4), 1552–1569.

Yu, K., Tresp, V., Schwaighofer, A. (2005). Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd international conference on Machine learning. ACM, pp. 1012–1019*.

Yuan, L., Liu, J., Ye, J. (2013). Efficient methods for overlapping group lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(9), 2104–2116.

Zhang, D., Shen, D., Alzheimer's Disease Neuroimaging Initiative (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, *59*(2), 895–907.

Zhang, Y., & Yeung, D.-Y. (2012a). A convex formulation for learning task relationships in multi-task learning. In *Conference on Uncertainty in Artificial Intelligence (UAI2010) 2010, pp. 733–742*.

Zhang, Y., & Yeung, D.-Y. (2012b). A convex formulation for learning task relationships in multi-task learning. arXiv:1203.3536.

Zhou, J., Chen, J., Ye, J. (2011). Clustered multi-task learning via alternating structure optimization. In *Advances in neural information processing systems, pp. 702–710*.

Zhou, J., Liu, J., Narayan, V.A., Ye, J., Alzheimer's Disease Neuroimaging Initiative. (2013). Modeling disease progression via multi-task learning. *NeuroImage*, *78*, 233–248.

Zhou, J.Y. Multi-task learning in crisis event classification. Technical report, Tech. Rep., http://www.public.asu.edu/jzhou29.

Zhu, X., Suk, H.-I., Lee, S.-W., Shen, D. (2016). Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. *IEEE Transactions on Biomedical Engineering*, *63*(3), 607–618.