Computerized Medical Imaging and Graphics 66 (2018) 100–114

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Modeling Alzheimer's disease cognitive scores using multi-task sparse group lasso

Xiaoli Liu[a,b,c], André R. Goncalves[d], Peng Cao[a,*], Dazhe Zhao[a,b], Arindam Banerjee[c], Alzheimer's Disease Neuroimaging Initiative

[a] College of Computer Science and Engineering, Northeastern University, Shenyang, China
[b] Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang, China
[c] Computing Science & Engineering, University of Minnesota, Twin Cities, USA
[d] Center for Research and Development in Telecommunications (CPqD), Brazil

ARTICLE INFO

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

ABSTRACT

Alzheimer's disease (AD) is a severe neurodegenerative disorder characterized by loss of memory and reduction in cognitive functions due to progressive degeneration of neurons and their connections, eventually leading to death. In this paper, we consider the problem of simultaneously predicting several different cognitive scores associated with categorizing subjects as normal, mild cognitive impairment (MCI), or Alzheimer's disease (AD) in a multi-task learning framework using features extracted from brain images obtained from ADNI (Alzheimer's Disease Neuroimaging Initiative). To solve the problem, we present a multi-task sparse group lasso (MT-SGL) framework, which estimates sparse features coupled across tasks, and can work with loss functions associated with any Generalized Linear Models. Through comparisons with a variety of baseline models using multiple evaluation metrics, we illustrate the promising predictive performance of MT-SGL on ADNI along with its ability to identify brain regions more likely to help the characterization Alzheimer's disease progression.

1. Introduction

Alzheimer's disease (AD) is a severe neurodegenerative disorder that results in a loss of mental function due to the deterioration of brain tissue, leading directly to death (Khachaturian, 1985). It accounts for 60–70% of age related dementia, affecting an estimated 30 million individuals in 2011 and the number is projected to be over 114 million by 2050 (Wimo et al., 2003). The cause of AD is poorly understood and currently there is no cure for AD. AD has a long preclinical phase, lasting a decade or more. There is increasing research emphasis on detecting AD in the pre-clinical phase, before the onset of the irreversible neuron loss that characterizes the dementia phase of the disease, since therapies/treatment are most likely to be effective in this early phase. The Alzheimer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/) has been facilitating the scientific evaluation of neuroimaging data including magnetic resonance imaging (MRI), positron emission tomography (PET), along with other biomarkers, clinical and neuropsychological assessments for predicting the onset and progression of MCI (mild cognitive impairment) and AD. Early diagnosis of AD is key to the development, assessment, and monitoring of new treatments for AD.

Recently, rather than predicting categorical variables as in classification, several studies begin to estimate continuous clinical variables from brain images. Therefore, instead of classify a subject into binary or multiple pre-determined categories or stages of the disease, regression focus on estimating continuous values which may help to assess patient's disease progression. The most commonly used cognitive measures are Alzheimer's Disease Assessment Scale cognitive total score (ADAS), Mini Mental State Exam score (MMSE) and Rey Auditory Verbal Learning Test (RAVLT). Regression analyses were commonly used to predict cognitive scores from imaging measures. The relationship between commonly used cognitive measures and structural changes with MRI has been previously studied by regression models and the results demonstrated there exist a relationship between baseline MRI features and cognitive measures (Wan et al., 2014; Stonnington et al., 2010). For example, Wan et al. has proposed an elegant regression model called CORNLIN that employs a sparse Bayesian learning algorithm to predict multiple cognitive scores based on 98 structural MRI regions of interests (ROIs) for Alzheimer's disease patients. The polynomial model used in CORNLIN can detect either a nonlinear or linear relationship between brain structure and cognitive decline (Wan et al., 2014). Stonnington et al. adopted relevance vector

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

https://doi.org/10.1016/j.compmedimag.2017.11.001
Received 29 October 2016; Received in revised form 10 June 2017; Accepted 14 November 2017

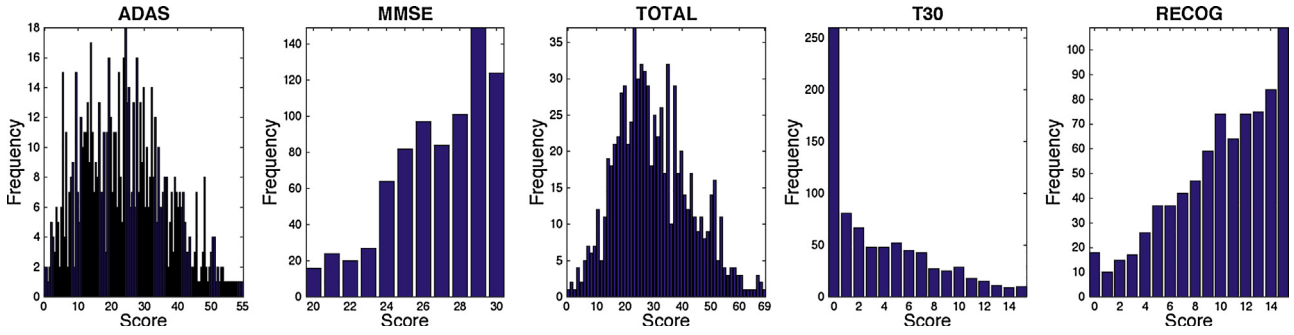


Fig. 1. Response profile (histograms) for cognitive scores (tasks) considered.

regression, a sparse kernel method formulated in a Bayesian framework, to predict four sets of cognitive scores using MRI voxel based morphometry measures (Stonnington et al., 2010). One of the biggest challenges in the prediction of inferring cognitive outcomes with MRI is the high dimensionality, which affects the computational performance and leads to a wrong estimation and identification of the relevant predictors. Sparse methods have attracted a great amount of research efforts in the neuroimaging field to reduce the high dimensionality and identify the relevant biomarkers due to its sparsity-inducing property. Ye et al. applied sparse logistic regression with stability selection to ADNI (Alzheimer's Disease Neuroimaging Initiative) data for robust feature selection (Ye et al., 2012), successfully predicted the conversion from MCI to probable AD and identified a small subset of bio-signatures. Recently, the multi-task learning (MTL) based feature learning methods with sparsity-inducing norm have been widely studied to select the discriminative feature subset from MRI features by incorporating inherent correlations among multiple clinical cognitive measures (Zhou et al., 2013; Wang et al., 2011; Zhang and Shen, 2012). For example, the $\ell_{2,1}$-norm regularization penalizes each row of parameters matrix as a whole and enforce sparsity among the rows, it is able to select the most discriminative features. Wang et al. (2011) and Zhang and Shen (2012) employed multi-task feature learning strategies for selecting biomarkers that could predict multiple clinical scores. Specially, Wang et al. (2011) considers some important features are only correlated to a subset of tasks, and adds an ℓ_1-norm regularizer to impose the sparsity among all elements and propose to use the combined $\ell_{2,1}$-norm and ℓ_1-norm regularizations to select features; Zhang proposed a multi-task learning with $\ell_{2,1}$-norm to select the common subset of relevant features for multiple variables from each modality by assuming that the related tasks share a common relevant feature subset. The most limitation of the popular learning models assume linear relationship between the MRI features and the cognitive outcomes. To model these more complicated but more flexible relationship between them, Zhang develop a multi-modal support vector regression (SVR) to fuse the above-selected features from all modalities with the selected feature subset (Zhang and Shen, 2012). Kernel methods have been studied to model the cognitive scores as nonlinear functions of neuroimaging measures. Recently, many kernel based classification or regression methods with faster optimization speed or stronger generalization performance have been proposed and investigated by theoretically analyzing and experimentally evaluating (Gu and Sheng, 2016; Gu et al., 2015). Suk et al. proposed a new sparse multi-task learning with an $\ell_{2,1}$-norm regularization (Suk et al., 2016). The multi-task learning is unlike the conventional multi-task learning methods, which treat all features equally. It utilizes the optimal regression coefficients learned in the lower hierarchy as context information to weight features adaptively. Most existing studies focus on only inferring the cognitive outcomes on single time-point of data (cross-sectional analysis), Ye et al. formulate the prediction problem as a multi-task regression problem by considering the prediction at each time point as a task, and propose a convex formulation with fused sparse group Lasso. The formulation allows to the simultaneous selection of a common set of biomarkers at all time points with ℓ_1-norm as well as the selection of a specific set of biomarkers at different time points with ℓ_1-norm, and in the meantime incorporates the temporal smoothness using the fused lasso penalty (Zhou et al., 2013).

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

A recent study proposed a prior knowledge guided regression model, using the group information to enforce the intra-group similarity with group sparse methods. In recent work, these existing ideas have been combined in Group-sparse Multitask Regression and Feature Selection (G-SMuRFS) (Yan et al., 2015; Wang et al., 2012) which takes into account coupled feature and group sparsity across tasks and uses vertex-based cortical surface measures in an anatomically meaningful manner. Since brain structures tend to work together to achieve a certain function, brain imaging measures are often correlated with each other. It assumes (1) possible partition exists among predictors, and (2) predictors within one partition should have similar weights. However, there exists three limitations of G-SMuRFS: (1) G-SMuRFS allows to learn a common subset of brain regions across all the tasks simultaneously with a Group $\ell_{2,1}$-norm. This assumption is too restrictive since different tasks may prefer different brain regions. It is desirable to select the specific ROIs for different tasks. (2) All scores are modeled with Gaussian (least squares) regression in G-SMuRFS, whereas it is not appropriate for all the scores. From Fig. 1, it can be seen that the distribution of scores of TOTAL and ADAS are Gaussian and three scores (T30, RECOG and MMSE) are Poisson. (3) The optimization of G-SMuRFS was done based on an iterative alternative optimization (AO) algorithm, which is an approximate gradient (not sub-gradient) descent method to handle sparse coefficient blocks and results in an inaccurate solution.

In order to solve these limitations, we propose a multi-task sparse group lasso (MT-SGL) method which encourages individual feature selection coupled with group selection with sparsity-inducing norm. Instead of learning a shared representation from the level of feature and region across all the tasks simultaneously, the MT-SGL formulation which encourages (a) individual feature selection based on the utility of the features across all tasks and (b) task specific group selection based on the utility of the group to decouple the ROIs sharing across tasks allowing for more flexibility. Moreover, the proposed MT-SGL framework can use general loss functions, including losses derived from

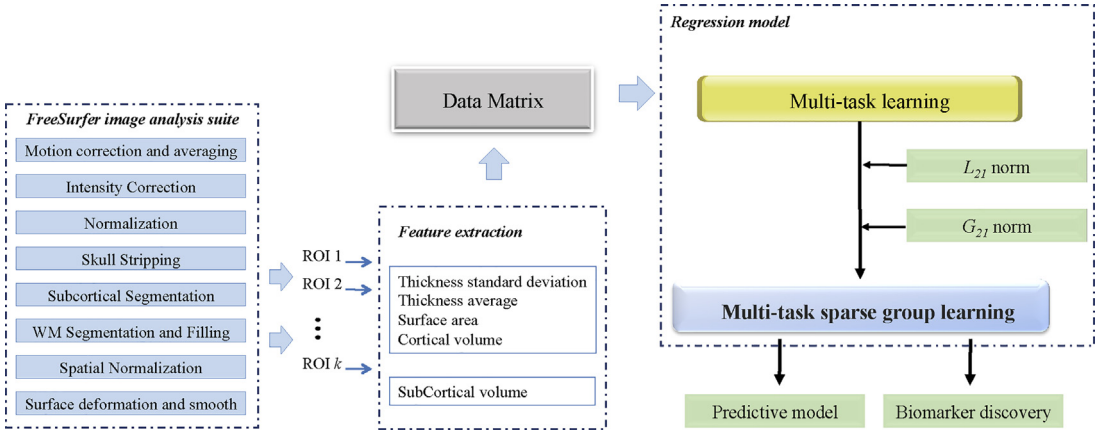


Fig. 2. Flow chart of the proposed MT-SGL method. Given a set of MRI images, we preprocess the images and parcellate a brain into ROIs. From each ROI, multiple features are extracted to measure the atrophy information involving cortical thickness, surface area and volume from gray matters and white matters in this study. Two sparsity-inducing norm ($\ell_{2,1}$ and $G_{2,1}$) regularizations are incorporated into our MTL method to model the task relatedness and group structure of features. Our framework not only provides cognitive scores prediction, but also identifies which are the brain areas more affected by the disease.

generalized linear models (GLMs). In our experiments, we consider MT-SGL models corresponding to Gaussian regression (least squares) as well as Poisson regression, inspired by the response profiles of some cognitive scores. Fig. 2 illustrates a schematic diagram of the proposed framework for cognitive score prediction and biomarker discovery.

The proposed formulation is, however, challenging to solve since the structured sparsity-inducing norms are non-smooth. In order to solve the new objective function, we consider two different approaches: proximal averaging, which takes the average the solutions from the proximal operator for the individual regularizers and has provable guarantees of convergence (Bauschke et al., 2008; Yu, 2013a); and proximal composition, which the proximal operator for the composite regularizer is the composition of the proximal operators for individual regularizers (Yu, 2013b). We consider accelerated versions of these methods based on suitable FISTA-style (Beck and Teboulle, 2009) application of accelerated gradient descent. Compared with the optimization algorithm in G-SMuRFS, the AGM leads to a fast and correct algorithm for the optimization.

Through empirical evaluation and comparison with five different baseline methods on data from ADNI, we illustrate that MT-SGL outperforms other baseline methods, including ridge regression, lasso, group lasso (Yuan and Lin, 2006) applied independently to each task, and multi-task group lasso (MT-GL) based on $\ell_{2,1}$-norm regularization (Liu et al., 2009). Improvements are statistically significant for most scores (tasks). MT-SGL showed similar results to G-SMuRFS, although MT-SGL has an efficient optimization method, besides having more general formulation which allows it to tackle a wider spectrum of problems.

We also present a discussion on the top ROIs identified by MT-SGL, that is, the ROIs that mostly explain the scores. We found that the selected ROIs corroborate with studies in neuroscience (Devanand et al., 2007; de Toledo-Morrell et al., 2004) as the areas of the brain that are more affected by the Alzheimer's disease. It indicates that MT-SGL can be a useful tool to guide further investigation on the ROIs pointed by the algorithm.

The rest of the paper is organized as follows. Section 2 discusses the MT-SGL formulation and optimization strategies are presented in Section 3. Experimental analysis is performed in Section 4 and results are compared with baseline methods. We conclude in Section 5.

2. Multi-task sparse group lasso

To identify the correlations between cognitive performance scores and MRI features, the linear (least square) regression method is a standard way in medical image analysis research. One of the biggest challenge in the prediction of inferring cognitive outcomes with MRI is the high dimensionality, which affects the computational performance and leads to a wrong estimation and identification of the relevant predictors. Sparse methods have attracted a great amount of research efforts in the neuroimaging field to reduce the high dimensionality and identify the relevant biomarkers due to its sparsity-inducing property. Moreover, the multi-task learning (MTL) methods with sparsity-inducing norm based on MRI features have been widely studied to investigate the prediction power of neuroimaging measures by incorporating inherent correlations among multiple clinical cognitive, and it has been commonly used to obtain better generalization performance than learning each task individually. It is known that there exist inherent correlations among multiple clinical cognitive variables of a subject. However, many works do not model dependence relation among multiple tasks and neglect the correlation between clinical tasks which is potentially useful. When the tasks are believed to be related, learning multiple related tasks jointly can improve performance relative to learning each task separately. The proposed work on multi-task sparse group lasso (MT-SGL) builds on the existing literature on linear regression models with sparsity structures over the regression coefficients. Our work, on the other hand, builds on the literature on sparse multi-task learning (Argyriou et al., 2007; Evgeniou and Pontil., 2004), which encourages related tasks to have similar sparsity structures.

We start with a basic description of the MT-SGL model. Consider a multi-task learning (MTL) setting with k tasks. Let p be the number of covariates, shared across all the tasks, and 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 $\theta_{\cdot h} \in \mathbb{R}^p$ corresponding to task h, $h = 1, \ldots, k$, and row $\theta_{i\cdot} \in \mathbb{R}^k$ corresponding to feature i, $i = 1, \ldots, p$.

The MTL problem can be set-up as one of estimating the parameters based on suitable regularized loss function:

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

where $L(\cdot)$ denotes a convex loss function and $R(\cdot)$ is a convex and possibly nonsmooth regularization function. In the context of least squares regression, for example, the loss function is defined as follows,

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

where $\mathbf{y}_i \in \mathbb{R}^{1\times k}$, $\mathbf{x}_i \in \mathbb{R}^{1\times p}$ are the ith rows of Y, X, respectively, corresponding to the multi-task response and covariates for the ith sample. We note that the MTL framework can be easily extended to other loss

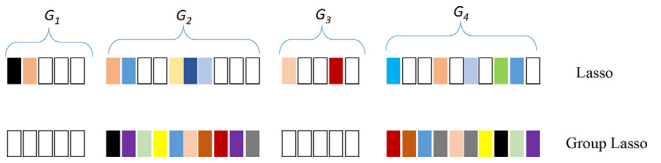


Fig. 3. The difference between lasso and group lasso. The different colors in the square boxes indicate the weights of the features and white color means zero-valued elements. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

functions, especially losses corresponding to generalized linear models (GLMs) (Nelder and Baker, 1972). In particular, based on the response profile for some tasks (see Fig. 1 in Section 4.3), one could consider loss functions based Poisson regression given by,

$$L_{\text{Poisson}}(\Theta) = [[\exp(X\Theta)]] - \langle X\Theta, Y\rangle, \tag{3}$$

where $[[\exp(\cdot)]]$ denotes sum over element-wise exponentiation of the matrix argument.

For the MTL regularization $R(\Theta)$, different choices encourage different structures in the estimated parameters, e.g., unstructured sparsity (Lasso) with $R(\Theta) = \|\Theta\|_1$, feature sparsity with $R(\Theta) = \|\Theta\|_{2,1}$ (Liu et al., 2009) and structured sparsity (Yuan and Lin, 2006). Group regularizers like group lasso (Yuan and Lin, 2006) via an $\ell_{2,1}$ regularization assumes covarying variables in groups, and have been extensively studied in the multi-task feature learning. The regularization $\ell_{2,1}$-norm ($R(\Theta) = \|\Theta\|_{G_{2,1}}$), $\|\Theta\|_{G_{2,1}}$ uses the ℓ_2-norm within a group and the ℓ_1-norm between groups. The difference of lasso and group lasso is illustrated in Fig. 3. The key assumption behind the group lasso regularizer is that if a few features in a group are important, then most of the features in the same group should also be important. Group lasso regularized multi-task learning (MT-GL) aims to improve the generalization performance by exploiting the shared features among tasks (Liu et al., 2009; Gong et al., 2012). It can identify important biomarkers, which potentially play the key roles in memory and cognition circuitry. The MT-GL algorithm and its extensions have been successfully applied to capture the biomarkers having affects across most or response in the application of AD prediction, since multiple cognitive assessment scores are essentially influenced by the same underlying pathology and only a subset of brain regions are relevant to these scores (Guerrero et al., 2017; Zhu et al., 2016; Yan et al., 2015). The MT-GL model via the $\ell_{2,1}$-norm regularization considers

$$R(\Theta) = \left\|\Theta\right\|_{2,1} = \sum_{i=1}^{k} \left\|\theta_{i.}\right\|_2, \tag{4}$$

and is suitable for simultaneously enforcing sparsity over features for all tasks.

We assume the p covariates to be divided into q disjoint groups $\mathcal{G}_\ell$, $\ell = 1, \ldots, q$, with each group having m_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, m_ℓ, ranges from 1 to 4, and the number of groups q can be in the hundreds. Then we introduce a $G_{2,1}$-norm according to the relationship between the brain regions (ROIs) and cognitive tasks and encourage a task-specific subset of ROIs. The $G_{2,1}$-norm $\|\Theta\|_{G_{2,1}}$ is defined as:

$$\left\|\Theta\right\|_{G_{2,1}} = \sum_{\ell=1}^{q} \sum_{h=1}^{k} w_\ell \left\|\theta_{\mathcal{G}_\ell,h}\right\|_2. \tag{5}$$

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

Plugging $G_{2,1}$-norm and $\ell_{2,1}$-norm to the formulation in Eq. (1), the objective function of multi-task sparse group lasso (MT-SGL) is given in the following optimization problem:

$$\min_{\Theta\in\mathbb{R}^{p\times k}} L(\Theta; Y, X) + \lambda_1 \left\|\Theta\right\|_{2,1} + \lambda_2 \left\|\Theta\right\|_{G_{2,1}}. \tag{6}$$

where $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ are the regularization parameters.

MT-SGL encourages (a) *individual feature selection* based on the utility of the features *across all tasks* with $\ell_{2,1}$-norm and (b) *task specific group selection* based on the utility of the group with $G_{2,1}$-norm, i.e., brain regions of interest (ROI) for that task. Unlike basic SGL for regression (Chatterjee et al., 2012; Liu and Ye., 2010; Friedman et al., 2010), MT-SGL has a parameter coupling across tasks because of $\|\theta_{j.}\|_2$ which encourages simultaneous sparsity across tasks for individual feature selection. Further, in the proposed MT-SGL, the group sparsity as determined by $\|\Theta\|_{G_{2,1}}$ is task specific, so that different tasks can use different groups if needed.

The proposed MT-SGL framework is related to the recently proposed G-SMuRFS (Yan et al., 2015; Wang et al., 2012) with three key differences: (i) unlike G-SMuRFS, MT-SGL regularization decouples the group sparse regularization across tasks allowing for more flexibility; (ii) MT-SGL allows the loss function to be based on generalized linear models (GLMs), rather than just square loss which corresponds to a Gaussian model, and (iii) the optimization in MT-SGL is done using FISTA (Beck and Teboulle, 2009) which leads to a fast and correct algorithm for the optimization. The motivation behind considering GLMs is that the responses in the context of AD are often non-Gaussian variables, e.g., the number of words an individual can remember after half hour, which can be potentially better modeled by a Poisson distribution or other distributions over discrete counts. We will study the effectiveness of using a GLM based MT-SGL in Section 4. Further, the formulation makes MT-SGL applicable to more general problems and data types. Further, while G-SMuRFS (Yan et al., 2015; Wang et al., 2012) considers a related model, the optimization was done based on an approximate gradient (not sub-gradient) descent method to handle sparse coefficient blocks. In contrast, we directly use an accelerated method based on FISTA (Beck and Teboulle, 2009) which is provably correct and faster. The difference between the formulations of MT-SGL and G-SMuRFS is illustrated in Fig. 4.

3. Efficient optimization for MT-SGL

The optimization problem for MT-SGL as in (6) is a convex optimization problem with a composite objective with a smooth term corresponding to the square loss and a non-smooth term corresponding to the regularizer. The composite minimization problem where the objective consists of a smooth loss function and a sum of nonsmooth functions, have received increasing attention due to the arise of structured sparsity (Bach et al., 2012), such as the graph-guided fused lasso (Kim and Xing, 2009), fused sparse group lasso (Zhou et al., 2013) and some others. These structured regularizers although greatly enhance our modeling capability, introduce significant new computational challenges as well (Yu, 2013a). In this section, we present a FISTA-style (Beck and Teboulle, 2009) algorithm for efficiently solving the MT-SGL problem.

Consider a general convex optimization problem with a composite objective given by

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}), \tag{7}$$

where $\mathbf{x} \in \mathbb{R}^d$, f: $\mathbb{R}^d \mapsto \mathbb{R}$ is a smooth convex function of the type $C^{1,1}$, i.e., continuously differentiable with Lipschitz continuous gradient so that $\|f(\mathbf{x}) - f(\mathbf{w})\| \leq \kappa \|\mathbf{x} - \mathbf{w}\|$ where κ denotes the Lipschitz constant, and g: $\mathbb{R}^d \mapsto \mathbb{R}$ is a continuous convex function which is possibly non-smooth. A well studied idea in efficient optimization of such composite objective functions is to start with a quadratic approximation of the form:

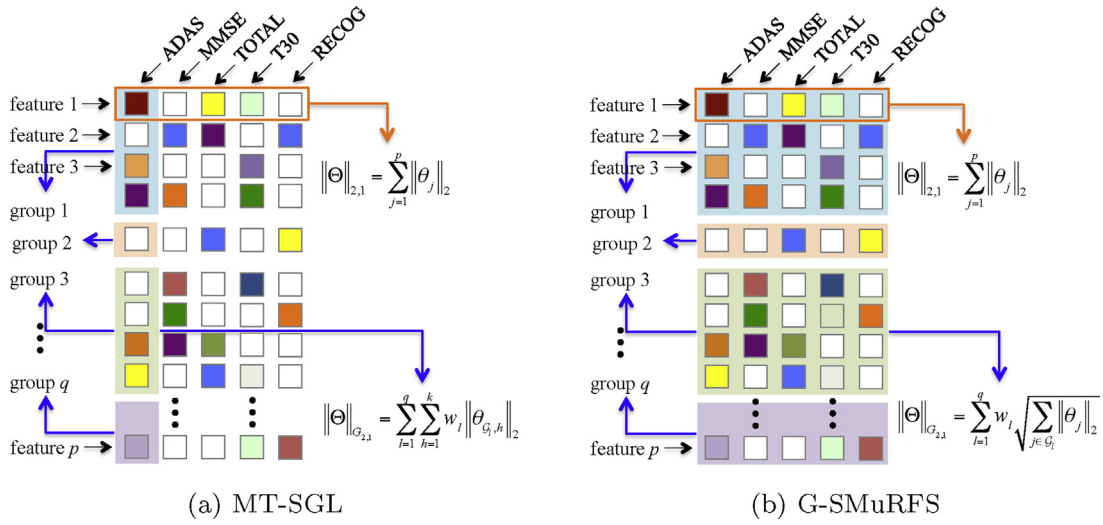


Fig. 4. The illustration of two group guided methods: MT-SGL and G-SMuRFS. Each column of Θ is corresponding to a single task and each row represents a feature dimension. The MRI measure features in each region belong to a group. We assume the p features to be divided into q disjoint groups G_l, $l = 1, \ldots, q$, with each group having m_l features respectively. For each element in Θ, white color means zero-valued elements and color indicates non-zero values. The different colors in the square boxes indicate the weights of the feature for the corresponding task. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$Q_\kappa(\mathbf{x}, \mathbf{x}^{(t)}) := f(\mathbf{x}^{(t)}) + \left\langle \mathbf{x} - \mathbf{x}^{(t)}, \nabla f(\mathbf{x}^{(t)}) \right\rangle + \frac{\kappa}{2}\left\|\mathbf{x} - \mathbf{x}^{(t)}\right\|^2 + g(\mathbf{x}). \tag{8}$$

Ignoring constant terms in $\mathbf{x}^{(t)}$, the unique minimizer of the above expression can be written as

$$P^\kappa_{f,g}(\mathbf{x}^{(t)}) = \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{\kappa}{2}\left\| \mathbf{x} - \left(\mathbf{x}^{(t)} - \frac{1}{\kappa}\nabla f(\mathbf{x}^{(t)})\right)\right\|^2 \right\}, \tag{9}$$

which can be viewed as a proximal operator corresponding to the non-smooth function $g(\mathbf{x})$. A popular approach to solving problems such as (7) is to simply do the following iterative update:

$$\mathbf{x}^{(t+1)} = P^\kappa_{f,g}(\mathbf{x}^{(t)}), \tag{10}$$

which can be shown to have a $O(1/t)$ rate of convergence (Nesterov, 2005; Parikh and Boyd, 2013).

For our purposes, we consider a refined version of the iterative algorithm inspired by Nesterov's accelerated gradient descent (Nesterov, 2005; Parikh and Boyd, 2013). The main idea, as studied in the literature as FISTA-style algorithms (Beck and Teboulle, 2009), is to iteratively consider the proximal operator $P^\alpha_{f,g}$ at a specific linear combination of the previous two iterates $\{\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}\}$, in particular at

$$\mathbf{z}^{(t+1)} = \mathbf{x}^{(t)} + \alpha^{(t+1)}(\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}), \tag{11}$$

instead of at just the previous iterate $\mathbf{x}^{(t)}$. The choice of $\alpha^{(t+1)}$ follows Nesterov's accelerated gradient descent (Nesterov, 2005; Parikh and Boyd, 2013) and is detailed in Algorithm 1. The iterative algorithm simply updates

$$\mathbf{x}^{(t+1)} = P^\kappa_{f,g}(\mathbf{z}^{(t+1)}). \tag{12}$$

As shown in (Beck and Teboulle, 2009), the algorithm has a rate of convergence of $O(1/t^2)$.

A key building block in MT-SGL is the computation of the proximal operator in (12) when $g(\cdot) \equiv R^{\lambda_1}_{\lambda_2}(\cdot)$ is the multi-task sparse group lasso regularizer given by

$$R^{\lambda_1}_{\lambda_2}(\Theta) = \lambda_1\|\Theta\|_{2,1} + \lambda_2\|\Theta\|_{G_{2,1}}. \tag{13}$$

For MT-SGL, the iterates $\mathbf{x}^{(t)} \equiv \Theta^{(t)}$ are matrices, and the proximal operator is computed at $\mathbf{z}^{(t+1)} \equiv Z^{(t+1)} = \Theta^{(t)} + \alpha^{(t+1)}(\Theta^{(t)} - \Theta^{(t-1)})$. For the loss function $L(\cdot)$ corresponding to Gaussian (least squares) and Poisson regression, $f_{\text{Gauss}}(Z^{(t+1)}) = \|Y - XZ^{(t+1)}\|^2_F$ and $f_{\text{Poisson}}(Z^{(t+1)}) = \|Y - XZ^{(t+1)}\|^2_F$, respectively. The loss gradients are given by

$$\begin{aligned} \nabla f_{\text{Gauss}}(Z^{(t+1)}) &= X^T(XZ^{(t+1)} - Y), \\ \nabla f_{\text{Poisson}}(Z^{(t+1)}) &= X^T(\exp(XZ^{(t+1)}) - Y), \end{aligned} \tag{14}$$

with $V^{(t+1)} = Z^{(t+1)} - \frac{1}{\kappa}\nabla f(Z^{(t+1)})$, the problem of computing the proximal operator $P^\kappa_{f,g}(Z^{(t+1)})$: $= P^{\lambda_1/\kappa}_{\lambda_2/\kappa}(V^{(t+1)})$ is given by

$$\begin{aligned} P^{\lambda_1/\kappa}_{\lambda_2/\kappa}(V^{(t+1)}) &= \operatorname*{argmin}_{\Theta\in\mathbb{R}^{p\times k}} \left\{ R^{\lambda_1/\kappa}_{\lambda_2/\kappa}(\Theta) + \tfrac{1}{2}\|\Theta - V^{(t+1)}\|^2 \right\} \\ &= \operatorname*{argmin}_{\Theta\in\mathbb{R}^{p\times k}} \left\{ R^{\lambda_1}_{\lambda_2}(\Theta) + \tfrac{\kappa}{2}\|\Theta - V^{(t+1)}\|^2 \right\}. \end{aligned} \tag{15}$$

The goal is to be able to compute $\Theta^{(t+1)} = P^{\lambda_1/\kappa}_{\lambda_2/\kappa}(V^{(t+1)})$ efficiently. For simple regularizers, such as ℓ_p-regularization with $p \in \{1, 2, \infty\}$, proximal operators are available in closed-form and can easily be computed (Combettes and Pesquet, 2011). On the other hand, proximal operators for complex regularizers are non-trivial and usually resort to inner iterative subroutine, which becomes frustratingly slow (Yu, 2013a).

When complex regularizers are composed of the sum of simple regularizers, researchers have looked for ways to leverage the fact that proximal operators of the summands are easy to compute. In the following, two proximal operators combination strategies are discussed: *shape proximal average* (Bauschke et al., 2008; Yu, 2013a) and *shape proximal composition* (Jenatton et al., 2011; Yu, 2013b).

Proximal average: It simply averages the solutions from the proximal operator for each simple regularizer, that is, $P_{\sum_i^k f_i} \approx \sum_i^k \alpha_i P_{f_i}$. Besides having nice properties (Bauschke et al., 2008), it has been shown that proximal average acts as a surrogate function and is a good approximation for the original composite regularizer (Yu, 2013a). Additionally, as proximal operators can be computed independently, it is suitable for parallel proximal gradient algorithms.

For the MT-SGL, the proximal average operator is computed as

$$\begin{aligned} U_1^{(t+1)} &= P^{\lambda_1/\kappa}(V^{(t+1)}), \\ U_2^{(t+1)} &= P_{\lambda_2/\kappa}(V^{(t+1)}), \\ \Theta^{(t+1)} &= \frac{U_1^{(t+1)} + U_2^{(t+1)}}{2} = P_{\lambda_2/\kappa}^{\lambda_1/\kappa}(V^{(t+1)}) \end{aligned} \tag{16}$$

where P^{λ_1}/κ and $P_{\lambda_2/\kappa}$ are the proximal operators for the $\ell_{2,1}$ and $G_{2,1}$ regularizers, respectively, and are discussed later in this section.

Proximal composition: Yu (2013) investigated the proximal operator of the sum of multiple non-smooth functions as the composition of the proximal operators for individual regularizers, that is, $P_{\sum_i^k f_i} = P_{f_1} \circ \cdots \circ P_{f_k}$. Yu shows sufficient conditions for the decomposition hold (see Theorem 1 in (Yu, 2013b)). Although he also shows that it needs not hold in general, as we show in Section 4.2, such strategy works well in practice.

Proximal composition for the MT-SGL composite regularizer (6) is expressed as: $P_{\lambda_2/\kappa}^{\lambda_1/\kappa}(\Theta) = P_{\lambda_2/\kappa} \circ P^{\lambda_1/\kappa}(\Theta)$, or more specifically

$$\begin{aligned} U^{(t+1)} &= P^{\lambda_1/\kappa}(V^{(t+1)}), \\ \Theta^{(t+1)} &= P_{\lambda_2/\kappa}(U^{(t+1)}) = P_{\lambda_2/\kappa}^{\lambda_1/\kappa}(V^{(t+1)}). \end{aligned} \tag{17}$$

Both *shape proximal average* and *shape proximal composition* take the proximal operators for the individual regularization terms and make a combination of them. Next, we discuss the proximal operators of the individual regularizers: $\ell_{2,1}$ and $G_{2,1}$, which compose the MT-SGL formulation. We also show that they can be executed efficiently using suitable extensions of soft-thresholding.

The proximal map of the $\ell_{2,1}$ regularization can be written as

$$U^{(t+1)} = \underset{U \in \mathbb{R}^{p \times k}}{\operatorname{argmin}} \left\{ \frac{\lambda_1}{\kappa} \left\| U \right\|_{2,1} + \frac{1}{2} \| U - V^{(t+1)} \|_F^2 \right\}. \tag{18}$$

Since $\left\| U \right\|_{2,1} = \sum_{j=1}^{p} \left\| \mathbf{u}_j \right\|_2$, the problem decomposes over the rows $\mathbf{u}_j$, $j = 1, \ldots, p$. Following (Liu et al., 2009), the row-wise updates can be done by soft-thresholding as

$$\mathbf{u}_j = \frac{\max\left\{ \|\mathbf{v}_j\|_2 - \frac{\lambda_1}{\kappa}, 0 \right\}}{\|\mathbf{v}_j\|_2} \mathbf{v}_j, \quad j = 1, \ldots, p, \tag{19}$$

where $\mathbf{u}_j$, $\mathbf{v}_j$ are the jth rows of $U^{(t+1)}$, $V^{(t+1)}$ respectively.

As for the $G_{2,1}$ regularization, the proximal map is defined as

$$\Theta^{(t+1)} = \underset{\Theta \in \mathbb{R}^{p \times k}}{\operatorname{argmin}} \left\{ \frac{\lambda_2}{\kappa} \left\| \Theta \right\|_{G_{2,1}} + \frac{1}{2} \left\| \Theta - U^{(t+1)} \right\|^2 \right\}. \tag{20}$$

Since $\left\| \Theta \right\|_{G_{2,1}} = \sum_{j=1}^{q} \sum_{h=1}^{k} w_j \left\| \theta_{\mathscr{G}_j,h} \right\|_2$, the problem decomposes as updates over the task specific groups $\theta_{\mathscr{G}_j,h}$, $j = 1, \ldots, q$, $h = 1, \ldots, k$. Following (Yuan et al., 2013), the task-specific group updates can be done by soft-thresholding as

$$\Theta_{\mathscr{G}_j,h} = \frac{\max\left\{ \|\Theta_{\mathscr{G}_j,h}\|_2 - \frac{\lambda_2}{\kappa}, 0 \right\}}{\|\Theta_{\mathscr{G}_j,h}\|_2} U_{\mathscr{G}_j,h}, \quad j = 1, \ldots, q, \quad h = 1, \ldots, k, \tag{21}$$

where $\Theta_{\mathscr{G}_j,h}$, $U_{\mathscr{G}_j,h}$ are parameters for task-specific groups for group j and task h in $\Theta^{(t+1)}$ and $U^{(t+1)}$, respectively.

In practice, since the Lipschitz constant κ may be unknown, we perform a backtracking line-search procedure to ensure function minimization. The pseudocode of MT-SGL is summarized in Algorithm 1, where $F(\Theta)$ denotes the objective function of MT-SGL as in Eq. (6), $Q_\kappa(\Theta_1, \Theta_2)$ denotes the quadratic approximation as in Eq. (8) for the MT-SGL objective, and $P_{\lambda_2/\kappa}^{\lambda_1/\kappa}$ denotes the proximal operator for the MT-SGL regularization as in Eq. (16) or (17). The algorithm can be stopped if the change of the function values corresponding to adjacent iterations is within a small value, say 10^{-4}.

Algorithm 1. The MT-SGL algorithm

$F(P_R^{\bar{\kappa}}(Z^{(t)})) \leq Q_{\bar{\kappa}}(P_R^{\bar{\kappa}}(Z^{(t)}), Z^{(t)})$

Require
$\gamma > 0$ ▷ Regularization parameter
X, Y ▷Data for all tasks
$\lambda_1 \longleftarrow \gamma \lambda_1^{\max}$ ▷$\lambda_1^{\max}$ is computed as (22a)
$\lambda_2 \longleftarrow \gamma \lambda_2^{\max}$ ▷$\lambda_2^{\max}$ is computed as (22b)
Set $\beta^{(0)} \longleftarrow 1$ and $\kappa^{(0)} \longleftarrow 1$ ▷Define initial β and κ
Set $\Theta^{(0)} \longleftarrow \mathscr{W}(0, 1)$ ▷Define initial parameter matrix Θ
Set $Z^{(1)} \longleftarrow \Theta^{(0)}$
$t \longleftarrow 1$
While convergence criterion not met **do**
Find the smallest nonnegative integers i_t such that with $\bar{\kappa} = 2^{i_t}\kappa^{(t-1)}$ $F(P_R^{\bar{\kappa}}(Z^{(t)})) \leq Q_{\bar{\kappa}}(P_R^{\bar{\kappa}}(Z^{(t)}), Z^{(t)})$
$\kappa^{(t)} \longleftarrow 2^{i_t}\kappa^{(t-1)}$
$V^{(t)} \longleftarrow Z^{(t)} - \frac{1}{\kappa^{(t)}} \nabla f(Z^{(t)})$
$\Theta^{(t)} \longleftarrow P_{\lambda_2/\kappa^{(t)}}^{\lambda_1/\kappa^{(t)}}(V^{(t)})$ ▷Proximal average or proximal composition
$\beta^{(t)} \longleftarrow \frac{1 + \sqrt{1 + 4(\beta^{(t-1)})^2}}{2}$
$Z^{(t)} \longleftarrow \Theta^{(t)} + \frac{\beta^{(t-1)} - 1}{\beta^{(t)}}(\Theta^{(t)} - \Theta^{(t-1)})$
$t \longleftarrow t + 1$
end while

In the MT-SGL, two regularization parameters need to be specified: λ_1 and λ_2. Using recent results on norm regularization (Banerjee et al., 2014), it is possible to express both parameters via a single parameter as follows: $\lambda_1 = \gamma \lambda_1^{\max}$ and $\lambda_2 = \gamma \lambda_2^{\max}$ (Meier et al., 2008; Banerjee et al., 2014), where $\lambda_1^{\max}$ and $\lambda_2^{\max}$ are computed as:

$$\lambda_1^{\max} = \|X^T Y\|_\infty, \tag{22a}$$

$$\lambda_2^{\max} = \underset{\ell, j \in \mathscr{G}_\ell}{\operatorname{argmax}} \frac{1}{\sqrt{m_\ell}} \|\max\{(|X_j^T Y| - \gamma \lambda_1^{\max}), 0\}\|_2. \tag{22b}$$

The choices follow from the current understanding in the literature of the correct form these parameters, in particular, in terms of the dual norm of the gradient of the objective (Banerjee et al., 2014; Liu and Ye., 2010). Thus, the only parameter to be empirically chosen in MT-SGL is the scaling γ.

Python codes of the proposed algorithm are available at: https://bitbucket.org/XIAOLILIU/mtl-sgl.

4. Experimental results

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

4.1. Experimental setting

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 has been 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

Table 1
Cortical features (n = 275).

Number	ROI	Laterality	Type
1	Caudal Anterior Cingulate Cortex	L, R	CV, SA, TA, TS
2	Caudal Middle Frontal Gyrus	L, R	CV, SA, TA, TS
3	Cuneus Cortex	L, R	CV, SA, TA, TS
4	Entorhinal Cortex	L, R	CV, SA, TA, TS
5	Frontal Pole	L, R	CV, SA, TA, TS
6	Fusiform Gyrus	L, R	CV, SA, TA, TS
7	Inferior Parietal Cortex	L, R	CV, SA, TA, TS
8	Inferior Temporal Gyrus	L, R	CV, SA, TA, TS
9	Insula	L, R	CV, SA, TA, TS
10	IsthmusCingulate	L, R	CV, SA, TA, TS
11	Lateral Occipital Cortex	L, R	CV, SA, TA, TS
12	Lateral Orbital Frontal Cortex	L, R	CV, SA, TA, TS
13	Lingual Gyrus	L, R	CV, SA, TA, TS
14	Medial Orbital Frontal Cortex	L, R	CV, SA, TA, TS
15	Middle Temporal Gyrus	L, R	CV, SA, TA, TS
16	Paracentral Lobule	L, R	CV, SA, TA, TS
17	Parahippocampal Gyrus	L, R	CV, SA, TA, TS
18	Pars Opercularis	L, R	CV, SA, TA, TS
19	Pars Orbitalis	L, R	CV, SA, TA, TS
20	Pars Triangularis	L, R	CV, SA, TA, TS
21	Pericalcarine Cortex	L, R	CV, SA, TA, TS
22	Postcentral Gyrus	L, R	CV, SA, TA, TS
23	Posterior Cingulate Cortex	L, R	CV, SA, TA, TS
24	Precentral Gyrus	L, R	CV, SA, TA, TS
25	Precuneus Cortex	L, R	CV, SA, TA, TS
26	Rostral Anterior Cingulate Cortex	L, R	CV, SA, TA, TS
27	Rostral Middle Frontal Gyrus	L, R	CV, SA, TA, TS
28	Superior Frontal Gyrus	L, R	CV, SA, TA, TS
29	Superior Parietal Cortex	L, R	CV, SA, TA, TS
30	Superior Temporal Gyrus	L, R	CV, SA, TA, TS
31	Supramarginal Gyrus	L, R	CV, SA, TA, TS
32	Temporal Pole	L, R	CV, SA, TA, TS
33	Transverse Temporal Cortex	L, R	CV, SA, TA, TS
34	Hemisphere	L, R	SA
35	Total Intracranial Volume	Bilateral	CV

AD progression will help researchers and clinicians develop new treatments and monitor their effectiveness. Further, being able to understand 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 the imaging data from the ADNI database processed by a team from UCSF (University of California at San Francisco), who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) according to the atlas generated in Desikan et al. (2006). The FreeSurfer software was employed to automatically label cortical and subcortical tissue classes for the structural MRI scan of each subject, and to extract thickness measures of cortical regions of interests (ROIs) and volume measures of cortical and subcortical.

Briefly, this processing includes motion correction and 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, 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; Segonne 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).

Totally, 48 cortical regions and 44 subcortical regions are generated and the number of features in each group is typically 1 or 4. The names

Table 2
Subcortical features (n = 44).

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

of 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, subcortical volume was calculated as features. The SA of left and right hemisphere and total intracranial volume (ICV) were also included. This yielded a total of p = 319 MRI features extracted from cortical/subcortical ROIs in each hemisphere (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, repeatedly over a 6-month or 1-year interval. The date when the subjects are scheduled to perform the screening 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 investigate the prediction performance of our method for inferring cognitive outcomes in a number of neuropsychological assessments at baseline time. 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 are categorized into 3 baseline diagnostic groups: Cognitively Normal (CN, n_1 = 225), Mild

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 ± sd)	75.87 ± 5.04	74.75 ± 7.39	75.42 ± 7.25
Education (y, ag ± sd)	16.03 ± 2.85	15.67 ± 2.95	14.65 ± 3.17

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

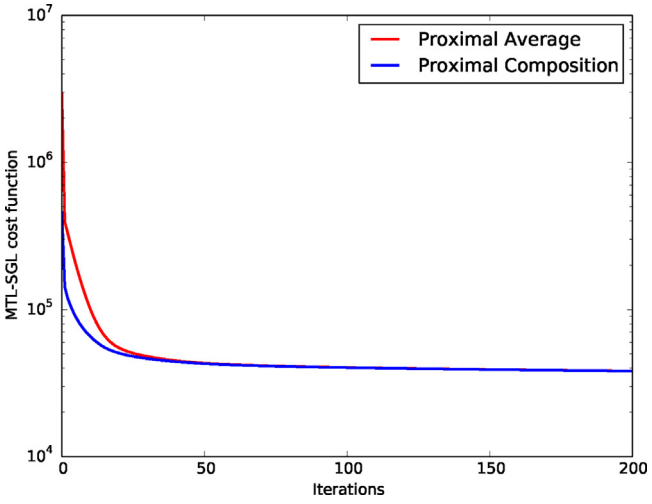


Fig. 5. Convergence of MT-SGL with *shape proximal average* and *shape proximal composition*. Both strategies show similar results.

Cognitive Impairment (MCI, $n_2 = 390$), and Alzheimer's Disease (AD, $n_3 = 173$). Table 3 lists the demographics information of all these subjects, including age, gender and education. The 788 baseline scans in the ADNI database were used for evaluation and a leave 5% out cross validation was adopted as in Wolz et al. (2011). 5% of the evaluation subjects were regarded as the test set, and the remaining 95% of the subjects were used to train a regression model which was then applied to the test set. This was repeated 50 times, each time selecting randomly the test set subjects. Finally, the average (avg) and standard deviation (std) of performance measures across the 50 repetitions were calculated and shown as avg ± std for each experiment. In each run, we ensured that all methods received exactly the same train and test set. For all experiments, a 5-fold nested cross validation procedure is employed to tune the regularization parameters in each trial, with parameter values in the range [1e-4,1e4] Data was z-scored before applying regression methods. The reported results were the best results of each method with the optimal parameter.

For predictive modeling, we focus on 5 widely used cognitive measures (Yan et al., 2015; Li et al., 2012), which to the $k = 5$ tasks in our setting. In particular, the cognitive scores used in our analysis are: Alzheimer's Disease Assessment Scale – cognitive total score (ADAS), Mini Mental State Exam score (MMSE), Rey Auditory Verbal Learning Test (RAVLT) total score (TOTAL), RAVLT 30 minutes delay score (T30) and RAVLT recognition score (RECOG). ADAS is the gold standard in AD drug trial for cognitive function assessment, which 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 is a measure of episodic memory and used for the diagnosis of memory disturbances, which consists of eight recall trials and a recognition test.

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

$$\mathrm{rMSE}(\mathrm{y}, \hat{\mathrm{y}}) = \frac{\|\mathrm{y} - \hat{\mathrm{y}}\|_2^2}{n} \tag{23}$$

$$\mathrm{Corr}(\mathrm{y}, \hat{\mathrm{y}}) = \frac{\mathrm{cov}(\mathrm{y}, \hat{\mathrm{y}})}{\sigma(\mathrm{y})\sigma(\hat{\mathrm{y}})} \tag{24}$$

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

$$\mathrm{nMSE}(Y, \hat{Y}) = \frac{\sum_{h=1}^{k} \frac{\| Y_h - \hat{Y}_h \|_2^2}{\sigma(Y_h)}}{\sum_{h=1}^{k} n_h} \tag{25}$$

$$\mathrm{wR}(Y, \hat{Y}) = \frac{\sum_{h=1}^{k} \mathrm{Corr}(Y_h, \hat{Y}_h) n_h}{\sum_{h=1}^{k} n_h} \tag{26}$$

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

4.2. Proximal-average and proximal-composition empirical convergence analysis

In Section 3 we discussed two proximal gradient methods for solving the optimization problem associated with MT-SGL formulation, namely *shape proximal average* and *shape proximal composition*. The methods differ how the proximal operator of the composite regularizer of MT-SGL (6) are computed. In this section we empirically investigate their convergences and compare their performances.

For this analysis, we assumed Gaussian GLMs for all tasks. Regularization parameter γ was chosen by cross-validation, with values in the range $\gamma \in [1e-5, 1e3]$. We initially look at the convergence of the proposed MT-SGL using both methods. Similar convergence curves can be observed in Fig. 5, although proximal composition presented a smoother curve.

Table 4 shows the RMSE performance of MT-SGL using both proximal operator computation strategies for three groups(AD,CN and MCI). A first glance at the results shows that MT-SGL with proximal

Table 4
RMSE performance of MT-SGL with proximal gradient methods using *shape proximal average* and *shape proximal composition* strategies for three groups(AD,CN and MCI).

	Method	ADAS	MMSE	RAVLT			nMSE
				Total	T30	RECOG	
AD	Prox. average	8.36 ± 2.12	**2.29 ± 0.55**	**7.89 ± 1.96**	**1.95 ± 0.40**	**3.83 ± 0.77**	6.18 ± 1.65
	Prox. composition	**7.71 ± 2.08**	2.49 ± 0.56	7.91 ± 1.96	2.08 ± 0.44	3.88 ± 0.80	**6.07 ± 1.63**
MCI	Prox. average	**5.66 ± 0.67**	2.14 ± 0.32	**8.45 ± 1.35**	**2.89 ± 0.47**	3.44 ± 0.44	**4.52 ± 0.67**
	Prox. composition	**5.66 ± 0.67**	**2.03 ± 0.33**	8.55 ± 1.36	2.91 ± 0.47	**3.41 ± 0.49**	**4.52 ± 0.75**
CN	Prox. average	**6.64 ± 1.30**	2.51 ± 0.48	11.93 ± 2.59	4.73 ± 0.82	3.62 ± 0.53	9.76 ± 2.76
	Prox. composition	6.78 ± 1.31	**2.02 ± 0.29**	**11.44 ± 2.66**	**4.45 ± 0.84**	**3.31 ± 0.53**	**8.65 ± 2.54**
ALL	Prox. average	6.72 ± 0.80	2.31 ± 0.21	9.62 ± 1.13	3.39 ± 0.39	3.62 ± 0.30	4.38 ± 0.52
	Prox. composition	**6.59 ± 0.79**	**2.16 ± 0.18**	**9.50 ± 1.13**	**3.32 ± 0.38**	**3.54 ± 0.31**	**4.19 ± 0.54**

Table 5
RMSE: Baseline methods vs. MT-SGL. Superscript symbols † and * indicate that MT-SGL[5G] and MT-SGL[2G3P], respectively, significantly outperformed that method on that score. Student's t-test at a level of 0.05 was used.

Method	ADAS	MMSE	RAVLT			nMSE
			TOTAL	T30	RECOG	
Ridge	7.19 ± 0.90$^{\dagger *}$	2.55 ± 0.27$^{\dagger *}$	10.68 ± 1.14$^{\dagger *}$	3.82 ± 0.43$^{\dagger *}$	3.99 ± 0.43$^{\dagger *}$	5.34 ± 0.67$^{\dagger *}$
Lasso	6.66 ± 0.78$^{\dagger *}$	2.20 ± 0.19$^{\dagger *}$	9.54 ± 1.12*	3.43 ± 0.35$^{\dagger *}$	3.57 ± 0.31$^{\dagger *}$	4.28 ± 0.49$^{\dagger *}$
Group lasso	6.68 ± 0.80$^{\dagger *}$	2.23 ± 0.18$^{\dagger *}$	9.58 ± 1.13$^{\dagger *}$	3.42 ± 0.40$^{\dagger *}$	3.57 ± 0.31$^{\dagger *}$	4.32 ± 0.52$^{\dagger *}$
MT-GL	6.73 ± 0.77$^{\dagger *}$	2.16 ± 0.18	9.55 ± 1.11*	3.34 ± 0.36†	3.54 ± 0.31	4.25 ± 0.48*
G-SMuRFS	6.69 ± 0.80$^{\dagger *}$	2.16 ± 0.19	9.66 ± 1.11†	3.36 ± 0.38†	3.57 ± 0.33	4.30 ± 0.54$^{\dagger *}$
MT-SGL[5G]comp	**6.59 ± 0.79**	2.16 ± 0.18	**9.50 ± 1.13**	**3.32 ± 0.38**	3.54 ± 0.31	**4.19 ± 0.54**
MT-SGL[2G3P]comp	6.82 ± 0.81	**2.14 ± 0.20**	9.68 ± 1.19	3.34 ± 0.35	**3.53 ± 0.31**	4.38 ± 0.52

Table 6
CC: Baseline methods vs. MT-SGL. Superscript symbols † and * indicate that MT-SGL[5G] and MT-SGL[2G3P], respectively, significantly outperformed that method on that score. Student's *t*-test at a level of 0.05 was used.

Method	ADAS	MMSE	RAVLT			wR
			TOTAL	T30	RECOG	
Ridge	0.61 ± 0.09$^{\dagger *}$	0.38 ± 0.15$^{\dagger *}$	0.44 ± 0.11$^{\dagger *}$	0.41 ± 0.11$^{\dagger *}$	0.32 ± 0.14$^{\dagger *}$	0.43 ± 0.08$^{\dagger *}$
Lasso	0.65 ± 0.08$^{\dagger *}$	0.50 ± 0.14$^{\dagger *}$	0.53 ± 0.10	0.51 ± 0.10$^{\dagger *}$	0.42 ± 0.13$^{\dagger *}$	0.52 ± 0.08$^{\dagger *}$
Group lasso	0.65 ± 0.08†	0.49 ± 0.14$^{\dagger *}$	0.53 ± 0.11$^{\dagger *}$	0.53 ± 0.10$^{\dagger *}$	0.41 ± 0.12$^{\dagger *}$	0.52 ± 0.08$^{\dagger *}$
MT-GL	0.65 ± 0.08$^{\dagger *}$	0.51 ± 0.14†	0.53 ± 0.10	0.53 ± 0.10†	0.43 ± 0.13	0.53 ± 0.08†
G-SMuRFS	0.65 ± 0.08†	0.51 ± 0.14$^{\dagger *}$	0.52 ± 0.11$^{\dagger *}$	0.52 ± 0.09$^{\dagger *}$	0.42 ± 0.11$^{\dagger *}$	0.52 ± 0.08$^{\dagger *}$
MT-SGL[5G]comp	**0.66 ± 0.08**	**0.52 ± 0.14**	**0.54 ± 0.11**	**0.54 ± 0.10**	**0.44 ± 0.13**	**0.54 ± 0.08**
MT-SGL[2G3P]comp	0.64 ± 0.08	**0.52 ± 0.14**	0.53 ± 0.10	**0.54 ± 0.10**	**0.44 ± 0.13**	0.53 ± 0.08

composition achieves better performances than proximal average in the group of AD, CN and ALL, which indicates that proximal composition optimization is more effective than the proximal average algorithm. Based on this result, MT-SGL with proximal composition is considered for the comparison with the baseline methods in the next section.

4.3. Comparison with baseline MTL methods

To validate the effectiveness of the proposed method, we first compared MT-SGL with 5 different regression methods, including: ridge regression, lasso, group lasso (Yuan and Lin, 2006), which are applied independently to each task, multi-task group lasso (MT-GL) based on $\ell_{2,1}$-norm regularization (Liu et al., 2009), and Group-sparse Multitask Regression and Feature Selection (G-SMuRFS) (Yan et al., 2015), which is one of the state-of-the-art methods for characterizing AD progression. Additionally, we compared the performance of the MT-SGL using two different loss functions settings derived from the GLM family: MT-SGL [5G], where all scores are modeled with Gaussian (least squares) regression; and MT-SGL[2G3P], where two scores are modeled with Gaussian (TOTAL and ADAS) and three scores (T30, RECOG and MMSE) with Poisson regression. The use of Poisson model is motived by the response profiles of some cognitive scores, particularly T30, RECOG and MMSE, as shown in Fig. 1.

Regularization parameters for methods are chosen using a nested cross-validation strategy on the training data, with search grid in the range of 5×10^{-3} to 5×10^{3} using a log-scale for MT-SGL and G-SMuRFS, and in the range of 10^{-4} to 10^{4} using a log-scale (Liu et al., 2009) for the other 4 methods. Prediction performance results, measured by RMSE and CC of MT-SGL and 5 different regression methods under 5 cognitive scores are shown in Table 5 and 6.

A first glance at the results shows that our MT-SGL method achieved the best performance compared to the competing methods. From the t-test results, we can observe that MT-SGL(5G) and MT-SGL(2G3P) are statistically significantly better than the competing methods with respect to nMSE and wR. Specifically, we observe the following:

(1) The results show that sparse learning methods (Lasso, Group Lasso, MT-GL, G-SMuRFS, and MT-SGL) are significantly more effective than ridge regression on predicting all scores. Lasso and group lasso are single-task learning methods being applied independently on each task, whereas MT-GL, G-SMuRFS and MT-SGL are multi-task learning methods.
(2) The multi-task learning methods: MT-SGL and MT-GL showed smaller nMSEs and higher wR than single-task learning methods, as they can exploit possible commonalities among scores. G-SMuRFS is worse than lasso and MTL-GL, the observation is same as the results in Yan et al. (2015), where G-SMuRFS did not show clear performance improvement over the MTL-GL based on the vertex-based surface measures, although it considers grouping the relevant surface features together according to anatomic structure. The reason may be that: (i) the strong assumption that the both features and ROIs are shared across multiple score tasks, and (ii) the iterative alternative optimization (AO) algorithm used in G-SMuRFS is not appropriate for the optimization of the formulation with nonsmooth structured regularization.
(3) We investigate the effect of group penalty in our model by comparing the results of MT-GL. MT-SGL(5G) outperforms MT-GL in terms of all the metrics. The traditional MT-GL considered only the sparsity of the regression coefficients, thus failing to capture the group structure of features in the data. When multiple features are extracted to measure the atrophy of each imaging biomarker, we can find that it can further improve the prediction performance by capturing of inherent feature structures.
(4) Compared with MT-SGL(5G), MT-SGL(2G3P) improves the regression performance on the scores of MMSE, T30 and RECOG, of which the distributions are Poisson. The results indicate that the GLM in MT-STL can provide a more flexible analysis approach for analyzing the data. Modeling the relationship between MRI features and the value of the cognitive score for each specific score task can help improve the performances for the multi-task learning.

4.4. Selection of ROI's

In Alzheimer's disease studies, researchers are not only interested in providing better cognitive scores prediction, but mainly to identify which are the brain areas more affected by the disease, which can help

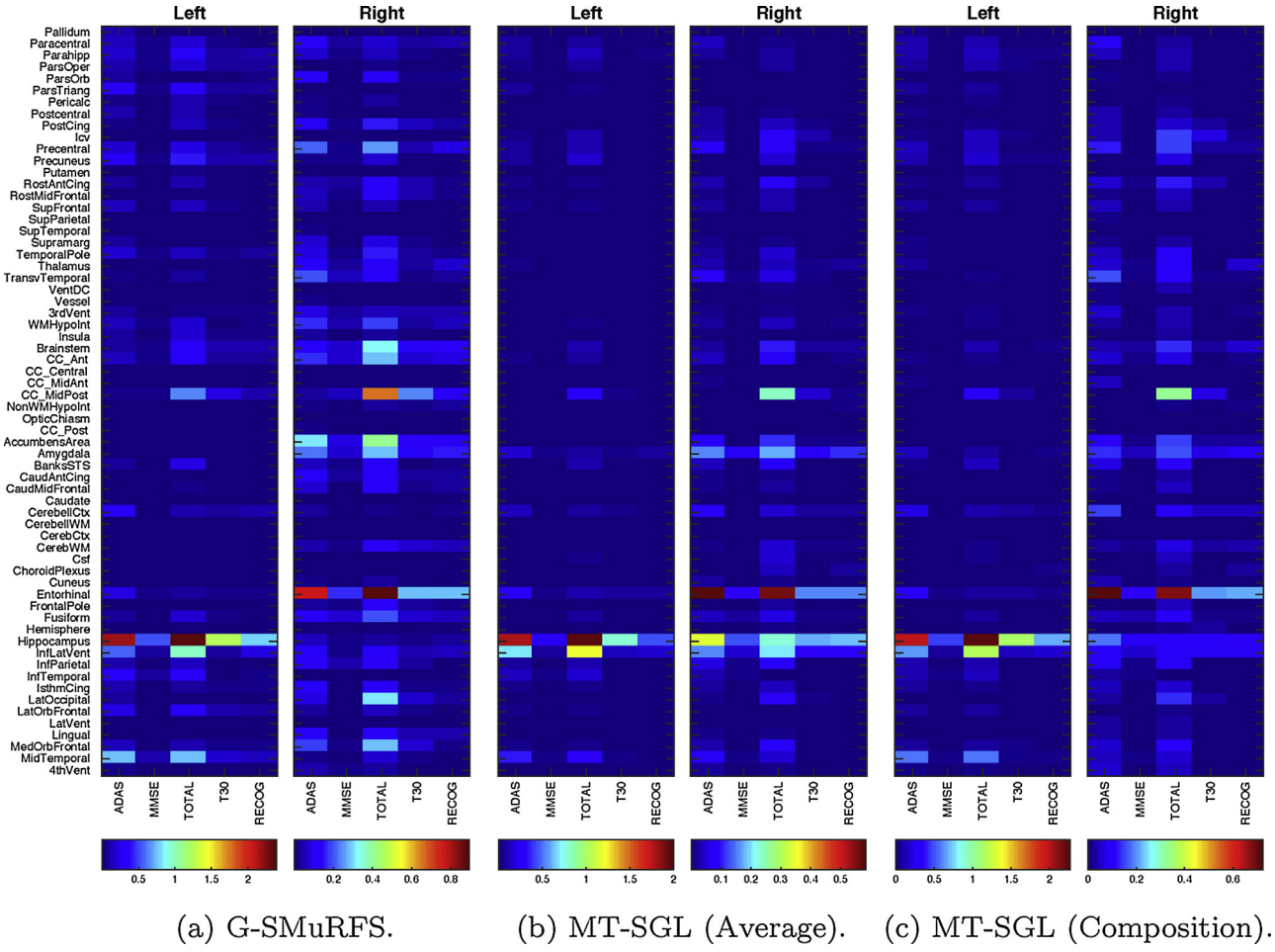


Fig. 6. Heat maps of regression coefficients of 50 trials on different splits of data. (a) G-SMuRFS, (b) proximal average, (c) proximal composition.

to diagnose early stages of the disease and how it spreads. We, then, turn our analysis now to the identification of MRI biomarkers. Both MT-SGL and G-SMuRFS are group sparse models which are able to identify a compact set of relevant neuroimaging biomarkers from the region level due to the group lasso on the features, which would provide us with better interpretability of the brain region.

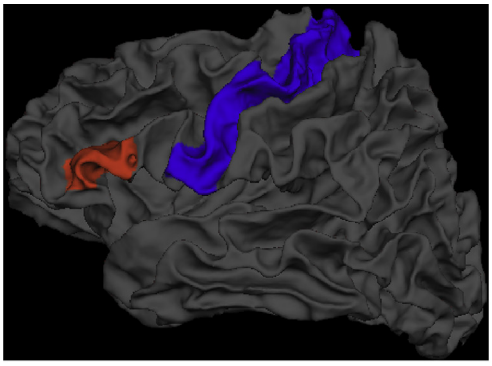
Fig. 6 are the heat maps of the regression weights (or coefficients) of all ROIs in each hemisphere for each cognitive score at the baseline time calculated by MT-SGL with two optimization strategies and G-SMuRFS methods through 50 trials. The value of each item (i, j) in the heat map indicates the weight of the ith ROI for the jth task, and is calculated by $\frac{1}{w_i}\sqrt{\sum_{k\in\mathcal{G}_i}\left\|\theta_{\mathrm{ki}}\right\|_2}$, where k is the the kth MRI feature. The larger the absolute value of a coefficient, the more important its corresponding brain region is in predicting the corresponding cognitive score. The figure illustrates that the proposed MT-SGL clearly presented a much better sparsity across all the cortical measures than G-SMuRFS from the level of ROI, where a small portion of the brain region was identified to be relevant to the cognitive outcome. The sparse ROIs make the results easier to interpret. The heat maps of MT-SGL with two optimization strategies are nearly the same. Based on the heat map, we selected the top 10 features according to the regression weights. The top 10 selected MRI features and brain regions (ROI) are shown in Table 7.

Table 7
Top 10 selected ROIs by G-SMuRFS, proximal average and proximal composition.

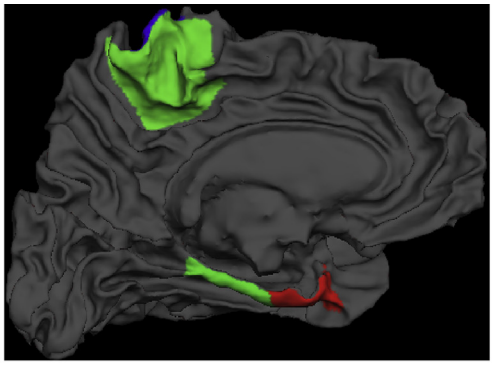
Numbers	Regions		
	G-SMuRFS	MT-SGL (average)	MT-SGL (composition)
1	L.Hippocampus	L.Hippocampus	L.Hippocampus
2	R.Entorhinal	L.InfLatVent	L.InfLatVent
3	L.InfLatVent	R.Entorhinal	R.Entorhinal
4	L.MidTemporal	R.Hippocampus	L.MidTemporal
5	CC_MidPost	L.MidTemporal	CC_MidPost
6	L.Precuneus	L.Entorhinal	L.Entorhinal
7	R.AccumbensArea	R.Amygdala	L.CerebellCtx
8	L.ParsTriang	L.InfTemporal	L.Precuneus
9	L.InfTemporal	R.InfLatVent	R.Hippocampus
10	L.LatOrbFrontal	L.Parahipp	R.Amygdala

Table 8
Top 10 ROIs selected via stability selection by G-SMuRFS, proximal average and proximal composition.

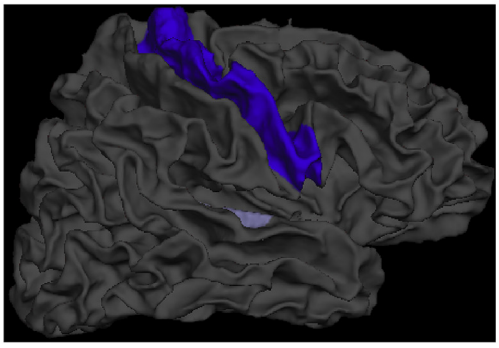
Regions		
G-SMuRFS	MT-SGL (average)	MT-SGL (composition)
L.Hippocampus	L.Hippocampus	L.Entorhinal
L.InfLatVent	L.InfLatVent	L.Hippocampus
R.Entorhinal	L.Parahipp	L.InfLatVent
L.ParsTriang	L.ParsOper	L.MidTemporal
R.Precentral	R.AccumbensArea	L.Parahipp
L.Paracentral	R.Amygdala	R.Amygdala
L.Precentral	R.Entorhinal	R.Entorhinal
L.Parahipp	R.Fusiform	R.Hippocampus
L.Entorhinal	R.Hippocampus	R.InfLatVent
R.TransvTemporal	R.InfLatVent	L.ParsOper



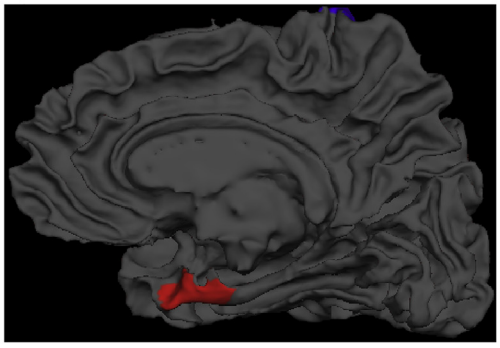
(a) Left-Hemisphere: *Left ParsTriang, Left Precentral.*



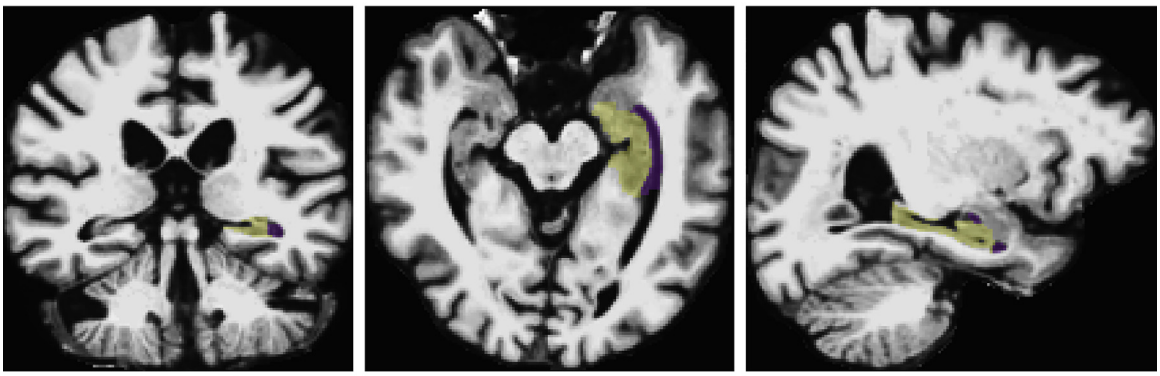
(b) Left-Hemisphere (view from inside out): *Left Paracentral, Left Parahipp, Left Entorhinal.*



(c) Right-Hemisphere: *Right Precentral, Right TransvTemporal.*



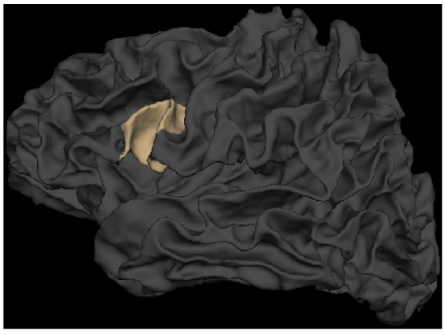
(d) Right-Hemisphere (view from inside out): *Right Entorhinal.*



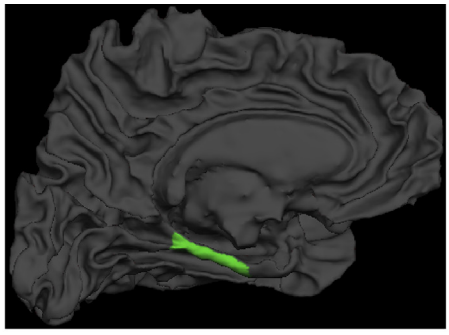
(e) Subcortical: *Left Hippocampus, Left InfLatVent.*

Fig. 7. Plots show the top 10 ROI's selected by G-SMuRFS. These were the most relevant areas for predicting all cognitive scores jointly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

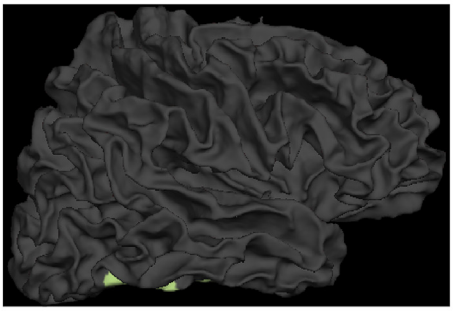
The first six brain regions selected by our MT-SGL are Hippocampus (Zhu et al., 2016; Braak and Braak, 1985; Van Hoesen et al., 1991), Entorhinal (Yan et al., 2015), Inferior lateral ventricle (Gutman et al., 2015; Wan et al., 2014) and Middle Temporal (Yan et al., 2015; Xu et al., 2016; Visser et al., 2002; Zhu et al., 2016), which are highly relevant to the cognitive impairment. These findings are in accordance with the known knowledge that in the pathological pathway of AD. These identified brain regions have been pointed out in the previous literatures and have been also shown to be highly related to clinical functions. For example, Hippocampus are located in the temporal lobe of the brain, which are the role of the memory and spatial navigation. The Hippocampi are the first damaged regions in AD, showing loss of memory and spatial or Entorhinalientation. Entorhinal cortex has long been considered as a relevant and reliable measure to identify individuals at risk for Alzheimer's disease. As entorhinal is a part of the memory system, the damage caused by Alzheimer's disease play a prominent role in the memory deficits. Moreover, Devanand et al. (2007) showed that the reduction of hippocampal and entorhinal cortex



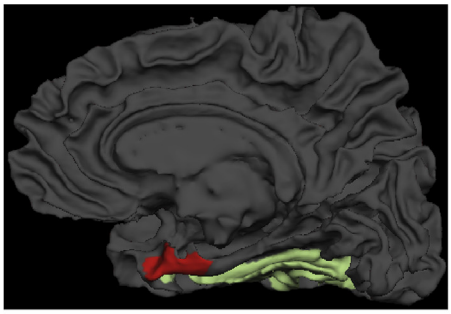
(a) Left-Hemisphere: *Left ParsOper* is highlithed.



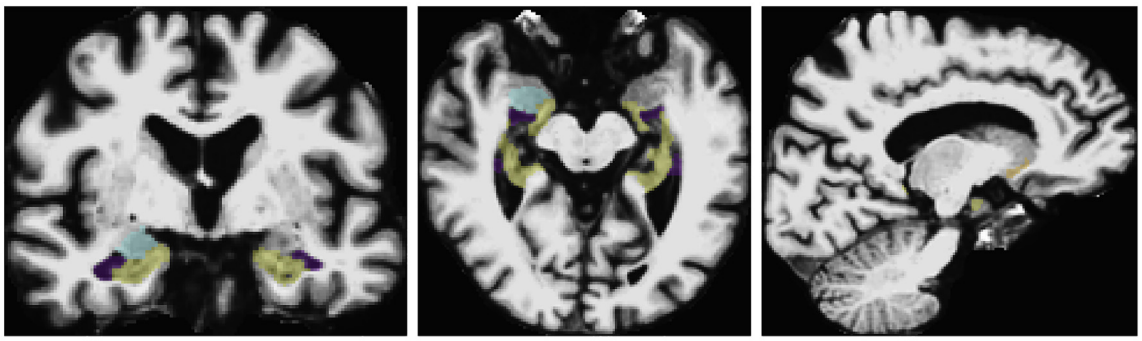
(b) Left-Hemisphere (view from inside out): *Left Parahipp.*



(c) Right-Hemisphere: No relevant ROI has been selected.



(d) Right-Hemisphere (view from inside out): *Right Entorhinal, Right Fusiform.*



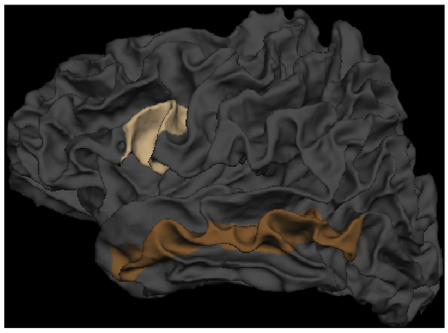
(e) Subcortical: *Right Amygdala, Right InfLatVent, Right Hippocampus, Left InfLatVent, Left Hippocampus, Right AccumbensArea.*

Fig. 8. Plots show the top 10 ROI's selected by MT-SGL with proximal average approach. These were the most relevant areas for predicting all cognitive scores jointly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

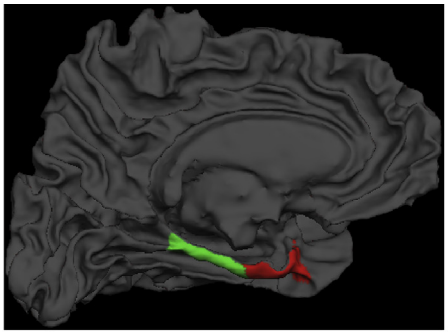
volumes contribute to the conversion of patients from MCI to AD. Additionally, changes in thickness of the inferior parietal lobule are occurring early in the progression from normal to MCI, and related to neuropsychological performance (Greene et al., 2010). So, these regions are important biomarkers for AD, as also identified by MT-SGL.

In order to minimize falsely select variables, we used a stability selection procedure described in Meinshausen and Bühlmann (2010). The stability selection allows to mitigate possible spurious selected variables due to noise and/or random fluctuation of the data. As in our model groups of features correspond to a single ROI, to be able to identify the most relevant ROI's for predicting cognitive scores we performed a *shape group stability selection*, which is described in the following. Let n be the number of data samples and Γ is the set of considered regularization parameter $\gamma \in \Gamma \subseteq \mathbb{R}$.

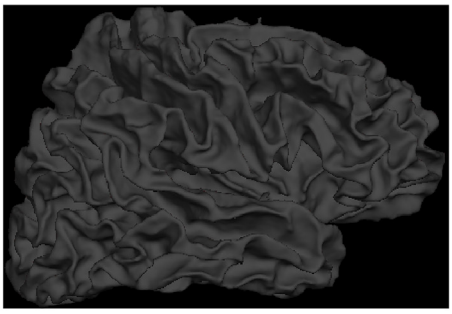
1. For each value of $\gamma \in \Gamma$, do:
   (a) From the data, generate N sub-samples of size $\lfloor n/2 \rfloor$ without replacement;
   (b) For each sub-sample i, run MT-SGL with parameter γ and obtain the variables selection set $\hat{S}_i^{\gamma}$;
   (c) With the selection sets, compute the (empirical) probability of each variable k being selected by MT-SGL:



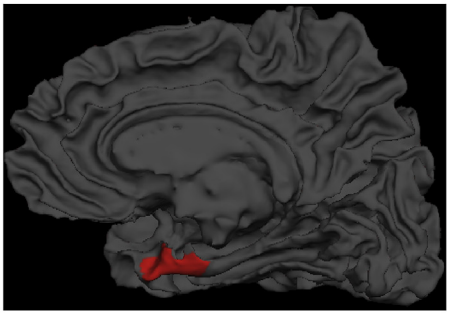
(a) Left-Hemisphere: *Left Middle Temporal, Left ParsOper.*



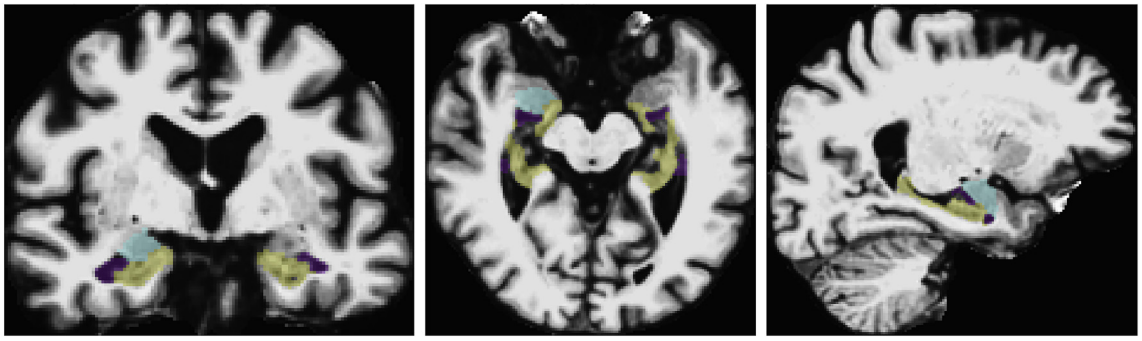
(b) Left-Hemisphere (view from inside out): *Left Entorhinal, Left Parahipp.*



(c) Right-Hemisphere: No relevant ROI has been selected.



(d) Right-Hemisphere (view from inside out): *Right Entorhinal.*



(e) Subcortical: *Right Amygdala, Right InfLatVent, Right Hippocampus, Left InfLatVent, Left Hippocampus.*

Fig. 9. Plots show the top 10 ROI's selected by MT-SGL with proximal composition approach. These were the most relevant areas for predicting all cognitive scores jointly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\mathbb{P}_k^{\gamma} = \mathbb{P}(k \in \hat{S}^{\gamma}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(k \in \hat{S}_i^{\gamma}), \tag{27}$$

where $\mathbb{I}(\cdot)$ is the indicator function;

(d) Compute the average of probabilities $\mathbb{P}_k^{\gamma}$ of the variables belonging to each group $\mathcal{G}_\ell$, $\ell = 1, ..., q$, that is, $\Pi_{\mathcal{G}_\ell}^{\gamma} = \frac{1}{m_\ell}\sum_{k\in\mathcal{G}_\ell} \mathbb{P}_k^{\gamma}$;

2. Given the probability of each group been selected by MT-SGL for all $\gamma \in \Gamma$, $\Pi_{\mathcal{G}_\ell}^{\gamma}$, the set of stable groups (ROIs) are those who satisfies the following definition:

$$\mathcal{G}^{\text{stable}} = \{\mathcal{G}_\ell : \max_{\gamma\in\Gamma} \Pi_{\mathcal{G}_\ell}^{\gamma} \geq \pi_{\text{thr}}\} \tag{28}$$

where $0 < \pi_{\text{thr}} < 1$ is cutoff threshold.

The intuition behind definition (28), is that stable ROIs are those who have been selected by MT-SGL with high probability. Using the group stability selection procedure described with $\Gamma \in$ [1e-5,1], $N = 50$, and $\pi_{\text{thr}} = 0.9$, MT-SGL was performed using both proximal average and proximal composition.

Table 8 shows the top 10 ROIs identified by the three methods through stability selection procedure. Fig. 7 shows the top 10 ROIs selected by G-SMuRFS and Figs. 8 and 9 show the top 10 ROIs selected by MT-SGL with the two different optimization methods. We observe that, the results are accordance with the one obtained by heat map, such as Hippocampus, InfLatVent, Entorhinal, Parahipp and Amygdala, which are also selected by heat map. Moreover, it is found that the most discriminative regions selected by G-SMuRFS and MT-SGL are not

completely the same. For example, Amygdala and R.Hippocampus, which are not identified by G-SMuRFS. Some results suggest that the magnitude of amygdala atrophy is comparable to that of the hippocampus in the earliest clinical stages of AD, and is related to global illness severity (Poulin et al., 2011).

5. Conclusion

Many clinical/cognitive measures have been designed to evaluate the cognitive status of the patients and such measures have been used as criteria for clinical diagnosis of probable AD. In this paper, we propose a multi-task learning framework for predictive modeling of such cognitive measures based on MRI data from ADNI. Our proposed MT-SGL framework considers structured sparsity of parameters with both coupling across tasks and group selection for individual tasks, can work with general loss functions and GLMs, and optimization is done using an efficient FISTA-style method.

Experiments and comparisons with baseline methods illustrate that MT-SGL is at par and usually outperforms existing methods. The method was also able to identify key brain areas for AD progression that corroborate with earlier studies in AD literature. MT-SGL has shown a promising tool not only to predict cognitive scores but also to provide inputs for domain experts towards the understanding of AD progression. In the current work, only priori group information is incorporated into multi-task predictive model, but lack the ability of learning the feature groups automatically. In future work we are interested in investigations of other structure in features, such as graph structure, whichcan help gain additional insights to understand and interpret data.

References

Argyriou, A., Evgeniou, T., Pontil, M., 2007. Multi-task feature learning. Advances in Neural Information Processing Systems (NIPS) 41–48.

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

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012. Structured sparsity through convex optimization. Stat. Sci. 27 (4), 450–468.

Banerjee, A., Chen, S., Fazayeli, F., Sivakumar, V., 2014. Estimation with norm regularization. Advances in Neural Information Processing Systems (NIPS) 1556–1564.

Bauschke, H.H., Goebel, R., Lucet, Y., Wang, X., 2008. The proximal average: basic theory. SIAM J. Optim. 19 (2), 766–785.

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2 (1), 183–202.

Braak, E., Braak, H., 1985. On areas of transition between entorhinal allocortex and temporal isocortex in the human brain. Normal morphology and lamina-specific pathology in Alzheimer's disease. Acta Neuropathol. 68 (4), 325–332.

Chatterjee, S., Steinhaeuser, K., Banerjee, A., Chatterjee, S., Ganguly, A.R., 2012. Sparse group lasso: consistency and climate applications. SIAM International Conference on Data Mining 47–58.

Combettes, P., Pesquet, J., 2011. Proximal splitting methods in signal processing. Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer, pp. 185–212.

Dale, A., Sereno, M., 1993. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. J. Cogn. Neurosci. 5 (2), 162–176.

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

de Toledo-Morrell, L., Stoub, T., Bulgakova, M., Wilson, R., Bennett, D., Leurgans, S., Wuu, J., Turner, D., 2004. MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. Neurobiol. Aging 25 (9), 1197–1203.

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., Albert, M.S., Killiany, R.J., 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.

Devanand, D., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G., Honig, L., Mayeux, R., Stern, Y., Tabert, M., de Leon, M., 2007. Hippocampal and entorhinal atrophy in mild cognitive impairment prediction of Alzheimer's disease. Neurology 68 (11), 828–836.

Evgeniou, T., Pontil, M., 2004. Regularized multi-task learning. ACM SIGKDD Conferences on Knowledge Discovery and Data Mining 109–117.

Fischl, B., Liu, A., Dale, A., 2001. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. IEEE Trans. Med. Imaging 20, 70–80.

Fischl, B., Salat, D., 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., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Fischl, B., Salat, D., van der Kouwe, A.J., Makris, N., Segonne, F., Quinn, B., Dale, A., 2004. Sequence-independent segmentation of magnetic resonance images. Neuroimage 23, S69–S84.

Friedman, J., Hastie, T., Tibshirani, R., 2010. A Note on the Group Lasso and a Sparse Group Lasso. arXiv:1001.0736.

Gong, P., Ye, J., Zhang, C.-s., 2012. Multi-stage multi-task feature learning. Advances in Neural Information Processing Systems 1988–1996.

Greene, S.J., Killiany, R.J., 2010. Subregions of the inferior parietal lobule are affected in the progression to alzheimer's disease. Neurobiol. Aging 31 (8), 104–111.

Gu, B., Sheng, V.S., 2016. A robust regularization path algorithm for ν-support vector classification. IEEE Trans. Neural Netw. Learn. Syst. PP, 1–8.

Gu, B., Sheng, V.S., Tay, K.Y., Romano, W., Li, S., 2015. Incremental support vector learning for ordinal regression. IEEE Trans. Neural Netw. Learn. Syst. 26 (7), 1403–1416.

Guerrero, R., Ledig, C., Schmidt-Richberg, A., Rueckert, D., Alzheimer's Disease Neuroimaging Initiative (ADNI and others), 2017. Group-constrained manifold learning: application to AD risk assessment. Pattern Recognit. 63, 570–582.

Gutman, B.A., Wang, Y., Yanovsky, I., Hua, X., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M., 2015. Empowering imaging biomarkers of Alzheimer's disease. Neurobiol. Aging 36, S69–S80.

Jenatton, R., Mairal, J., Obozinski, G., Bach, F., 2011. Proximal methods for hierarchical sparse coding. J. Mach. Learn. Res. 12 (July), 2297–2334.

Khachaturian, Z., 1985. Diagnosis of Alzheimer's disease. Arch. Neurol. 42 (11), 11097–11105.

Kim, S., Xing, E.P., 2009. Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genet. 5 (8), e1000587.

Li, T., Wana, J., Zhang, Z., Yan, J., Kim, S., Risacher, S., Fang, S., Beg, M., Wang, L., Saykin, A., Shen, L., 2012. Hippocampus as a predictor of cognitive performance: comparative evaluation of analytical methods and morphometric measures. NIBAD's Workshop at MICCAI 133–144.

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

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

Meier, L., van de Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. J. Roy. Stat. Soc. Ser. B 70 (1), 53–71.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc.: Ser. B 72 (4), 417–473.

Nelder, J., Baker, R., 1972. Generalized Linear Models. Wiley Online Library.

Nesterov, Y., 2005. Smooth minimization of non-smooth functions. Math. Program. 103 (1), 127–152.

Parikh, N., Boyd, S., 2013. Proximal algorithms. Found. Trends Optim. 1 (3), 127–239.

Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., Initiative, A.D.N., et al., 2011. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res.: Neuroimag. 194 (1), 7–13.

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

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

Segonne, F., Pacheco, J., Fischl, B., 2007. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. IEEE Trans. Med. Imaging 26, 518–529.

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

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

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.D.N., et al., 2016. Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. Brain Struct. Funct. 221 (5), 2569–2587.

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

Visser, P., Verhey, F., Hofman, P., Scheltens, P., Jolles, J., 2002. Medial temporal lobe atrophy predicts alzheimer's disease in patients with minor cognitive impairment. J. Neurol. Neurosurg. Psychiatry 72 (4), 491–497.

Wan, J., Zhang, Z., Rao, B.D., Fang, S., Yan, J., Saykin, A.J., Shen, L., 2014. Identifying the neuroanatomical basis of cognitive impairment in alzheimer's disease by correlation-and nonlinearity-aware sparse Bayesian learning. IEEE Trans. Med. Imaging 33 (7), 1475–1487.

Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Shen, L., ADNI, 2011. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. International Conference on Computer Vision 6–13.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S., Saykin, A.J., Shen, L., 2012. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics 28 (2), 229–237.

Weiner, M., Aisen, P., Jack, C., Jagust, W., Trojanowski, J., Shaw, L., Saykin, A., Morris, J., Cairns, N., Beckett, L., et al., 2010. The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimer's Dement. 6 (3), 202–211.

Wimo, A., Winblad, B., Aguero-Torres, H., von Strauss, E., 2003. The magnitude of dementia occurrence in the world. Alzheimer Dis. Assoc. Disord. 17 (2), 63–67.

Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J., 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PloS ONE 6 (10), e25446.

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. J. Alzheimer's Dis. 51 (4), 1045–1056.

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. Neurobiol. Aging 36, S185–S193.

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

Yu, Y., 2013a. Better approximation and faster algorithm using the proximal average. Advances in Neural Information Processing Systems (NIPS). pp. 458–466.

Yu, Y., 2013b. On decomposing the proximal map. Advances in Neural Information Processing Systems (NIPS). pp. 91–99.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 68 (1), 49–67.

Yuan, L., Liu, J., Ye, J., 2013. Efficient methods for overlapping group lasso. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 35 (9), 2104–2116.

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

Zhou, J., Liu, J., Narayan, V.A., Ye, J., 2013. Modeling disease progression via multi-task learning. NeuroImage 78, 233–248.

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