Modeling Alzheimer's Disease Progression with Fused Laplacian Sparse Group Lasso

XIAOLI LIU, Northeastern University and University of Minnesota
PENG CAO, Northeastern University
ANDRÉ R. GONÇALVES, Lawrence Livermore National Laboratory
DAZHE ZHAO, Northeastern University
ARINDAM BANERJEE, University of Minnesota

Alzheimer's disease (AD), the most common type of dementia, not only imposes a huge financial burden on the health care system, but also a psychological and emotional burden on patients and their families. There is thus an urgent need to infer trajectories of cognitive performance over time and identify biomarkers predictive of the progression. In this article, we propose the multi-task learning with fused Laplacian sparse group lasso model, which can identify biomarkers closely related to cognitive measures due to its sparsity-inducing property, and model the disease progression with a general weighted (undirected) dependency graphs among the tasks. An efficient alternative directions method of multipliers based optimization algorithm is derived to solve the proposed non-smooth objective formulation. The effectiveness of the proposed model is demonstrated by its superior prediction performance over multiple state-of-the-art methods and accurate identification of compact sets of cognition-relevant imaging biomarkers that are consistent with prior medical studies.

CCS Concepts: **General and reference; • Applied computing → Imaging;**

Additional Key Words and Phrases: Alzheimer's disease, disease progression, multi-task learning, graph laplacian, ADMM

ACM Reference format:
Xiaoli Liu, Peng Cao, André R. Gonçalves, Dazhe Zhao, and Arindam Banerjee. 2018. Modeling Alzheimer's Disease Progression with Fused Laplacian Sparse Group Lasso. *ACM Trans. Knowl. Discov. Data* 12, 6, Article 65 (August 2018), 35 pages.
https://doi.org/10.1145/3230668

1 INTRODUCTION

Dementia poses a serious challenge to the aging society. Alzheimer's disease (AD) is the most common cause of dementia. AD is a gradually progressive syndrome that mainly affects memory

The research was supported by the National Natural Science Foundation of China (No.61502091), the Fundamental Research Funds for the Central Universities (Nos. N161604001 and N150408001). The research was also supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, and CNS- 1314560.
Authors' addresses: X. Liu, P. Cao, and D. Zhao, College of Computer Science and Engineering, Northeastern University, Shenyang, China; emails: neuxiaoliliu@gmail.com, caopeng@cse.neu.edu.cn, zhaodz@neusoft.com; A. R. Gonçalves, Lawrence Livermore National Laboratory, CA; email: goncalves1@llnl.gov; A. Banerjee, Computing Science & Engineering, University of Minnesota, Twin Cities; email: banerjee@cs.umn.edu.

https://doi.org/10.1145/3230668

function, ultimately culminating in a dementia state where all cognitive functions are affected. It is a devastating disease for those who are affected and presents a major burden to caregivers and society. The worldwide prevalence of AD is predicted to quadruple from 46.8 million in 2016 (Association 2016) to 131.5 million by 2050 according to Alzheimer's Disease Neuroimaging Initiative (ADNI's) World Alzheimer Report (Batsch and Mittelman 2015). Dementia also has a huge economic impact. Today, the total estimated worldwide cost of dementia is US $818 billion, and it will become a trillion dollar disease by 2018. The huge price of caring for AD patients has made it one of the most costly diseases in the developed countries. Caring for the disease also causes great physical, as well as psychological suffering on the caregivers.

Accurate diagnosis or cognitive performance prediction of AD is key to the development, assessment, and monitoring of new treatments for AD (Xu et al. 2015; Gao et al. 2017). Several studies have identified strong connection between patterns of brain atrophy and AD progression, measured by means of patient's cognitive characterization (Misra et al. 2009; Weiner et al. 2017). Many clinical/cognitive measures such as Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale cognitive sub-scale (ADAS-Cog) have been designed to evaluate the cognitive status of the patients and they have been used as an important criteria for clinical diagnosis of probable AD.

In the literature, machine learning models have been widely studied with focus on the prediction power of neuroimaging measures as biomarkers for inferring cognitive outcomes or tracking disease progression in the AD study. These data-oriented approaches seek to infer patient's cognitive and functional abilities from neuroimaging data, such as magnetic resonance imaging (MRI), positron emission tomography (PET) along with other biomarkers. The associated learning problem has commonly been posed as a classification, survival analysis, or a regression problem.

Classification-based models (Misra et al. 2009; Liu et al. 2013) aim to classify the patient state into a pre-defined set of disease stages, usually categorized as: *Cognitively Normal* (CN), *Mild Cognitive Impairment* (MCI) and AD. With survival analysis models (Doody et al. 2010; Vemuri et al. 2009), it is possible to answer different questions such as when a patient stage will turn from MCI to AD or patient's survival time.

There is a sizable literature on applying regression methods in the context of AD (Ye et al. 2012, 2008; Zhou et al. 2011, 2013). Several previous works have studied the relationship between the cognitive scores and possible risk factors such as age, ApoE gene, years of education, and gender (Tombaugh 2005; Ito et al. 2011). The relationship between cognitive scores and imaging markers based on MRI has been explored by correlating these features with baseline (BL) MMSE scores (Stonningtona et al. 2010; Frisoni et al. 2010). However, several existing models do not model correlation among multiple tasks, where the tasks can be different cognitive scores, or the same cognitive scores over time. When the tasks and their corresponding models are believed to be related, it is desirable to learn all of the models jointly rather than treating each task as independent of each other and fitting each model separately. It is known that there exist inherent correlations among different cognitive scores or the same cognitive score over time, since the underlying pathology is the same and there is disease progression over time. Thus, joint modeling of multiple tasks is expected to lead to better predictive ability.

Multi-task learning (MTL) is a statistical learning framework, which seeks to learn models for several tasks jointly. The idea of MTL is to utilize the intrinsic relationships among multiple related tasks in order to improve the generalization performance (Caruana 1997; Argyriou et al. 2008; Gonçalves et al. 2016). It has been commonly used to obtain better generalization performance than learning each task individually in the field of AD. The critical issues in MTL are to identify how the tasks are related and build learning models to capture such task relatedness with different assumptions. One approach to modeling multi-task relationship is to assume that all tasks

are related and the respective models are similar to each other. In Zhang et al. (2012), the prediction of different types of targets such as MMSE and ADAS-Cog is modeled as a MTL problem and all models are constrained to share a common set of features. Jing et al. (Wan et al. 2012) proposed a new sparse Bayesian MTL approach, which adaptively learns and exploits the correlation structure within each coefficient row in the multiple measurement vector model. Our previous work adapted sparse group lasso (SGL) to consider two-level hierarchy with feature-level and group-level sparsity and parameter coupling across tasks (Liu et al. 2016).

The focus of the current article is on MTL in the context of AD, where the tasks involve accurately predicting a given same cognitive score over multiple time steps, i.e., each task focuses on modeling a given cognitive score at a given time step, and different tasks focus on different time steps for the same cognitive score. For AD, such longitudinal data usually consists of measurements at a starting time point ($t = 0$), after 6 months ($t = 6$), after 12 months ($t = 12$), after 24 months ($t = 24$), and so on usually up to 48 months ($t = 48$). MTL with such longitudinal data has been considered in the literature. In Zhou et al. (2011), a MTL is used to model the longitudinal disease progression using the temporal group lasso (TGL) regularization to capture task relatedness. TGL constrains the models at all time points to select a common set of features, and hence may miss the temporal patterns and variability of the biomarkers during disease progression. Zhou et al. (2013) proposed convex fused sparse group lasso (cFSGL), which allows the simultaneous selection of a common set of biomarkers at all time points and the selection of a specific set of biomarkers at different time points using the SGL penalty, and in the meantime incorporates the temporal smoothness using the fused lasso penalty (Tibshirani et al. 2005). The proposed formulation is challenging to solve due to the use of several non-smooth penalties. The authors show that the proximal operator associated with the proposed formulation exhibits a certain decomposition property and can be computed efficiently; thus cFSGL can be solved using a suitable variant of the accelerated gradient method (AGM). Results demonstrate the effectiveness of the proposed MTL formulations for disease progression in comparison with single-task learning algorithms, including ridge and lasso regression. The limitation of TGL and cFSGL is that the fused lasso only consider two successive time points, potentially missing out on helpful task dependencies beyond the immediate neighbors. In essence, if every task (time step) is viewed as nodes of a graph and edges determine task dependencies, cFSGL use a graph where there are edges between tasks, t and $(t + 1)$, $t = 1, \ldots, T - 1$, but there are no other edges.

In this article, we present a general framework called Fused Laplacian Sparse Group Lasso (FL-SGL), which in principle allows more general weighted (undirected) dependency graphs among the tasks. We consider a regularized MTL formulation encouraging related tasks to have similar parameters, where the regularization depends on suitable structured sparsity based on the graph Laplacian of the task dependency matrix. In this article, we consider weighted task dependency graphs based on a Gaussian kernel over the time steps, which yields a fully connected graph with decaying weights. We consider different bandwidths for the Gaussian kernel yielding qualitatively different task dependencies. In particular, for small bandwidths, we obtain task dependencies mainly among nearby neighbors, and for large bandwidths, dependencies are across all neighbors. Note that while we specifically focus on Gaussian kernels for this article, one can consider fused Laplacian MTL formulations for any task dependency graph, including recent approaches where the task dependency is also learned from the data (Zhang and Yeung 2010; Rai et al. 2012; Gonçalves et al. 2016).

The FL-SGL MTL formulation outlined above are in the form of non-smooth optimization problems. We present two alternating direction method of multipliers (ADMM)-type algorithms for solving the formulations. In recent years, ADMM has become popular, since it is often easy to parallelize such algorithms. Further, ADMM has been successfully applied to a variety of

nonconvex optimization problems, including L_1-regularization (Yang and Zhang 2011), group lasso-regularization (Deng et al. 2013), and total variation (TV) regularization problems (Wang et al. 2008). In this article, we consider two variants, respectively, based on multi-block ADMM and traditional two-block ADMM (Boyd et al. 2011; He et al. 2012). In both variants, we use inexact ADMM, which yields fast closed form updates in each iteration and which have been shown to has the same rate-of-convergence as exact updates (Boyd et al. 2011). While the algorithms are applied to Gaussian kernel based task dependency structures considered in this article, the ADMM algorithms can be applied to FL-SGL formulations with any graph capturing the task dependency structure.

We perform extensive experiments using longitudinal data from the ADNI. Five types of cognitive scores are considered. We, then, empirically evaluate the performance of the proposed FL-SGL methods along with several BL methods, including ridge regression, lasso, and the recently developed cFSGL (Zhou et al. 2013). The quantitative results indicate that FL-SGL outperforms the BLs on the aggregated performance, i.e., predictive performance on the entire longitudinal data for a test subject, and the improvements are statistically significant. Further, based on the FL-SGL models, stable MRI features which significant predictive power are identified using stability selection (Meinshausen and Bühlmann 2010), and keyregions of interest (ROIs) contributing these MRI features are discussed. We also present brain maps highlighting the top ROIs selected by the FL-SGL algorithm. Finally, in addition to the MRI features, we use demographic and genetic information for FL-SGL as well as the BL models. While the additional features improve the predictive performance of all the models, FL-SGL shows substantial improvements and continue to significantly outperform the BLs on the aggregated performance.

The rest of the article is organized as follows. In Section 2, we present the FL-SGL model for MTL. In Section 3, we discuss details of the two ADMM algorithms proposed for the FL-SGL models. We present experimental results on ADNI data in Section 4, and conclude in Section 5.

2 MULTI-TASK LEARNING WITH FUSED LAPLACIAN SPARSE GROUP LASSO (FL-SGL)

Consider a MTL problem over T tasks, where each task corresponds to a time point $t = 1, \ldots, T$. For each time point t, we consider a regression task based on data $(\mathbf{y}_t, X_t)$, where $X_t \in \mathbb{R}^{n \times p}$ denotes the matrix of covariates, p is the number of covariates and n is the number of samples, shared across all the tasks, and $\mathbf{y}_t \in \mathbb{R}^n$ is the matrix of responses. Let $\Theta \in \mathbb{R}^{p \times T}$ denote the regression parameter matrix over all tasks, so that column $\boldsymbol{\theta}_t \in \mathbb{R}^p$ corresponds to the parameters for the task in time step t. In the context of AD, $\mathbf{y}_t$ corresponds to a specific cognitive score at time step t for n patients, so the responses $\mathbf{y}_t, t = 1, \ldots, T$ over time (tasks) measure the progression of the cognitive score. The question of interest is: can we model the progression of the cognitive score based on the covariates, which are based on suitable brain images and other features?

We pose the MTL problem for Θ such that two goals are accomplished: each θ_t accomplishes low regression error for each task t, and "nearby" θ_t are coupled to be similar, since the "nearby" tasks are temporarily related. The notion of "nearby" needs to be suitably defined, and the current work makes novel contributions on this aspect. For MTL problems, where the tasks are over time t, a popular choice is to use a fused lasso style coupling where one encourages the difference $\boldsymbol{\gamma}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}$ to be sparse (Tibshirani et al. 2005). It is reasonable to assume that the scores between two successive time points should be close (Huang et al. 2016). However, in clinical practice, this assumption may not always hold. Figure 1 shows how the real ADAS, MMSE, and RAVLT.TOTAL scores of several subjects from our dataset changed over the years. Steady periods and sharp declines intertwined with occasional improvements. This indicates that longitudinal clinical scores may have more complex evolution than a simple linear trend with local temporal

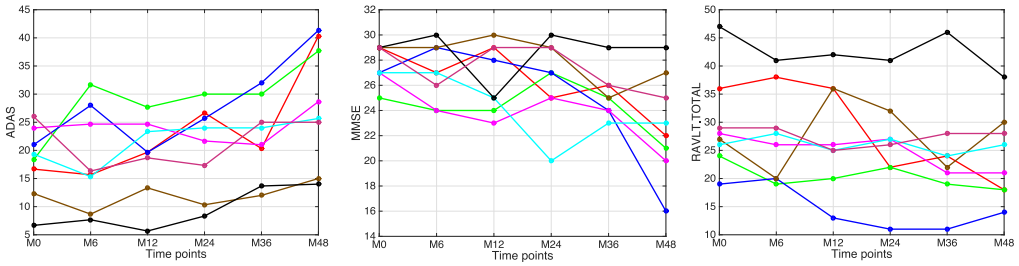


Fig. 1. The change patterns of several patients' cognitive scores over the six time points: (a) ADAS, (b) MMSE, and (c) RAVLT.TOATL. The different colors indicate different patients from our dataset.

correlations. In this article, we take a more general perspective inspired by local non-parametric regression, in particular kernel-based linear smoothers such as the Nadaraya–Watson kernel estimator (Wasserman 2006). From such a perspective, we model the local approximation as

$$\hat{\boldsymbol{\theta}}_t = \sum_{\substack{\ell=1 \\ \ell \neq t}}^{T} w_{\ell,t} \boldsymbol{\theta}_\ell, t = 1, \ldots, T \tag{1}$$

$$w_{\ell,t} = \frac{\exp\left(-\frac{(\ell-t)^2}{\sigma^2}\right)}{\sum_{\substack{\ell'=1 \\ \ell' \neq t}}^{T} \exp\left(-\frac{(\ell'-t)^2}{\sigma^2}\right)}, \ell = 1, \ldots, T, \ell \neq t, \tag{2}$$

where $\sigma \geq 0$ is a constant bandwidth parameter. Based on such an approximation, our general MTL formulations focus on encouraging sparsity on the residual

$$\boldsymbol{\gamma}_t = \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell \neq t}}^{T} w_{\ell,t} \boldsymbol{\theta}_\ell, t = 1, \ldots, T. \tag{3}$$

With $\Gamma \in \mathbb{R}^{p \times T}$ denoting the matrix of residuals with columns $\boldsymbol{\gamma}_t \in \mathbb{R}^p$, we pose the MTL problem as the following constrained optimization problem:

$$\begin{aligned} &\min_{\Theta, \Gamma} \sum_{t=1}^{T} \|\mathrm{y}_t - X_t \boldsymbol{\theta}_t\|^2 + R_{\lambda_2}^{\lambda_1}(\Theta) + \lambda_3 \|\Gamma\|_1 \\ &\text{s.t. } \boldsymbol{\gamma}_t = \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell \neq t}}^{T} w_{\ell,t} \boldsymbol{\theta}_\ell, t = 1, \ldots, T, \end{aligned} \tag{4}$$

where $R_{\lambda_2}^{\lambda_1}(\Theta)$ is the combination of lasso and group lasso penalties, also known as the SGL penalty, which allows simultaneous joint feature selection for all tasks and selection of a specific set of features for each task (Yuan et al. 2013). In particular

$$R_{\lambda_2}^{\lambda_1}(\Theta) = \lambda_1 \|\Theta\|_1 + \lambda_2 \|\Theta\|_{2,1}, \tag{5}$$

where $\|\Theta\|_1$ is the Lasso penalty and $\|\Theta\|_{2,1} = \sum_{j=1}^{p} \|\boldsymbol{\theta}_j\|, \boldsymbol{\theta}_j \in \mathbb{R}^T$ is the group Lasso penalty considering groups across time for each feature j, encouraging the regression models at different time points to share a common set of features. In the formulation, $\lambda_1, \lambda_2, \lambda_3 > 0$ are the regularization parameters that are fixed, and will be chosen using cross validation.

In Equation (4), Γ is a linear transformation of Θ, introduced to capture the temporal smoothness of the cognitive scores at different time points. For ease of exposition, assume $p = 1$ so that

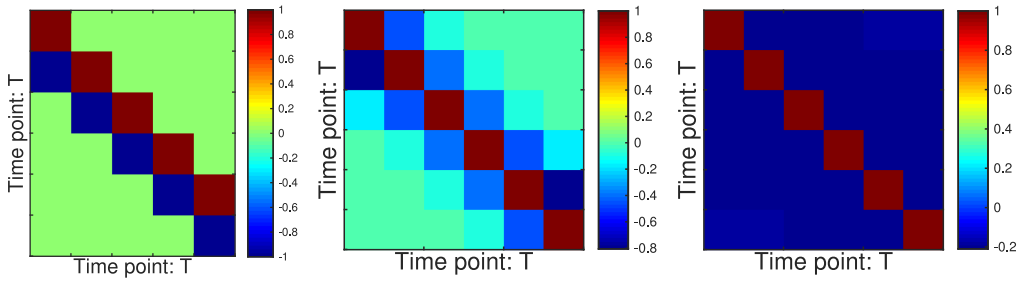


Fig. 2. [Best viewed in color] Illustration of three different fused regularizations: (a) Standard fused lasso, (b) Gaussian kernel weighted fused lasso with $\sigma = 1$, and (c) Gaussian kernel weighted fused lasso with $\sigma = 10$.

$\boldsymbol{\theta}_t, \gamma_t \in \mathbb{R}$ and $\Theta, \Gamma \in \mathbb{R}^{1\times T}$. Then, in order to penalizes large deviations between predictions at multiple time points, it can be defined as $\Gamma = \Theta\mathbb{D}$, $\mathbb{D} \in \mathbb{R}^{T\times T}$ is the transformation matrix.

In standard fused lasso penalty (Zhou et al. 2013), it is assumed that the difference of the cognitive scores between two successive time points is relatively small. In order to penalize large deviations between predictions at neighboring time points. The temporal smoothness term can be expressed as follows:

$$\boldsymbol{\gamma}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}, t = 1, \ldots, T-1 \tag{6}$$

Here, the sparse matrix $\mathbb{H} \in \mathbb{R}^{T\times(T-1)}$ is defined as follows: $\mathbb{H}_{ti} = 1$ if $t = i$, $\mathbb{H}_{ti} = -1$ if $t = i+1, i = 1 \ldots T-1$, and $\mathbb{H}_{ti} = 0$ otherwise (see Figure 2(a)).

In our work, the SGL aspect of the formulation is evident from Equation (4), and the fused aspect comes from putting sparsity on the residual $\boldsymbol{\gamma}_t$. The proposed fused penalty can be written in terms of a graph Laplacian and we use $\mathbb{L}$ to denote the transformation matrix. When $p = 1$[1], one can write

$$\Gamma = \Theta\mathbb{L} \Rightarrow \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_T \end{bmatrix}^T = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_T \end{bmatrix}^T \begin{bmatrix} 1 & -w_{1,2} & -w_{1,3} & \cdots & w_{1,T} \\ -w_{2,1} & 1 & -w_{2,3} & \cdots & -w_{2,T} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -w_{T,1} & -w_{T,2} & -w_{T,3} & \cdots & 1 \end{bmatrix},$$

where the matrix $\mathbb{L}$ is symmetric, since $w_{t,\ell} = w_{\ell,t}$, being a function of $|t-\ell|$. Note that $\mathbb{L}$ corresponds the graph Laplacian matrix for an undirected (complete) graph with T nodes, with the edge between nodes t and ℓ having a weight $w_{t,\ell} = w_{|t-\ell|}$. In the sequel, we will drop the double subscripts and directly use $w_{|t-\ell|}$. Finally, note that the graph Laplacian perspective continues to hold when we consider the general case where $\boldsymbol{\theta}_t, \gamma_t \in \mathbb{R}^p$. We will return to this perspective in Section 3.3.

In our current work, we use a Laplacian where the weights are determined by a Gaussian kernel as defined in Equation (2), where σ is the kernel bandwidth that must be defined. When σ is small, the Gaussian curve decays quickly, and so the weights $w_{|t-\ell|}$ decay quickly as $|t-\ell|$ increases; on the other hand, when σ is large, the Gaussian curve decays gradually, and the weights $w_{|t-\ell|}$ decay slowly as $|t-\ell|$ increases. The graph Laplacian matrix $\mathbb{L}$ is illustrated in Figure 2(b) and (c). As shown in Figure 1, one time point is not simply linear related to other time points. We consider two values of σ, corresponding to these two regimes: $\sigma = 1$ (Figure 2(b)), where the weights decay

[1]Again, we are assuming $\gamma_t, \theta_t \in \mathbb{R}$ to illustrate the idea, noting that the same ideas hold for the general matrix case. We discuss the general matrix form in Section 3.3.

quickly, and $\sigma = 10$ (Figure 2(c)), where the weights decay gradually, and are practically uniform across the entire range.

In the next section, we develop ADMM-based algorithms to solve these formulations. We call the model corresponding to the two Laplacians in Figure 2(b) and (c) FL-SGL. To distinguish the two variants, we refer to the model with Laplacian matrix with $\sigma = 1$, as shown in Figure 2(b), as FL-SGL1, and the one with $\sigma = 10$, as shown in Figure 2(c), as FL-SGL2. The ADMM algorithms we describe next are applicable to both of these models. As discussed earlier, Figure 2(a) corresponds to the standard fused lasso, the corresponding MTL problem can be solved using cFSGL (Zhou et al. 2013). We consider an alternative approach, which we refer to as Fused Sparse Group Lasso (F-SGL), for solving the problem corresponding to standard fused lasso. Note that the formulations corresponding to cFSGL and F-SGL are the same, the only difference is the optimization procedures used. For cFSGL, the optimization problem is solved by the AGM, which computes the proximal operator in two stages including a fused lasso and a group lasso for multiple tasks. The fused lasso stage is solved by the fused lasso signal approximator (Liu et al. 2010), in which the Subgradient Finding Algorithm is used.

3 ADMM FOR LEARNING FL-SGL MODELS

The unconstrained optimization problem in Equation (4) can be difficult to optimize directly due to the non-smooth and coupling terms. A simple special case of the FL-SGL formulation is fused Lasso for which a variety of optimization algorithms have been studied (Beck and Teboulle 2009; Bach et al. 2012). In the current context, we propose ADMM-based algorithms for the general FL-SGL formulations, which can seamlessly handle different variations of the fused Laplacian. Further, we focus on algorithms which can suitably decouple the updates for $\boldsymbol{\theta}_t \in \mathbb{R}^p$, so that the updates for $\boldsymbol{\theta}_t$ can be done in parallel rather than jointly in terms of Θ.

Start by noting that the optimization problem in Equation (4) can be reformulated as the following linearly constrained optimization problem:

$$\begin{aligned} &\min_{\Theta,\Gamma,Q,\Pi} \sum_{t=1}^{T} \frac{1}{2}\|\mathbf{y}_t - X_t\boldsymbol{\theta}_t\|^2 + R^{\lambda_1}_{\lambda_2}(Q) + \lambda_3\|\Pi\|_1 \\ &\text{s.t. } \Theta - Q = 0,\ \boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\boldsymbol{\theta}_\ell - \boldsymbol{\gamma}_t = 0, t = 1,\ldots,T, \Gamma - \Pi = 0. \end{aligned} \tag{7}$$

Note that there are four variables $\Theta, \Gamma, Q, \Pi \in \mathbb{R}^{p\times T}$ in the optimization. In the feasible set, we have $Q = \Theta$. Further, $\Gamma = [\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_T]$ captures the fused Laplacian residuals so that in the feasible set $\boldsymbol{\gamma}_t = \boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\boldsymbol{\theta}_\ell$ and also $\Pi = \Gamma$.

The augmented Lagrangian for the problem is given by

$$\begin{aligned} L_\rho(\Theta,\Gamma,Q,\Pi,S,U,V) &= \sum_{t=1}^{T} \frac{1}{2}\|\mathbf{y}_t - X_t\boldsymbol{\theta}_t\|^2 + R^{\lambda_1}_{\lambda_2}(Q) + \lambda_3\|\Pi\|_1 \\ &+ \mathrm{Tr}(S^T(\Theta - Q)) + \frac{\rho}{2}\|\Theta - Q\|^2 \\ &+ \sum_{t=1}^{T}\left\{\mathbf{u}_t^T\left(\boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\boldsymbol{\theta}_\ell - \boldsymbol{\gamma}_t\right) + \frac{\rho}{2}\left\|\boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\boldsymbol{\theta}_\ell - \boldsymbol{\gamma}_t\right\|^2\right\} \\ &+ \mathrm{Tr}(V^T(\Gamma - \Pi)) + \frac{\rho}{2}\|\Gamma - \Pi\|^2, \end{aligned} \tag{8}$$

where $S, U, V \in \mathbb{R}^{p \times T}$ are the Lagrangian multipliers corresponding to the constraints $\Theta - Q = 0$, $\boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell \neq t}}^{T} w_{|t-\ell|}\boldsymbol{\theta}_\ell - \boldsymbol{\gamma}_t = 0$, and $\Gamma - \Pi = 0$, and $\rho > 0$ is penalty parameter effectively determining the step-size for dual ascent in ADMM (Boyd et al. 2011).

For convenience, let

$$h(\Theta) = \frac{\rho}{2} \sum_{t=1}^{T} \left\| \boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell \neq t}}^{T} w_{|t-\ell|}\boldsymbol{\theta}_\ell - \boldsymbol{\gamma}_t \right\|^2. \tag{9}$$

While the original objective function does not have terms with interactions between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_\ell, \ell \neq t$, the augmented Lagrangian do have such terms, and such interactions are captured in $h(\Theta)$. In order to decouple the $\boldsymbol{\theta}_t$ updates, we will perform the ADMM updates by suitably linearizing $h(\Theta)$ around the current iterate Θ^k. Let $\mathbf{h}_t^k = \nabla_{\boldsymbol{\theta}_t} h(\Theta^k)$ denote the gradient with respect to $\boldsymbol{\theta}_t$. Recent work on Bregman ADMM (Wang and Banerjee 2014) and related work on inexact ADMM (Yang and Zhang 2011; Boyd et al. 2011) have shown that ADMM updates with such linearization continue to work.

3.1 Linearized Multi-Block ADMM

The main idea here is to linearize the $h(\Theta)$ term, so that the coupling between $\boldsymbol{\theta}_t$ is not there and the individual $\boldsymbol{\theta}_t$ can be updated in parallel. With such a linearization, we simply update the primal and dual variables based on a multi-block ADMM algorithm (Hong and Luo 2017; Wang et al. 2014). While theoretical performance of multi-block ADMM is still a topic of active research (Chen et al. 2016; Deng et al. 2017; Hong et al. 2014), in the context of the current work we focus on extensively evaluating the empirical performance of the algorithm for the AD dataset (see Section 4). We also consider more standard two-block ADMM (see Section 3.2) for the problem along with comparisons with the proposed multi-block approach both in optimization performance as well as predictive performance on AD data (see Section 4).

Update $\boldsymbol{\theta}_t^{k+1}$: From the augmented Lagrangian in Equation (8), with the linearization of $h(\Theta^K)$, the update for each $\boldsymbol{\theta}_t$ can be done in parallel. In particular, the update involves solving the following unconstrained quadratic objective, which can be done efficiently using Cholesky decomposition as discussed in Boyd et al. (2011)

$$\boldsymbol{\theta}_t^{k+1} = \underset{\boldsymbol{\theta}_t}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y}_t - X_t\boldsymbol{\theta}_t\|^2 + (\mathbf{s}_t^k)^T\boldsymbol{\theta}_t + \frac{\rho}{2}\|\boldsymbol{\theta}_t - \mathbf{q}_t\|^2 + (\mathbf{u}_t^k + \mathbf{h}_t^k)^T\boldsymbol{\theta}_t + \frac{\rho_1}{2}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^k\|^2, \tag{10}$$

where $\rho_1 > 0$ is a suitably chosen constant. In particular, since $h(\Theta)$ is smooth and has Lipschitz continuous gradients with constant ν under 2-norm, it suffices to have $\rho_1 \geq 2\nu$ (Wang and Banerjee 2014). As we show in Section 3.3, we can choose any $\rho_1 \geq 3\rho$ to satisfy the requirement. We use $\rho_1 = 3\rho$ for our experiments.

Update $\boldsymbol{\gamma}_t^{k+1}$: From the augmented Lagrangian in Equation (8), the update for Γ can be done in parallel for each $\boldsymbol{\gamma}_t$ and the updates need to solve the following unconstrained quadratic problems:

$$\boldsymbol{\gamma}_t^{k+1} = \underset{\boldsymbol{\gamma}_t}{\operatorname{argmin}} \frac{\rho}{2} \left\| \boldsymbol{\gamma}_t - \left(\boldsymbol{\pi}_t^k + \boldsymbol{\theta}_t^{k+1} - \sum_{\substack{\ell=1 \\ \ell \neq t}}^{T} w_{|t-\ell|}\boldsymbol{\theta}_\ell^{k+1} \right) \right\|^2 - (\mathbf{u}_t^k - \mathbf{v}_t^k)^T\boldsymbol{\gamma}_t. \tag{11}$$

Given the above form, note that the updates for $\boldsymbol{\gamma}_t$ can in fact be done in an element-wise parallel manner, i.e., by solving scalar unconstrained quadratic problems, which have closed-form solutions.

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

$$Q^{k+1} = \underset{Q}{\operatorname{argmin}} \frac{\rho}{2}\|Q - \Theta^{k+1}\|^2 - (S^k)^T Q + R^{\lambda_1}_{\lambda_2}(Q), \tag{12}$$

which is equivalent to computing the proximal operator for $R^{\lambda_2}_{\lambda_1}(\cdot)$. In particular, we need to solve

$$\Psi^{\lambda_1/\rho}_{\lambda_2/\rho}(O^{k+1}) = \underset{Q}{\operatorname{argmin}} \left\{ R^{\lambda_1/\rho}_{\lambda_2/\rho}(Q) + \frac{1}{2}\|Q - O^{k+1}\|^2 \right\}, \tag{13}$$

where $O^{k+1} = \Theta^{k+1} + \frac{1}{\rho}S^k$. The goal is to be able to compute $Q^{k+1} = \Psi^{\lambda_1/\rho}_{\lambda_2/\rho}(O^{k+1})$ efficiently. It can be shown (Yu 2013a, 2013b) that the proximal operator for the composite regularizer can be computed efficiently in two steps, as outlined below:

$$\Omega^{k+1} = \Psi^{\lambda_1/\rho}_0(O^{k+1}) \tag{14a}$$

$$Q^{k+1} = \Psi^0_{\lambda_2/\rho}(\Omega^{k+1}) = \Psi^{\lambda_1/\rho}_{\lambda_2/\rho}(O^{k+1}). \tag{14b}$$

Both of these steps can be executed efficiently using suitable extensions of soft-thresholding. The update in Equation (14a) can be computed by the soft-thresholding operator $\zeta_{\lambda_1/\rho}(O^{k+1})$, which is defined as

$$\zeta_\lambda(x) = \operatorname{sign}(x)\max(|x| - \lambda, 0).$$

Next, we focus on the update Equation (14b), which can be written as

$$Q^{k+1} = \underset{Q}{\operatorname{argmin}} \left\{ \frac{\lambda_2}{\rho}\|Q\|_{2,1} + \frac{1}{2}\|Q - \Omega^{k+1}\| \right\}.$$

The row-wise updates can be done by soft-thresholding as

$$q_i = \frac{\max\left\{\|\omega_i\|_2 - \frac{\lambda_2}{\rho}, 0\right\}}{\|\omega_i\|_2},$$

where q_i and ω_i are the ith rows of Q^{k+1} and Ω^{k+1}, respectively.

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

$$\Pi^{k+1} = \underset{\Pi}{\operatorname{argmin}} \frac{\rho}{2}\|\Pi - \Gamma^{k+1}\|^2 - \operatorname{Tr}((V^k)^T\Pi) + \lambda_3\|\Pi\|_1, \tag{15}$$

which is equivalent to computing the proximal operator for L_1-norm. In particular, the problem can be solved in closed form using soft-thresholding operator as

$$\Pi^{k+1} = \zeta_{\lambda_3/\rho}\left(\Gamma^{k+1} + \frac{1}{\rho}V^k\right). \tag{16}$$

Dual Updates for S, U, V: Following standard ADMM dual updates, the updates for the dual variables for our setting are as follows:

$$S^{k+1} = S^k + \rho(\Theta^{k+1} - Q^{k+1}) \tag{17}$$

$$\mathbf{u}^{k+1}_t = \mathbf{u}^k_t + \rho\left(\boldsymbol{\theta}^{k+1}_t - \sum_{\substack{\ell=1 \\ \ell\neq k}}^{T} w_{|t-\ell|}\boldsymbol{\theta}^{k+1}_\ell - \boldsymbol{\gamma}^{k+1}_t\right), t = 1, \ldots, T \tag{18}$$

$$V^{k+1} = V^k + \rho(\Gamma^{k+1} - \Pi^{k+1}). \tag{19}$$

All the dual updates can be done in an element-wise parallel manner.

3.2 Linearized Two-Block ADMM

The constrained optimization problem in Equation (7) can be equivalently posed as follows:

$$\begin{aligned} &\min_{\Theta,\Gamma,Q,\Pi} \sum_{t=1}^{T} \frac{1}{2}\|\mathbf{y}_t - X_t\boldsymbol{\theta}_t\|^2 + R_{\lambda_2}^{\lambda_1}(Q) + \lambda_3\|\Pi\|_1 \\ &\text{s.t. } \Theta - Q = 0,\ \mathbf{q}_t - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\mathbf{q}_\ell - \boldsymbol{\gamma}_t = 0,\ \Gamma - \Pi = 0. \end{aligned} \tag{20}$$

Note that the residual is now defined in terms of Q, not Θ, so there is no coupling between Θ and Γ. As a result, the above problem can be solved as a basic ADMM with two sets of variables $X = (\Theta, \Gamma)$, since Θ and Γ can be updated in parallel, and $Z = (Q, \Pi)$ since Q and Π can be updated independently.

The augmented Lagrangian for the problem is given by

$$\begin{aligned} L_\rho(\Theta,\Gamma,Q,\Pi,S,U,V) &= \sum_{t=1}^{T} \frac{1}{2}\|\mathbf{y}_t - X_t\boldsymbol{\theta}_t\|^2 + R_{\lambda_2}^{\lambda_1}(Q) + \lambda_3\|\Pi\|_1 \\ &+ \mathrm{Tr}(S^T(\Theta - Q)) + \frac{\rho}{2}\|\Theta - Q\|^2 \\ &+ \sum_{t=1}^{T}\left\{\mathbf{u}_t^T\left(\mathbf{q}_t - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\mathbf{q}_\ell - \boldsymbol{\gamma}_t\right) + \frac{\rho}{2}\left\|\mathbf{q}_t - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\mathbf{q}_\ell - \boldsymbol{\gamma}_t\right\|^2\right\} \\ &+ \mathrm{Tr}(V^T(\Gamma - \Pi)) + \frac{\rho}{2}\|\Gamma - \Pi\|^2. \end{aligned} \tag{21}$$

As before, we use linearization to simplify the Q updates. Let

$$h(Q) = \frac{\rho}{2}\sum_{t=1}^{T}\left\|\mathbf{q}_t - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\mathbf{q}_\ell - \boldsymbol{\gamma}_t\right\|^2. \tag{22}$$

Let $\mathbf{h}_t^k = \nabla_{\mathbf{q}_t} h(Q^k)$ denote the gradient w.r.t. $\mathbf{q}_t$, and let H^k be the collection of all such gradients.

Update $\boldsymbol{\theta}_t^{k+1}$: The update involves solving the following quadratic objective, which can be done efficiently using Cholesky decomposition as discussed in Boyd et al. (2011)

$$\boldsymbol{\theta}_t^{k+1} = \underset{\boldsymbol{\theta}_t}{\mathrm{argmin}}\ \frac{1}{2}\|\mathbf{y}_t - X_t\boldsymbol{\theta}_t\|^2 + (\mathbf{s}_t^k)^T\boldsymbol{\theta}_t + \frac{\rho}{2}\|\boldsymbol{\theta}_t - \mathbf{q}_t\|^2. \tag{23}$$

Update $\boldsymbol{\gamma}_t^{k+1}$: The update for $\boldsymbol{\gamma}_t$ needs to solve the following quadratic problem:

$$\boldsymbol{\gamma}_t^{k+1} = \underset{\boldsymbol{\gamma}_t}{\mathrm{argmin}}\ \frac{\rho}{2}\left\|\gamma_t - \left(\boldsymbol{\pi}_t^k + \mathbf{q}_t^k - \sum_{\substack{\ell=1 \\ \ell\neq t}}^{T} w_{|t-\ell|}\mathbf{q}_\ell^k\right)\right\|^2 - (\mathbf{u}_t^k - \mathbf{v}_t^k)^T\boldsymbol{\gamma}_t. \tag{24}$$

Note that the Γ^{k+1} update does not depend on Θ^{k+1}, and hence the two updates can be performed in parallel. We can treat the overall update as that of one variable $X^{k+1} = (\Theta^{k+1}, \Gamma^{k+1})$. Further, these updates can be executed in a component-wise parallel manner for each $\boldsymbol{\theta}_t^{k+1}$ and $\boldsymbol{\gamma}_t^{k+1}$, $t = 1, \ldots, T$. In fact, the $\boldsymbol{\gamma}_t^{k+1}$ updates can be done in an element-wise parallel manner.

Update Q: The update for Q is based on linearization and we need to compute the proximal operator for $R^{\lambda_2}_{\lambda_1}(\cdot)$:

$$Q^{k+1} = \underset{Q}{\operatorname{argmin}} \frac{\rho}{2}\|Q - \Theta^{k+1}\|^2 - \operatorname{Tr}((S^k + H^k + A^k)^T Q) + \frac{\rho_1}{2}\|Q - Q^k\|^2 + R^{\lambda_1}_{\lambda_2}(Q), \tag{25}$$

where A^k is such that $\operatorname{Tr}((A^k)^T Q) = \sum_{t=1}^{T} (\mathbf{u}_t^k)^T (\mathbf{q}_t - \sum_{\ell \neq t} w_{|t-\ell|}\mathbf{q}_\ell)$, and $\rho_1 > 0$ is a suitably chosen constant. In particular, since $h(Q)$ is smooth and has Lipschitz continuous gradients say with constant ν under 2-norm, it suffices to have $\rho_1 \geq 2\nu$ (Wang and Banerjee 2014). The overall expression in Equation (25) can be simplified to get in the form of a proximal operator computation for $R^{\lambda_2}_{\lambda_1}(Q)$, as discussed in Section 3.1.

Update Π: For the update Π, we need to compute the proximal operator for L_1-norm, as discussed in Section 3.1

$$\Pi^{k+1} = \underset{\Pi}{\operatorname{argmin}} \frac{\rho}{2}\|\Pi - \Gamma^{k+1}\|^2 - \operatorname{Tr}((V^k)^T \Pi) + \lambda_3\|\Pi\|_1. \tag{26}$$

Note that the Π^{k+1} update does not depend on Q^{k+1}, and hence the two updates can be performed in parallel. We can treat the overall update as that of one variable $Z^{k+1} = (Q^{k+1}, \Pi^{k+1})$.

Since the primal updates can be viewed as sequentially updating two variables $X = (\Theta, \Gamma)$ and $Z = (Q, \Pi)$, the resulting algorithm is just basic ADMM with a linearization, which has the same rate of convergence (Boyd et al. 2011).

Dual Updates S, U, V: Following standard ADMM dual updates, the updates for the dual variables for our setting are as follows:

$$S^{k+1} = S^k + \rho(\Theta^{k+1} - Q^{k+1}) \tag{27}$$

$$\mathbf{u}_t^{k+1} = \mathbf{u}_t^k + \rho\left(\mathbf{q}_t^{k+1} - \sum_{\substack{\ell=1 \\ \ell \neq k}}^{T} w_{|t-\ell|}\mathbf{q}_\ell^{k+1} - \boldsymbol{\gamma}_t^{k+1}\right), t = 1, \ldots, T \tag{28}$$

$$V^{k+1} = V^k + \rho(\Gamma^{k+1} - \Pi^{k+1}). \tag{29}$$

As before, all dual updates can be done in an element-wise parallel manner.

3.3 Lipschitz Constant for Linearization

Recall that $h(\Theta) = \frac{\rho}{2}\sum_{t=1}^{T} \|\boldsymbol{\theta}_t - \sum_{\substack{\ell=1 \\ \ell \neq t}}^{T} w_{|t-\ell|}\boldsymbol{\theta}_\ell - \boldsymbol{\gamma}_t\|^2$. For ease of exposition, we assume uniform weights, i.e., $w_{|t-\ell|} = \frac{1}{T-1}$ which corresponds to the bandwidth $\sigma \to \infty$. The analysis we present here can be straight-forwardly extended to general weights $w_{|t-\ell|}$ as defined in Equation (2).

Since $h(\Theta)$ has Lipschitz gradients, there is a constant ν such that for any $\Theta, \tilde{\Theta}$, we have

$$\|\nabla h(\Theta) - \nabla h(\tilde{\Theta})\|_2 \leq \nu\|\Theta - \tilde{\Theta}\|_2.$$

In this section, our focus is on characterizing ν, and to show that $\nu \leq 3\rho$ so that the linearizations used in Section 3.1 are well justified. Further, since the linearization used in 3.2 is essentially the same but based on $h(Q)$, the same function with Q as the argument, the analysis also holds for $h(Q)$.

Recall that $\Theta \in \mathbb{R}^{p\times T}$ and $\boldsymbol{\theta}_t \in \mathbb{R}^p$. A direct calculation shows

$$\frac{1}{\rho}\nabla_{\boldsymbol{\theta}_t} h(\Theta) = \left(\boldsymbol{\theta}_t - \frac{1}{T-1}\sum_{\substack{\ell=1\\ \ell\neq t}}^{T}\boldsymbol{\theta}_\ell - \boldsymbol{\gamma}_t\right) - \frac{1}{T-1}\sum_{\substack{\ell=1\\ \ell\neq t}}^{T}\left(\boldsymbol{\theta}_\ell - \frac{1}{T-1}\sum_{\substack{\ell'=1\\ \ell'\neq \ell}}^{T}\boldsymbol{\theta}_{\ell'} - \boldsymbol{\gamma}_\ell\right)$$
$$= \left(1+\frac{1}{T-1}\right)\left(\boldsymbol{\theta}_t - \frac{1}{T-1}\sum_{\substack{\ell=1\\ \ell\neq t}}^{T}\boldsymbol{\theta}_\ell\right) - \left(\boldsymbol{\gamma}_t - \frac{1}{T-1}\sum_{\substack{\ell=1\\ \ell\neq t}}^{T}\boldsymbol{\gamma}_\ell\right).$$

It is important to note that the second term involving $\boldsymbol{\gamma}_t, \boldsymbol{\gamma}_\ell, \ell \neq t$ will be exactly the same for $\frac{1}{\rho}\nabla_{\boldsymbol{\theta}_t} h(\Theta)$ and $\frac{1}{\rho}\nabla_{\tilde{\boldsymbol{\theta}}_t} h(\tilde{\Theta})$, and will hence cancel out when we consider $\frac{1}{\rho}\nabla_{\tilde{\boldsymbol{\theta}}_t} h(\Theta) - \frac{1}{\rho}\nabla_{\tilde{\boldsymbol{\theta}}_t} h(\tilde{\Theta})$. Hence, we will not consider these terms in the subsequent analysis.

Ignoring the constant terms, the overall gradient can be written in vectorized form as

$$\frac{1}{\rho}\operatorname{vec}(\nabla h(\Theta)) = \left(1+\frac{1}{T-1}\right)\mathbb{L}\operatorname{vec}(\Theta), \tag{30}$$

where $\operatorname{vec}(\nabla h(\Theta)), \operatorname{vec}(\Theta) \in \mathbb{R}^{p\times T}$ are vectorized versions of the $p \times T$ matrices $\nabla h(\Theta)$ and Θ; and $\mathbb{L}$ is given by

$$\mathbb{L} = \begin{bmatrix} \mathbb{I} & -\frac{1}{T-1}\mathbb{I} & -\frac{1}{T-1}\mathbb{I} & \cdots & -\frac{1}{T-1}\mathbb{I} \\ -\frac{1}{T-1}\mathbb{I} & \mathbb{I} & -\frac{1}{T-1}\mathbb{I} & \cdots & -\frac{1}{T-1}\mathbb{I} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{T-1}\mathbb{I} & -\frac{1}{T-1}\mathbb{I} & -\frac{1}{T-1}\mathbb{I} & \cdots & \mathbb{I} \end{bmatrix}. \tag{31}$$

We note that $\mathbb{L}$ is the (block) graph Laplacian of a complete graph with T vertices (Merris 1994; Chung 1997). The eigenvalues of a complete graph with T vertices are 0 and $\frac{T}{T-1}$ with a multiplicity of $(T-1)$ (Merris 1994).

Then,

$$\frac{1}{\rho}(\operatorname{vec}(\nabla h(\Theta)) - \operatorname{vec}(\nabla h(\tilde{\Theta}))) = \left(1+\frac{1}{T-1}\right)\mathbb{L}(\operatorname{vec}(\Theta) - \operatorname{vec}(\tilde{\Theta})).$$

Denoting $\mathbf{z} = \operatorname{vec}(\Theta) - \operatorname{vec}(\tilde{\Theta})$ for convenience, we have

$$\begin{aligned}\frac{1}{\rho}\|\operatorname{vec}(\nabla h(\Theta)) - \operatorname{vec}(\nabla h(\tilde{\Theta}))\|_2 &\leq \left(1+\frac{1}{T-1}\right)\|\mathbb{L}\mathbf{z}\|_2 \\ &\leq \left(1+\frac{1}{T-1}\right)\frac{T}{T-1}\|\mathbf{z}\|_2 \\ &\leq \left(\frac{T}{T-1}\right)^2\|\mathbf{z}\|_2.\end{aligned}$$

For the longitudinal AD datasets we use, $T \geq 3$, which makes $(\frac{T}{T-1})^2 \leq 3$. Thus, the Lipschitz constant ν for the function $h(\Theta)$ in our setting satisfies $\nu \leq 3\rho$.

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

4 EXPERIMENTAL RESULTS

In this section, we present experimental analysis to demonstrate the effectiveness of the proposed framework on characterizing AD progression using a dataset from the ADNI (Weiner et al. 2010). ADNI[2] is a multi-site study that aimed to improve clinical trials for the prevention and treatment

[2]http://adni.loni.usc.edu/, https://en.wikipedia.org/wiki/Alzheimer's_Disease_Neuroimaging_Initiative.

of AD. ADNI started in 2004, currently includes researchers from 63 research centers in the United States and Canada, and has resulted in inumerous scientific publications using the ADNI data. ADNI has been facilitating the scientific evaluation of neuroimaging data, including MRI, PET, along with other biomarkers, and clinical and neuropsychological assessments for predicting the onset and progression of MCI and AD. The study gathered and analyzed thousands of brain scans, genetic profiles, and biomarkers in blood and cerebrospinal fluid that are used to measure the progress of disease or the effects of treatment.

ADNI is the result of efforts of many researchers from a broad range of academic institutions and private corporations, which was designed to find more sensitive and accurate methods to detect AD at earlier stages and mark its progress through biomarkers. The initial goal of ADNI was to recruit 800 subjects, ages 55–90, including 200 normal controls, 400 individuals with MCI, and 200 subjects with mild AD at approximately 50 sites in the United States and Canada for longitudinal follow up. ADNI also aims to accurately track progression of the disease and devise tests to measure the effectiveness of potential interventions. Currently, the study involves over 1,000 participants, including people without memory problems, those with MCI, and patients with diagnosed AD. Early diagnosis of AD is key to the development, assessment, and monitoring of new treatments for AD. Approaches to characterize AD progression will help researchers and clinicians to 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.

4.1 Experimental Setting

The ADNI project is a longitudinal study, where selected subjects are categorized into three BL diagnostic groups: CN, MCI, and AD, repeatedly over a 6-month or 1-year interval. The date when the subjects are scheduled to perform the screening becomes BL after approval and the time point for the follow-up visits is denoted by the duration starting from the BL. We use the notation Month 6 (M6) to denote the time point half year after the first visit. Currently, ADNI has up to Month 48 follow-up data available for some patients. However, many patients drop out from the study for many reasons.

In this work, we conduct empirical evaluation for the proposed methods on MRI data. 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/). For each image, 71 cortical regions and 44 subcortical regions are generated after this pre-processing. 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 (SV) was calculated as feature. This yielded a total of $p = 319$ MRI features (including 275 cortical and 44 subcortical features) extracted from cortical and subcortical ROIs (see Tables 1 and 2 for details). In addition to the features corresponding to these cortical and sub-cortical regions, the SA of the left and the right hemispheres, and the total intracranial volume (ICV) were also included. Details of the analysis procedure are available at: http://adni.loni.ucla.edu/research/mri-post-processing/.

In this work, we remove features with more than 10% missing entries (for all patients and all time points), exclude patients without BL MRI records and complete the missing entries using the average value. This yields a total of $n = 788$ subjects (173 AD, 390 MCI, and 225 CN) for BL and for the M6, M12, M24, M36, and M48 time points the sample size are 718 (155 AD, 352 MCI, and 211 CN), 662 (134 AD, 330 MCI, and 198 CN), 532 (101 AD, 254 MCI, and 177 CN), 345 (1 AD, 189

Table 1. Cortical Features from the Following 71 (=35 × 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 × 2 × 4 + 1 × 2 × 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).

MCI, and 155 CN), 91 (0 AD, 42 MCI, and 49 CN), respectively. The amount of instances of each task is different in Table 3 since the datasets decrease in size due to the drop out of some patients for various reasons.

Cognitive scores: For predictive modeling, five sets of cognitive scores (Wan et al. 2014; Wang et al. 2012) are examined: ADAS, MMSE, Rey Auditory Verbal Learning Test (RAVLT), Cate-

Table 2. Subcortical Features from the Following 44 (=16 × 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).

Table 3. Description of the Cognitive Scores Considered in the Experiments

Time point	Category			Total
	CN	MCI	AD	
Baseline	225	390	173	788
Month 6	211	352	155	718
Month 12	198	330	134	662
Month 24	177	254	101	532
Month 36	155	189	1	345
Month 48	49	42	0	91

Table 4. Description of the Cognitive Scores Considered in the Experiments

Score name		Description
ADAS		Alzheimer's Disease Assessment Scale
MMSE		Mini-Mental State Exam
RAVLT	TOTAL	Total score of the first 5 learning trials
	TOT6	Trial 6 total number of words recalled
	T30	30 minute delay total number of words recalled
	RECOG	30 minute delay recognition
FLU	ANIM	Animal total score
	VEG	Vegetable total score
TRAILS	A	Trail making test A score
	B	Trail making test B score

gory Fluency (FLU), and Trail Making Test (TRAILS). 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. FLU is a measure of semantic memory (verbal fluency and language). The subject is asked to name different exemplars from a given semantic category. TRAILS is a test of processing speed and executive function, consists of two parts in which the subject is instructed to connect a set of 25 dots as fast as possible while still maintaining accuracy. Certain scores have different variants, yielding a total of 10 scores, and these are listed in Table 4.

In our setting, each of the 10 cognitive scores correspond to one MTL problem, where the different time steps are considered as distinct tasks. Thus, the MTL models, including FL-SGL focus on modeling the progression of these scores. Results will be reported on each of these 10 cognitive scores separately.

Evaluation metrics: 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 single time point. CC is used to calculate the value of R in Figures 5–7. Moreover, for aggregated performance over all time points, the normalized mean squared error (nMSE) (Argyriou et al. 2008; Zhou et al. 2013) and weighted R-value (wR) (Stonningtona et al. 2010) are used. The nMSE and wR are defined as follows:

$$\text{nMSE}(Y, \hat{Y}) = \frac{\sum_{h=1}^{k} \frac{\|Y_h - \hat{Y}_h\|_2^2}{\sigma(Y_h)}}{\sum_{h=1}^{k} n_h}, \tag{32}$$

$$\text{wR}(Y, \hat{Y}) = \frac{\sum_{h=1}^{k} \text{Corr}(Y_h, \hat{Y}_h) n_h}{\sum_{h=1}^{k} n_h}, \tag{33}$$

where Y and $\hat{Y}$ are the ground truth cognitive scores and the predicted cognitive scores, respectively. A smaller (higher) value of nMSE and rMSE (CC and wR) represents better regression performance. The average (avg) and standard deviation (std) of performance measures across 20

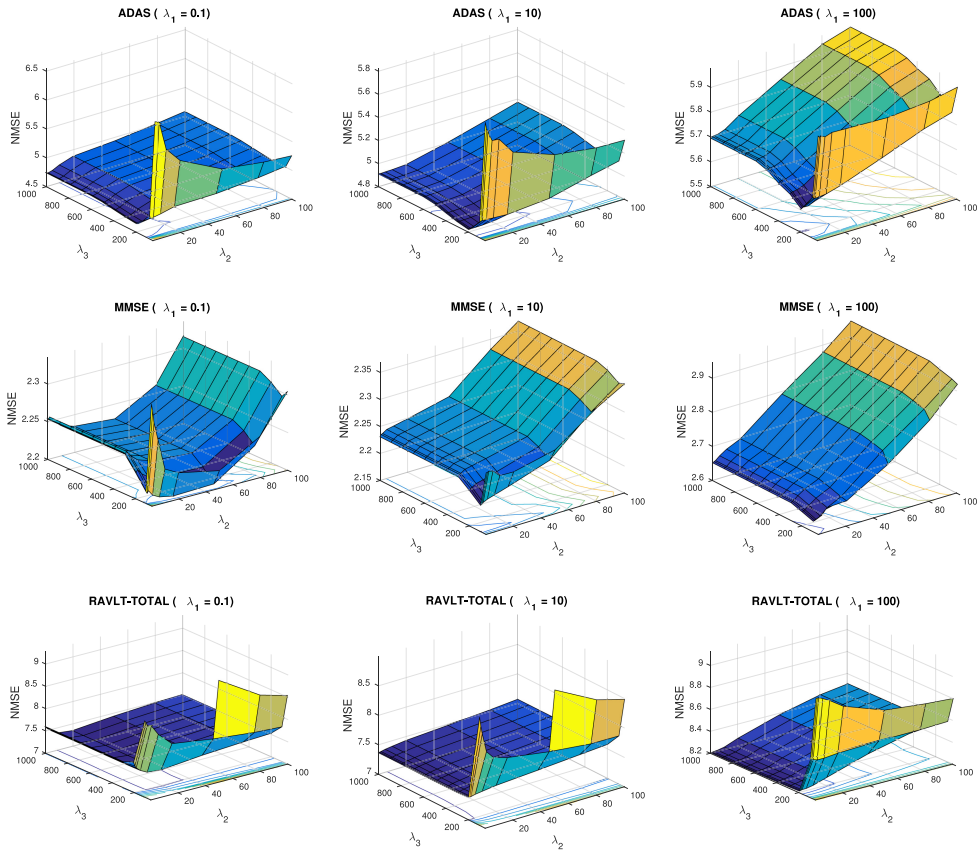


Fig. 3. Hyper-parameter sensitivity analysis: hyper-parameter λ_3 associated with the FL-SGL temporal smooth term plays an important role in the prediction performance and should not be neglected. Larger values for λ_2 (associated with group lasso penalty) has a clear tendency to worsen the results, particularly for larger values of λ_1.

runs on different splits of data are shown as avg ± std for each experiment. A Student's t-test at a significance level of 0.05 is performed to determine whether the performances difference are significant.

Experimental methodology: We randomly split the data into training and testing sets using a ratio 9:1 and repeat 20 trials, i.e., we build models on 90% of the data (train-set) and evaluate these models on the remaining 10% of the data (test-set). In each trial, a five-fold cross validation on the train-set is done to select the regularization parameters (hyper-parameters) $(\lambda_1, \lambda_2, \lambda_3)$, and the estimated model using these regularization parameters are used to predict on the test set. For the cross-validation, for a fixed set of hyper-parameters, four folds are used for training, one fold for evaluation using nMSE. For hyper-parameter selection, we consider a grid of regularization parameter values, where each regularization parameter is varied from 10^{-1} to 10^3 in log scale. The data was z-scored before applying regression methods.

Hyper-parameter sensitivity: To assess the sensitivity of the three hyper-parameters in the FL-SGL formulation (Equation (4)), we explored the three-dimension hyper-parameters space and plot

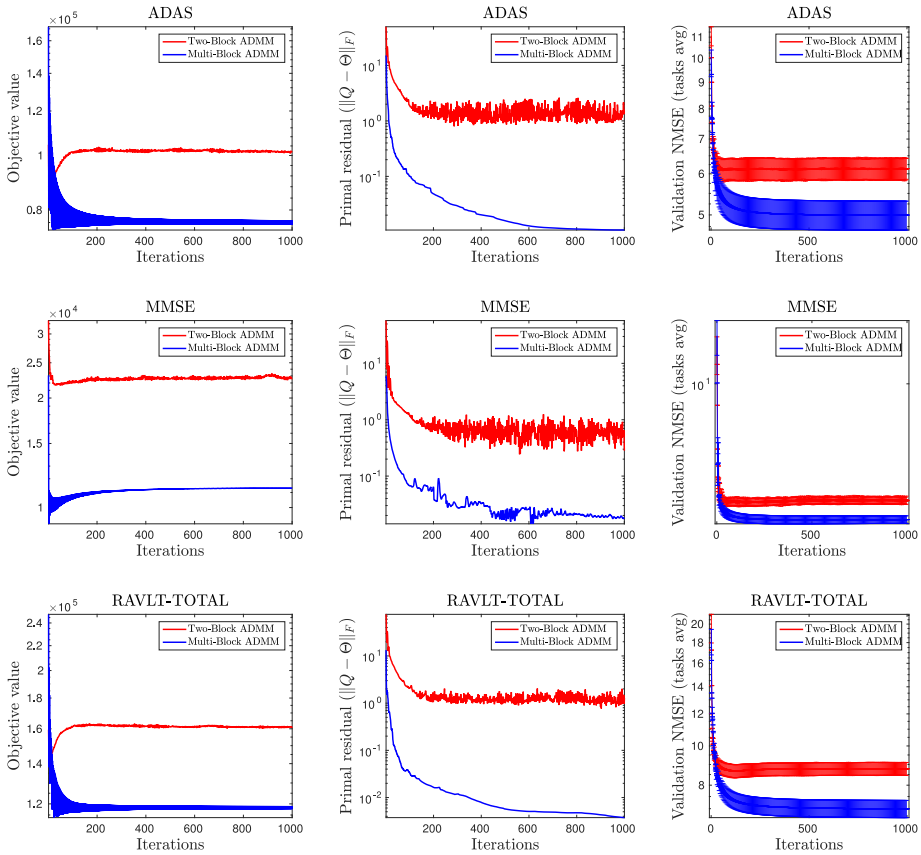


Fig. 4. Convergence of Two-Block and Multi-Block ADMM for FL-SGL formulation. The Multi-block version has a faster convergence rate and leads to a lower validation prediction error.

the NMSE metric for each combination of values. The sensitivity study is important to investigate the influence of each term in the FL-SGL formulation, and provide a guidance on how to properly set the hyper-parameters. The hyper-parameter space is defined as $\lambda_1 \in [0.1, 100]$, $\lambda_2 \in [0.1, 100]$, and $\lambda_3 \in [50, 1000]$. $\rho = 10$ was used in this experiment. The NMSE presented was computed in the test set. Due to space limitations, Figure 3 only shows plots for ADAS, MMSE, and RAVLT.TOTAL cognitive scores. It is possible to observe that for all cognitive scores, smaller values for λ_3 led to poor regression performance, indicating that the temporal smooth penalization term plays an important role in the prediction and should not be neglected. Larger values for λ_2 (associated with group lasso penalty) has a clear tendency to worsen the results, particularly for larger values of λ_1. As λ_1 increases, we enforce more sparsity on the $\boldsymbol{\theta}$ parameters, hence breaking the group structure present in the data.

4.2 Optimization: Multi-Block and Two-Block ADMM

In Section 3, we discussed two methods for solving the optimization problem associated with the FL-SGL formulation, namely Linearized Two-Block ADMM and Linearized Multi-Block ADMM. In this section, we empirically investigate and compare their performances based on the ADNI dataset described in Section 4.1.

Due to space limitations, we only show the convergence on three scores in Figure 4, namely ADAS, MMSE, and RAVLT.TOTAL. The same behavior was observed for the other scores. The (primal) objective value plots on the left column show that the Multi-block ADMM converged to solutions with lower cost function in the training phase. It is worth mentioning that the cost function does not monotonically decrease, because in the beginning of the optimization process the solution is not feasible, and the objective function curve may change as the solution becomes feasible. In ADMMs, the feasibility is reached during the optimization and it can be assessed by looking at the primal residual, which is illustrated in the middle column plots. To compute the validation curves in the right most column, 20% of the training data was used as a validation set to measure the prediction power of intermediate solutions of both ADMMs. We carried out 30 independent runs with different sets of validation data splits, and the curves show the mean and standard deviation NMSE over the multiple runs. For these experiments we used $\rho = 10$, which showed to lead to more stable behavior of the methods. Notably, Multi-block ADMM achieved solutions with higher generalization capacity. Guided by these results, we chose the Multi-Block ADMM as our optimization algorithm for the remaining experiments.

4.3 Prediction Performance Based on MRI Features

Prediction performance results of 10 cognitive scores are reported in Table 5. We compare the performance of FL-SGL with different regression methods, including ridge regression (Draper and Smith 2014), lasso (Liu and Ye 2009), which are applied independently to each time point, and cFSGL (Zhou et al. 2013), which is one of the state-of-the-art methods for characterizing longitudinal AD progression. Recall that each experiment focuses on a specific cognitive score, with different time points serving as different tasks for the MTL formulations. Since there are a total of 10 cognitive scores, we run experiments and report results individually for each score.

The average and standard deviation of performance measures are calculated by 20 iterations of trials on different splits of data, and are shown in Table 5. It is worth noting that we use the same training and test data across the experiments for all the methods for fair comparison.

The results show that multi-task temporal smoothness models (cFSGL, F-SGL, FL-SGL1, and FL-SGL2) provide more accurate predictions of the cognitive scores when compared to single-task learning models (ridge and lasso) in terms of both nMSE and wR over all scores. cFSGL outperforms F-SGL in terms of nMSE, wR and the rMSE in the most time points. As F-SGL and cFSGL formulations are equivalent, it suggests that the different results due to the different optimization methods. The results indicate the AGM is more effective than ADMM for optimizing the fused SGL-based formulation. However, FL-SGL1 and FL-SGL2 outperform cFSGL, which demonstrates the advantage of the proposed transform matrix taking into account all the time points. Between the two proposed methods, FL-SGL2 outperforms FL-SGL1 in six tasks (RAVLT.TOTAL, RAVLT.TOT6, RAVLT.RECOG, FLU.ANIM, FLU.VEG, and TRAILS.A) in terms of nMSE and seven tasks (ADAS, MMSE, RAVLT.TOT6, RAVLT. RECOG, FLU.ANIM, and FLU.VEG) in terms of wR. The statistical hypothesis test reveals that FL-SGL1 and FL-SGL2 are significantly better than the contenders for most of the scores.

We show the scatter plots of actual values versus predicted values on testing data. Due to lack of space, we only listed three scatter plots, including ADAS, MMSE, and RAVLT.TOTAL in Figures 5–7, respectively. Since the sample size at the M48 time point is small, we only show the scatter plots for the first four time points. The value of R in the figures is calculated by CC. From the scatter plots, we can see that the predicted values and actual values scores have similar high correlation for these three tasks. The scatter plots show that the prediction performance for ADAS is better

Table 5. Prediction performance results of 10 cognitive scores of six time points based on MRI features

	Ridge	Lasso	cFSGL	F-SGL	FL-SGL1	FL-SGL2
Score: ADAS						
nMSE	10.01±0.794$^{\dagger *}$	6.613±0.474$^{\dagger *}$	5.030±0.310$^{\dagger *}$	5.642±0.380$^{\dagger *}$	**4.969±0.318**	4.975±0.329
wR	0.581±0.033$^{\dagger *}$	0.628±0.024$^{\dagger *}$	**0.750±0.016**	0.695±0.019$^{\dagger *}$	0.749±0.017	0.749±0.018
BL rMSE	7.831±0.719	6.922±0.630	6.278±0.462	6.143±0.609	**6.033±0.496**	6.182±0.502
M6 rMSE	8.700±0.922	7.645±0.749	6.532±0.648	6.913±0.727	6.576±0.705	**6.520±0.686**
M12 rMSE	9.771±0.810	8.636±0.816	**7.240±0.630**	7.992±0.747	7.293±0.638	7.243±0.637
M24 rMSE	11.81±1.263	10.29±0.866	9.115±1.034	9.685±0.817	**8.950±0.958**	8.992±0.976
M36 rMSE	12.82±1.750	9.496±1.291	**8.183±0.854**	9.036±1.180	8.331±0.775	8.204±0.855
M48 rMSE	20.07±3.749	9.125±2.088	**8.104±1.296**	9.142±1.810	8.105±1.352	8.164±1.338
Score: MMSE						
nMSE	13.90±12.01$^{\dagger *}$	2.582±0.251$^{\dagger *}$	2.208±0.175$^{\dagger *}$	2.409±0.184$^{\dagger *}$	**2.136±0.165**	2.152±0.160
wR	0.424±0.038$^{\dagger *}$	0.575±0.033$^{\dagger *}$	0.652±0.031†	0.604±0.028$^{\dagger *}$	0.654±0.032	**0.651±0.031**
BL rMSE	2.716±0.283	2.205±0.188	**2.198±0.243**	2.133±0.207	2.207±0.226	2.223±0.205
M6 rMSE	3.410±0.223	2.875±0.272	**2.605±0.257**	2.733±0.255	2.619±0.259	2.624±0.259
M12 rMSE	3.888±0.363	3.191±0.328	2.852±0.273	3.036±0.297	**2.825±0.253**	2.833±0.263
M24 rMSE	4.951±0.423	3.886±0.483	3.540±0.468	3.840±0.499	**3.468±0.456**	3.491±0.451
M36 rMSE	5.901±1.202	3.299±0.660	3.065±0.525	3.295±0.705	3.045±0.497	**3.029±0.526**
M48 rMSE	29.94±1.466	4.334±1.200	4.240±1.399	4.533±1.586	**3.226±0.847**	3.297±0.864
Score: RAVLT.TOTAL						
nMSE	17.69±1.207$^{\dagger *}$	9.679±0.501$^{\dagger *}$	7.025±0.419$^{\dagger *}$	8.246±0.483$^{\dagger *}$	6.845±0.467	**6.842±0.454**
wR	0.406±0.047$^{\dagger *}$	0.506±0.031$^{\dagger *}$	0.675±0.023$^{\dagger *}$	0.588±0.028$^{\dagger *}$	**0.685±0.025**	0.684±0.024
BL rMSE	11.37±0.848	9.826±0.753	8.994±0.730	9.075±0.659	**8.860±0.815**	8.969±0.763
M6 rMSE	11.62±0.844	10.08±0.804	8.720±0.848	9.363±0.802	8.745±0.798	**8.694±0.791**
M12 rMSE	12.90±1.221	11.57±1.011	9.633±0.927	10.43±1.002	9.463±0.866	**9.415±0.873**
M24 rMSE	14.88±1.639	12.27±1.049	9.979±0.997	11.24±0.944	9.793±0.801	**9.776±0.816**
M36 rMSE	16.66±1.969	11.71±1.415	9.394±1.187	10.95±1.324	9.259±1.166	**9.162±1.181**
M48 rMSE	41.21±3.442	11.88±1.801	9.611±1.990	11.81±2.050	**9.069±2.318**	9.277±2.340
Score: RAVLT.TOT6						
nMSE	3.934±0.275$^{\dagger *}$	2.881±0.131$^{\dagger *}$	2.375±0.172$^{\dagger *}$	2.664±0.126$^{\dagger *}$	2.297±0.161	**2.280±0.170**
wR	0.459±0.044$^{\dagger *}$	0.542±0.031$^{\dagger *}$	0.655±0.031$^{\dagger *}$	0.583±0.029$^{\dagger *}$	0.666±0.031	**0.669±0.031**
BL rMSE	3.624±0.274	3.252±0.197	3.099±0.240	3.177±0.190	**3.056±0.226**	3.071±0.230
M6 rMSE	3.436±0.321	3.160±0.258	2.865±0.215	3.056±0.246	2.813±0.189	**2.800±0.176**
M12 rMSE	3.825±0.318	3.438±0.292	3.121±0.260	3.322±0.263	3.085±0.237	**3.053±0.250**
M24 rMSE	4.111±0.375	3.601±0.328	3.197±0.330	3.429±0.293	3.122±0.334	**3.115±0.323**
M36 rMSE	4.245±0.745	3.496±0.372	3.011±0.328	3.359±0.334	2.967±0.412	**2.916±0.421**
M48 rMSE	7.538±1.252	4.212±0.815	3.499±0.766	3.662±0.581	**3.368±0.780**	3.402±0.828
Score: RAVLT.T30						
nMSE	3.869±0.240$^{\dagger *}$	3.012±0.131$^{\dagger *}$	2.392±0.179$^{\dagger *}$	2.770±0.140$^{\dagger *}$	**2.369±0.169**	2.370±0.163
wR	0.456±0.043$^{\dagger *}$	0.539±0.032$^{\dagger *}$	0.667±0.033	0.586±0.034$^{\dagger *}$	**0.671±0.030**	0.670±0.029
BL rMSE	3.831±0.277	3.443±0.224	3.273±0.266	3.356±0.222	**3.216±0.258**	3.257±0.253
M6 rMSE	3.454±0.298	3.176±0.309	**2.881±0.192**	3.067±0.281	2.897±0.178	2.885±0.187

(Continued)

Table 5. Continued

	Ridge	Lasso	cFSGL	F-SGL	FL-SGL1	FL-SGL2
M12 rMSE	4.026±0.406	3.753±0.339	3.264±0.325	3.597±0.343	3.281±0.328	**3.245±0.318**
M24 rMSE	4.126±0.439	3.707±0.319	3.191±0.303	3.543±0.332	3.183±0.316	**3.177±0.308**
M36 rMSE	4.097±0.778	3.428±0.399	2.935±0.381	3.313±0.378	2.882±0.429	**2.860±0.414**
M48 rMSE	7.029±1.602	4.669±0.966	3.603±0.767	3.950±0.640	**3.598±0.756**	3.693±0.786
Score: RAVLT.RECOG						
nMSE	6.279±0.629†*	3.350±0.181†*	2.896±0.223†*	3.147±0.178†*	2.850±0.215	**2.818±0.215**
wR	0.341±0.037†*	0.470±0.033†*	0.578±0.036†*	0.512±0.031†*	0.588±0.033	**0.595±0.035**
BL rMSE	4.259±0.278	3.554±0.231	3.447±0.250	3.479±0.230	**3.394±0.277**	3.434±0.282
M6 rMSE	4.491±0.339	3.833±0.329	3.576±0.382	3.750±0.345	3.597±0.399	**3.532±0.392**
M12 rMSE	4.776±0.364	3.809±0.220	3.534±0.277	3.723±0.250	3.555±0.278	**3.503±0.272**
M24 rMSE	4.983±0.465	3.771±0.277	3.404±0.386	3.680±0.294	3.316±0.391	**3.289±0.366**
M36 rMSE	5.340±0.476	3.629±0.281	3.381±0.378	3.468±0.320	3.353±0.395	**3.333±0.415**
M48 rMSE	12.59±0.901	4.856±0.703	3.560±0.692	4.096±1.103	**3.300±0.561**	3.343±0.605
Score: FLU.ANIM						
nMSE	9.497±0.817†*	5.080±0.325†*	3.966±0.395†*	4.640±0.283†*	3.901±0.360	**3.896±0.361**
wR	0.300±0.061†*	0.402±0.048†*	0.592±0.046†*	0.479±0.041†*	0.602±0.040	**0.602±0.040**
BL rMSE	6.369±0.509	5.366±0.413	4.984±0.360	5.105±0.410	**4.858±0.370**	4.956±0.369
M6 rMSE	6.017±0.507	5.178±0.559	4.633±0.419	4.946±0.497	4.649±0.386	**4.602±0.381**
M12 rMSE	6.678±0.905	5.783±0.929	4.902±0.998	5.387±0.860	4.864±0.941	**4.816±0.962**
M24 rMSE	7.278±0.570	5.744±0.547	5.068±0.558	5.588±0.523	5.034±0.551	**5.025±0.545**
M36 rMSE	7.655±1.026	5.439±0.743	4.550±0.623	5.309±0.728	4.558±0.627	**4.535±0.649**
M48 rMSE	20.42±2.239	6.357±1.301	5.131±1.082	6.300±1.307	5.128±1.013	**5.014±1.019**
Score: FLU.VEG						
nMSE	6.622±0.447†*	3.477±0.197†*	2.830±0.169†*	3.273±0.206†*	2.800±0.187	**2.797±0.174**
wR	0.382±0.041†*	0.502±0.043†*	0.634±0.030*	0.544±0.035†*	0.637±0.031	**0.638±0.030**
BL rMSE	4.449±0.343	3.643±0.319	3.503±0.306	3.552±0.311	**3.445±0.309**	3.511±0.306
M6 rMSE	4.674±0.447	3.913±0.373	3.517±0.303	3.813±0.359	3.516±0.306	**3.481±0.301**
M12 rMSE	4.802±0.490	3.971±0.260	3.489±0.344	3.824±0.282	3.503±0.357	**3.476±0.350**
M24 rMSE	5.297±0.531	4.308±0.431	3.870±0.365	4.169±0.441	**3.832±0.333**	3.840±0.338
M36 rMSE	6.567±0.661	4.222±0.334	3.762±0.295	4.109±0.348	3.779±0.310	**3.717±0.297**
M48 rMSE	13.89±1.565	5.102±0.880	3.634±0.512	4.865±0.894	**3.513±0.544**	3.534±0.520
Score: TRAILS.A						
nMSE	33.31±3.152†*	23.62±1.922†*	18.99±1.645†*	21.41±1.688†*	18.38±1.696	**18.31±1.612**
wR	0.360±0.056†*	0.391±0.043†*	0.576±0.049†*	0.482±0.046†*	**0.597±0.052**	0.596±0.049
BL rMSE	27.60±3.255	24.98±3.246	22.36±2.611	23.55±2.839	**21.89±2.764**	22.26±2.696
M6 rMSE	28.10±3.267	24.93±3.403	22.76±2.593	23.94±3.277	22.64±2.716	**22.28±2.686**
M12 rMSE	28.56±3.394	25.66±4.284	21.98±2.288	24.26±3.295	21.25±2.013	**21.01±2.042**
M24 rMSE	30.68±5.957	27.91±5.404	23.83±4.680	26.22±4.878	23.29±4.382	**23.24±4.372**
M36 rMSE	31.67±5.915	23.89±5.073	**22.71±3.094**	23.46±4.503	22.81±3.081	22.78±3.121
M48 rMSE	51.79±10.68	27.12±13.39	23.61±7.991	25.46±10.91	**22.75±7.584**	23.31±7.906

(Continued)

Table 5. Continued

	Ridge	Lasso	cFSGL	F-SGL	FL-SGL1	FL-SGL2
Score: TRAILS.B						
nMSE	91.14±7.728$^{\dagger *}$	65.90±5.027$^{\dagger *}$	**55.54±4.238**	59.59±5.114$^{\dagger *}$	55.65±4.021	55.98±4.240
wR	0.404±0.044$^{\dagger *}$	0.451±0.040$^{\dagger *}$	0.568±0.030	0.515±0.037$^{\dagger *}$	**0.570±0.029**	0.567±0.031
BL rMSE	80.03±5.554	71.40±5.356	67.64±5.241	67.63±5.158	**67.28±5.344**	67.78±5.608
M6 rMSE	79.51±9.557	71.42±7.229	**66.07±7.492**	67.58±7.642	66.21±7.353	66.14±7.402
M12 rMSE	77.88±7.025	71.61±5.791	**64.75±5.848**	67.96±5.975	65.51±5.933	65.44±6.014
M24 rMSE	90.88±12.90	78.03±8.892	69.61±8.652	73.56±8.536	69.43±8.431	**69.20±8.455**
M36 rMSE	93.62±21.31	75.01±20.42	69.51±20.56	72.92±21.21	**69.39±19.04**	70.28±18.83
M48 rMSE	131.2±21.75	69.42±19.59	57.10±16.40	68.06±17.11	**56.23±16.46**	56.98±16.88

Note that the best results are boldfaced, superscript symbols † and * indicate that FL-SGL1 and FL-SGL2, respectively, significantly outperformed that method on that score. Student's t-test at a level of 0.05 was used.

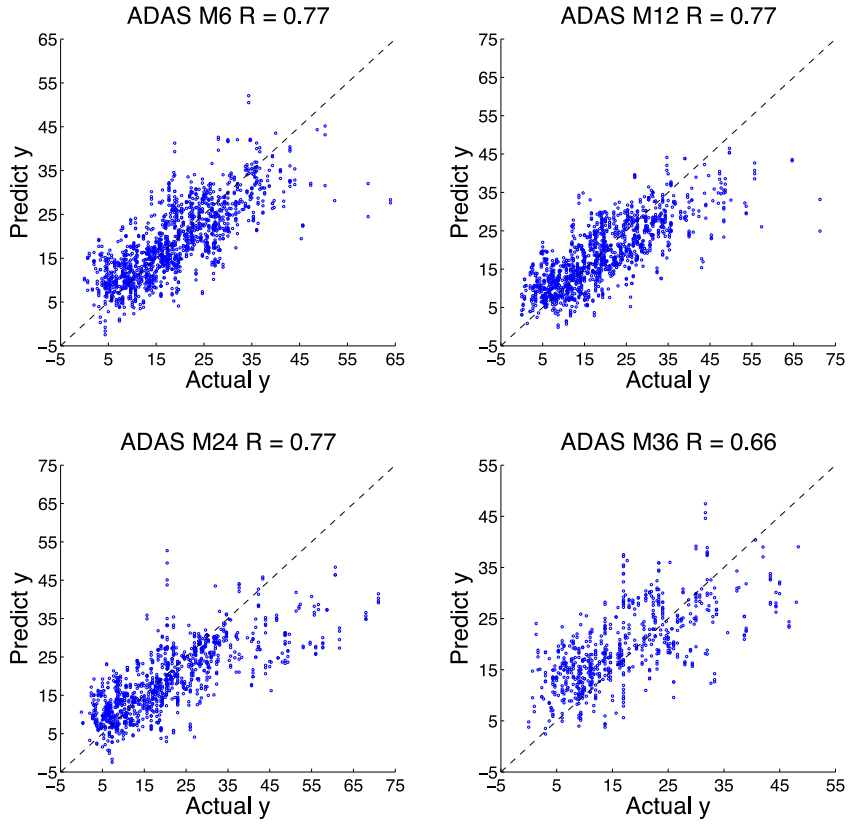


Fig. 5. Scatter plots of actual ADAS versus predicted values on testing data using FL-SGL2 based on MRI features. The value of R is calculated by Correlation Coefficient. Strong correlation is observed for the ADAS score in all time points.

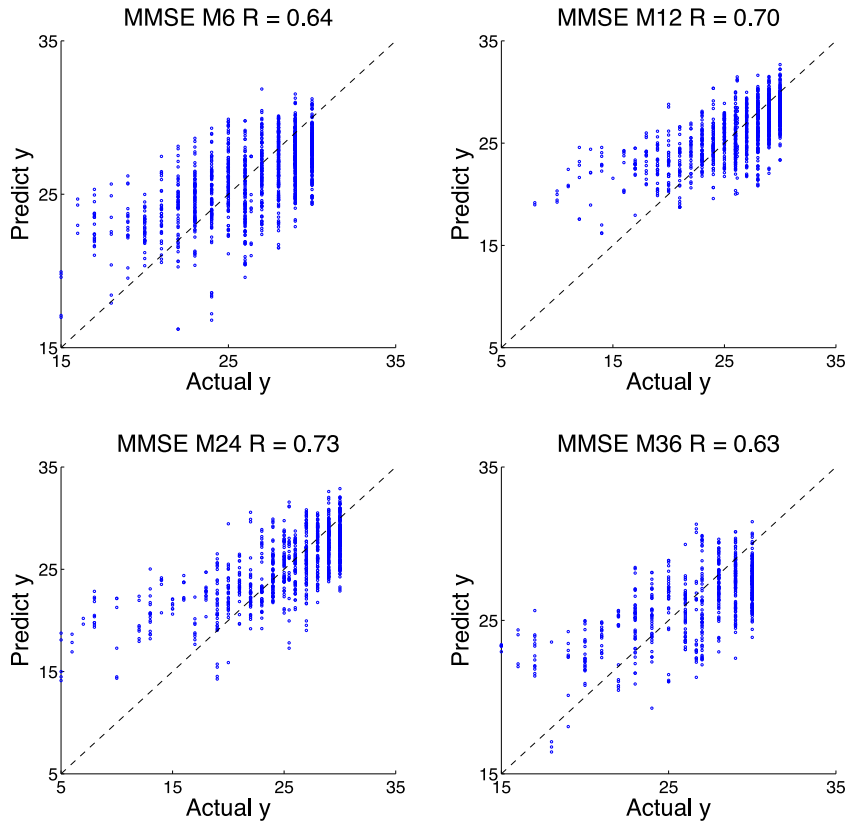


Fig. 6. Scatter plots of actual MMSE versus predicted values on testing data using FL-SGL2 based on MRI features. The value of R is calculated by Correlation Coefficient. Strong correlation is observed for the MMSE score in all time points.

than that of MMSE and RAVLT.TOTAL. In the future, we will add more modalities, such as PET, CSF to improve the performance.

4.4 Identification of Longitudinal MRI Biomarkers

One of the primary goals of our formulation is to identify the temporal imaging markers which are highly correlated to the longitudinal AD progression and are also clinically meaningful. In this subsection, we first identify statistically stable biomarkers (Section 4.4.1) and discuss their clinical relevance based on existing literature (Section 4.4.2).

4.4.1 Stable Longitudinal Biomarkers. We study the temporal imaging markers identified by our method using longitudinal stability selection (Zhou et al. 2013). Broadly speaking, the stability selection procedure consists in running FL-SGL to numerous random subsets of data and computing the frequency with which each feature was selected (corresponding model weight is greater than a threshold value) for each cognitive score and time point across the runs. A feature is claimed to be stable if it was selected in a large portion of the runs (high frequency). For more details, see Meinshausen and Bühlmann (2010) and Zhou et al. (2013). The computed frequency vector is referred here as stability vector.

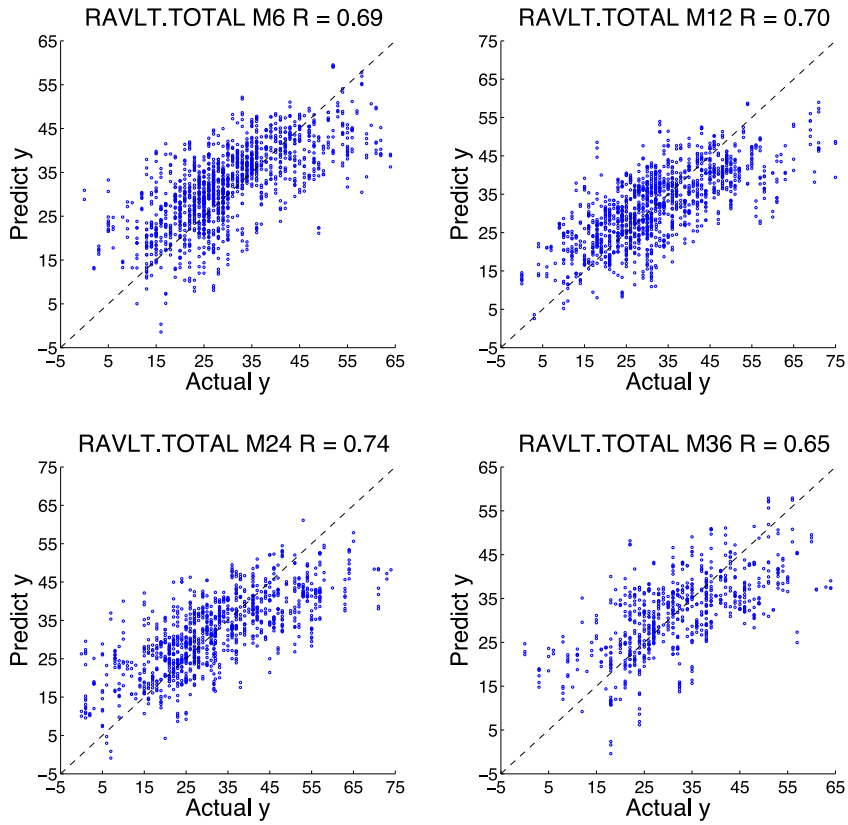


Fig. 7. Scatter plots of actual RAVLT.TOTAL versus predicted values on testing data using FL-SGL2 based on MRI features. The value of R is calculated by Correlation Coefficient. Strong correlation is observed for the RAVLT.TOTAL score in all time points.

Since there are few samples available for the last time point (M48), we only performed longitudinal stability selection for the first four time points. Due to lack of space, we only show three stability vectors with the top 30 stable features for ADAS, MMSE, and RAVLT.TOTAL by obtaining an average stability score for four time points, respectively, in Figure 8. We also listed the top 10 stable features for all the scores in Table 6.

The top 30 stable MRI features for ADAS score are shown in Figure 8(a). We note that most features provide significant information that span across all the time points, which demonstrates that these biomarkers are longitudinally stable due to the advantage of smooth temporal regularization. It is interesting to note that the selected stable features are consistent with respect to the stable score, which indicates that our method has a good performance for later time points where few samples are available. Moreover, it also demonstrates there exists a strong correlation among the multiple tasks of score prediction at multiple time points. SV of left hippocampus, cortical TA of left middle temporal, and CV of left pars opercularis have large longitudinal stability scores.

Figure 8(b) presents the top 30 stable MRI features for MMSE. We observe that the stable biomarkers have different patterns than for ADAS score. Notably, the stability vectors are not consistent across of multiple time points for MMSE as for ADAS, which suggests a weaker

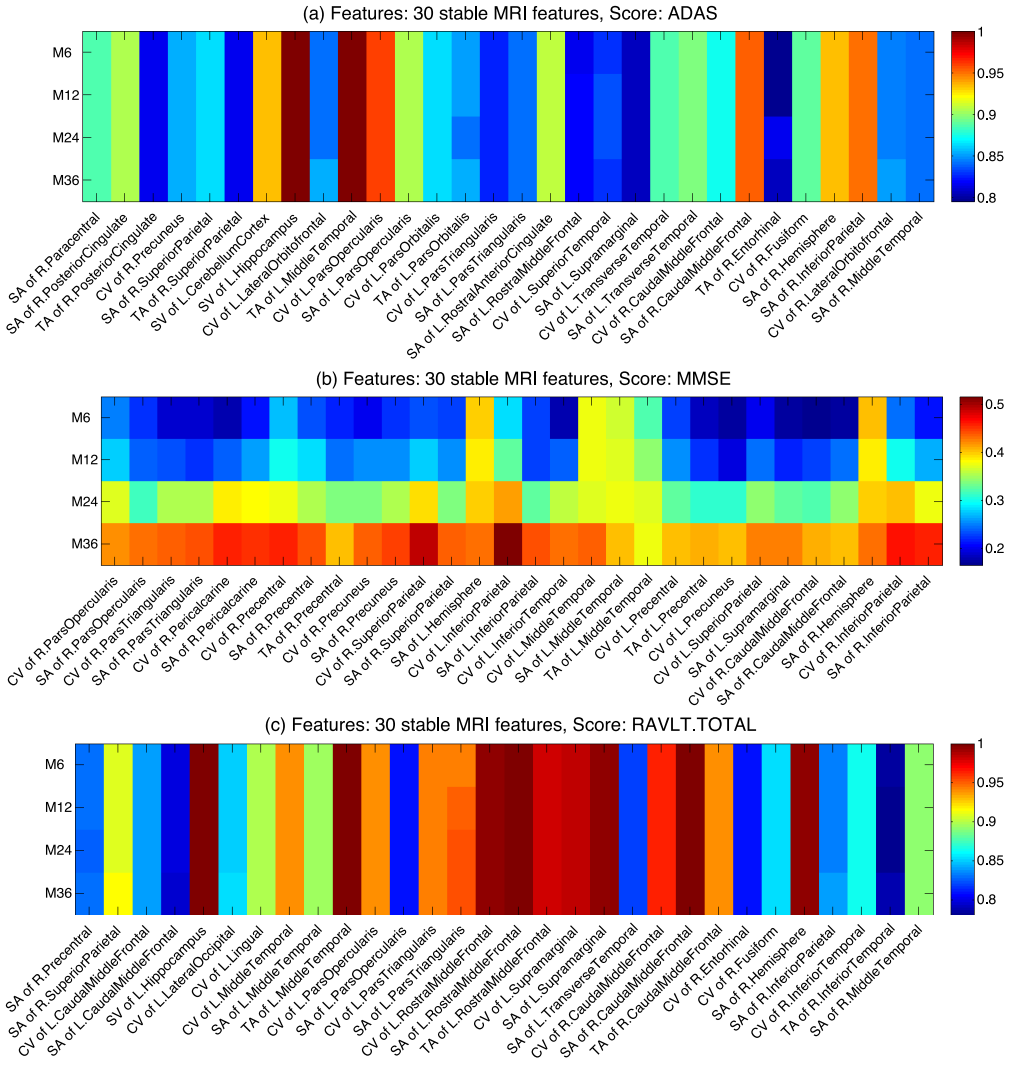


Fig. 8. Stability vectors of stable MRI features generated by FL-SGL2 for ADAS, MMSE, and RAVLT.TOTAL using longitudinal stability selection. The larger the value more stable is the feature.

correlation among multiple time points if compared to ADAS and RAVLT.TOTAl. Additionally, we note that most biomarkers provide significant information for the last stage (M36) and few of them are significant in the first two years, which possibly is the factor by which our method obtained higher performance in later time points. However, SA of left and right hemisphere, and CV and SA of left middle temporal are important biomarkers for all time points.

The stability vectors of the top 30 stable MRI features for RAVLT.TOTAL are shown in Figure 8(c). Similar to ADAS, most features are longitudinally stable, but more features have large longitudinal stability vectors than ADAS, such as SV of left hippocampus, cortical TA of left middle temporal, SA of left rostral middle frontal, right caudal middle frontal and right hemisphere. It can be the reason for the FL-SGL consistent results for all the times points. The identified temporal patterns of MRI biomarkers for these three scores suggest that different time points share

Table 6. The Top 10 Features Identified by Our FL-SGL2 Method for the 10 Prediction Tasks of Cognitive Scores

Score Name	Features
ADAS	SV of L.Hippocampus, TA of L.MiddleTemporal, CV of L.ParsOpercularis, SA of R.CaudalMiddleFrontal, SA of R.InferiorParietal, SV of L.CerebellumCortex, SA of R.Hemisphere, SA of L.RostralAnteriorCingulate, SA of R.PosteriorCingulate, SA of L.ParsOpercularis
MMSE	SA of R.Hemisphere, SA of L.Hemisphere, CV of L.MiddleTemporal, CV of L.InferiorParietal, SA of L.MiddleTemporal, CV of R.InferiorParietal, TA of L.MiddleTemporal, CV of R.Precentral, CV of R.SuperiorParietal, CV of R.ParsOpercularis
RAVLT.TOTAL	SV of L.Hippocampus, TA of L.MiddleTemporal, SA of L.RostralMiddleFrontal, SA of R.CaudalMiddleFrontal, CV of L.RostralMiddleFrontal, SA of L.Supramarginal, SA of R.Hemisphere, CV of L.Supramarginal, TA of L.RostralMiddleFrontal, CV of R.CaudalMiddleFrontal
RAVLT.TOT6	SV of L.Hippocampus, TA of L.MiddleTemporal, CV of L.RostralMiddleFrontal, SA of L.RostralMiddleFrontal, CV of R.CaudalMiddleFrontal, SA of R.CaudalMiddleFrontal, TA of R.CaudalMiddleFrontal, TA of R.Lingual, CV of L.ParsOpercularis, TA of L.RostralMiddleFrontal
RAVLT.T30	CV of R.RostralAnteriorCingulate, SV of L.Hippocampus, CV of L.RostralMiddleFrontal, SA of L.RostralMiddleFrontal, SV of R.InferiorLateralVentricle, TA of R.Lingual, TA of L.RostralMiddleFrontal, CV of R.Lingual, SA of R.CaudalMiddleFrontal, TA of L.MiddleTemporal
RAVLT.RECOG	SV of L.Hippocampus, CV of L.ParsOpercularis, SA of L.RostralMiddleFrontal, SA of L.ParsOpercularis, SA of L.ParsTriangularis, SA of R.CaudalMiddleFrontal, CV of L.ParsTriangularis, CV of R.CaudalMiddleFrontal, CV of L.RostralMiddleFrontal, TA of L.RostralMiddleFrontal
FLU.ANIM	TA of R.SuperiorParietal, SA of R.MiddleTemporal, SA of L.Supramarginal, SA of R.SuperiorParietal, CV of R.SuperiorFrontal, TA of L.MiddleTemporal, CV of R.MiddleTemporal, TA of R.MiddleTemporal, CV of L.Supramarginal, SA of L.MedialOrbitofrontal
FLU.VEG	CV of L.SuperiorFrontal, CV of L.ParsOpercularis, CV of L.CaudalMiddleFrontal, SV of L.Hippocampus, CV of R.Precuneus, SA of L.ParsOpercularis, SA of L.ParsTriangularis, SA of R.MiddleTemporal, TA of R.SuperiorParietal, SA of L.SuperiorFrontal
TRAILS.A	SA of L.MedialOrbitofrontal, CV of R.FrontalPole, SA of R.FrontalPole, CV of L.MedialOrbitofrontal, SA of R.CaudalMiddleFrontal, CV of R.CaudalMiddleFrontal, TA of L.SuperiorFrontal, CV of R.SuperiorTemporal, SA of R.SuperiorTemporal, CV of L.ParsOrbitalis
TRAILS.B	TA of L.MiddleTemporal, TA of L.Insula, TA of R.Entorhinal, TA of L.InferiorParietal, SV of R.LateralVentricle, SA of L.SuperiorFrontal, TA of R.InferiorParietal, TA of R.Lingual, SV of FourthVentricle, TS of L.ParsTriangularis

Many features have been simultaneously identified as stable features for multiple cognitive measures.

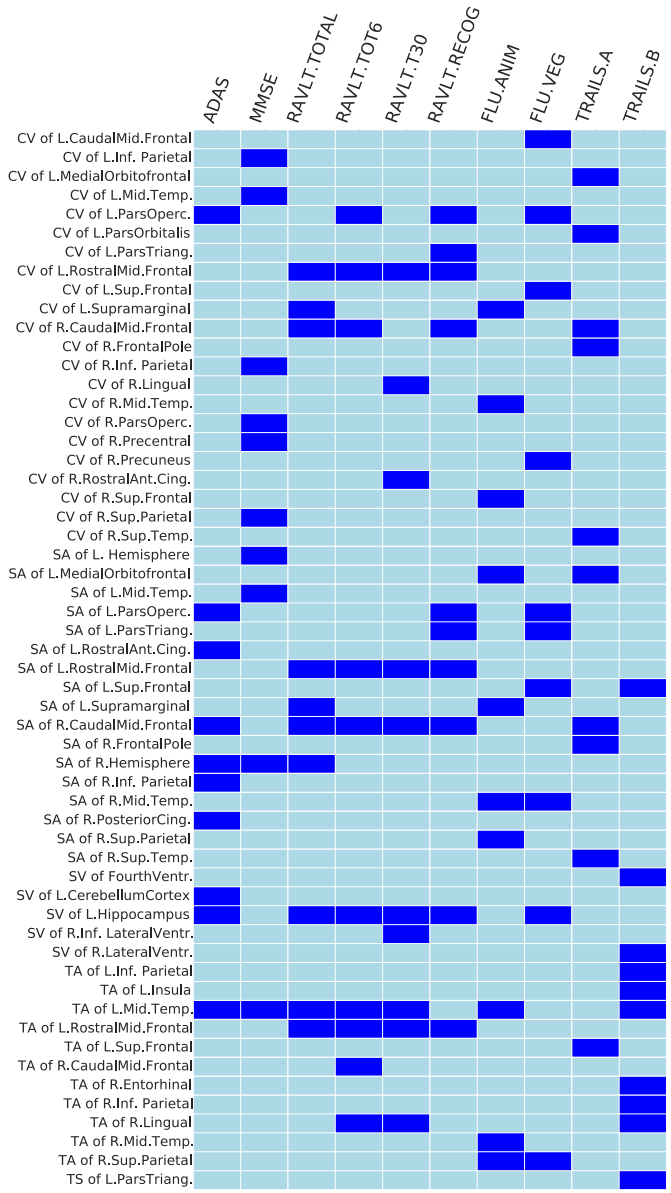


Fig. 9. An alternative view from Table 6. The top 10 features identified by our FL-SGL2 method for the 10 prediction tasks of cognitive scores. Several features have been identified as an important marker by different scores.

similar features, which demonstrates that these biomarkers are longitudinally important due to the advantage of smooth temporal regularization.

The top 10 stable features selected by ranking the average of the four stability vectors from time points M6, M12, M24, and M36 are listed in Table 6. And the alternative view from Table 6 is shown in Figure 9. The total number of features for the 10 scores is 56. This is due to the fact that some features were identified as stable for many scores, which suggests that different scores share similar features.

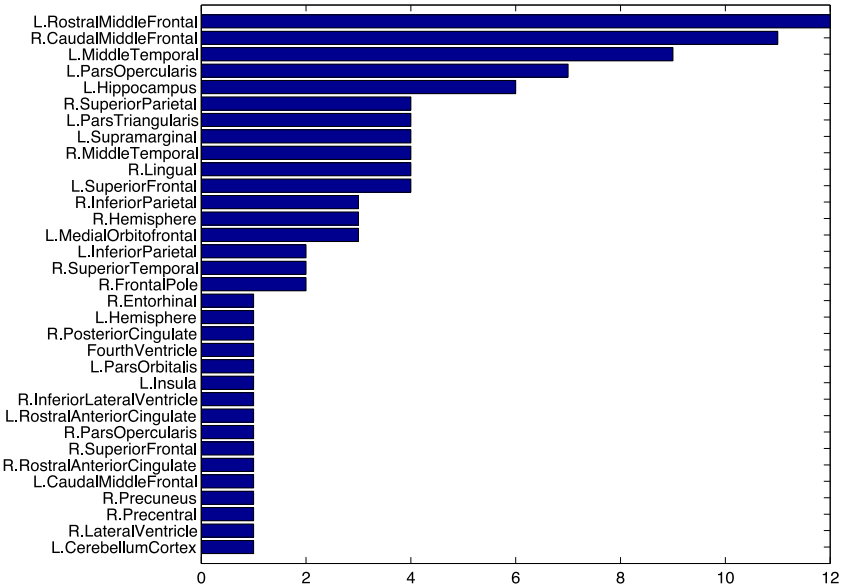


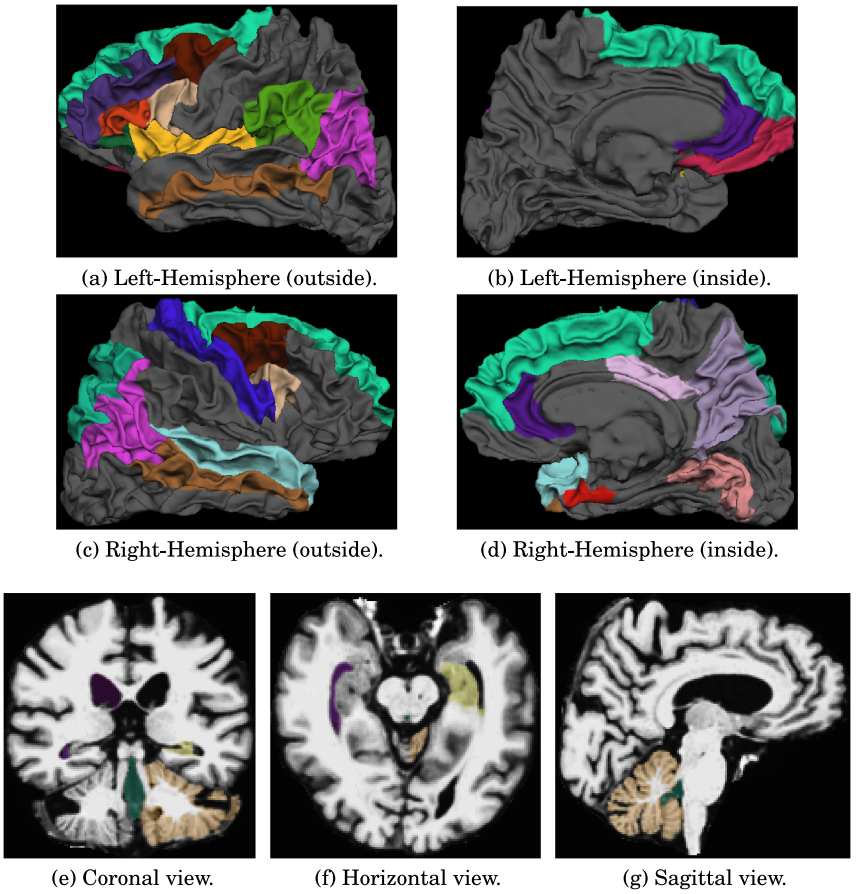
Fig. 10. The histogram of the top ROIs by stability selection. Middle temporal, caudal middle frontal, and hippocampus are selected, relevant according to existing AD domain knowledge.

To clear summarize the biomarkers identified by our method for multiple cognitive scores, we assume the p covariates to be divided into q disjoint groups $\mathcal{G}_\ell, \ell = 1, \ldots, q$, with each group having v_ℓ covariates, respectively. In the context of AD, each group corresponds to a 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, v_ℓ, is 1 or 4, and the number of groups q can be in the hundreds. Figure 10 illustrates the histogram of the ROIs name from the perspective of regions.

4.4.2 Clinical Relevance of Identified ROIs. We briefly discuss the clinical relevance of the ROIs identified based on statistical stability. The identified regions include middle temporal, caudal middle frontal, hippocampus (Risacher et al. 2009; Wang et al. 2009; Apostolova et al. 2006), which have been found to be predicted during disease progression. Furthermore, we show the brain maps of the top ROIs in Figure 11, including cortical ROIs and sub-cortical ROIs. These findings are consistent with their atrophy pattern and prediction power of AD found in the literature (Wang et al. 2012; Wan et al. 2012; Zhou et al. 2013).

We observe that different cognitive scores also share similar ROIs, which demonstrate that there exists a strong correlation among the multiple tasks of score prediction at multiple time points. For example, the number of the common top 10 features is 19 for the score of RAVLT, including TOTAL, TOT6, T30, and RECOG, which implies that these four scores are strongly correlated. Interestingly, the top 10 features of TRAILS.A and TRAILS.B are exactly different, indicating that they are weakly correlated.

On the whole, some important brain regions are selected by our method, 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), Entorhinal (Yan et al. 2015), Inferior lateral ventricle (Gutman et al. 2015; Wan et al. 2014), and Parahipp (Echávarri et al. 2011), which are highly relevant to the cognitive impairment. These findings are in accordance with the known knowledge that in the pathological pathway of AD.



(a) Left-Hemisphere (outside). (b) Left-Hemisphere (inside).

(c) Right-Hemisphere (outside). (d) Right-Hemisphere (inside).

(e) Coronal view. (f) Horizontal view. (g) Sagittal view.

Fig. 11. Brain maps of the top ROIs selected by FL-SGL2 using stability selection. (a)–(d) are cortical ROIs selected and (e)–(g) are sub-cortical ROIs selected.

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 is located in the temporal lobe of the brain, which are the role of the memory and spatial navigation. The Entorhinal cortex is the first area of the brain to be affected in AD, and it is the most heavily damaged cortex in AD (Van Hoesen et al. 1991). Hippocampus and entorhinal cortex have already been identified as areas steadily affected by AD (Braak and Braak 1985; Van Hoesen et al. 1991). Recent studies (Devanand et al. 2007; Khan et al. 2014; López et al. 2014) suggest that these are the first areas damaged by AD, therefore, can be considered as an important biomarker for diagnosing AD in early stages. Both hippocampus and entorhinal are part of the memory system, then it is expected that such areas relate to AD as memory loss is one of the primaries AD clinical signs (Burns and Iliffe 2009). Moreover, some recent results stress the importance of parahippocampal atrophy as an early biomarker of AD, since parahippocampal volume discriminates better than hippocampal volume between cases of healthy aging, MCI, and mild AD, in particular, in the early phase of the disease (Echávarri et al. 2011). Additionally, results also suggest that 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).

Table 7. Prediction performance results of 10 cognitive scores of six time points based on MRI features, demographic information, and ApoE genotyping information

	Ridge	Lasso	cFSGL	F-SGL	FL-SGL1	FL-SGL2
Score: ADAS						
nMSE	9.655±0.896$^{\dagger *}$	6.095±0.421$^{\dagger *}$	4.705±0.278$^{\dagger *}$	5.217±0.306$^{\dagger *}$	4.612±0.311	**4.584±0.301**
wR	0.599±0.028$^{\dagger *}$	0.665±0.022$^{\dagger *}$	0.769±0.016$^{\dagger *}$	0.723±0.018$^{\dagger *}$	0.773±0.017	**0.774±0.017**
BL rMSE	7.456±0.594	6.565±0.617	5.992±0.392	5.830±0.547	**5.758±0.424**	5.848±0.410
M6 rMSE	8.632±0.950	7.397±0.824	6.431±0.695	6.696±0.744	6.457±0.781	**6.383±0.734**
M12 rMSE	9.550±0.798	8.389±0.736	6.989±0.584	7.791±0.677	7.015±0.560	**6.936±0.578**
M24 rMSE	11.74±1.193	9.829±0.807	8.710±0.996	9.257±0.745	**8.547±0.895**	8.607±0.893
M36 rMSE	12.55±1.842	9.103±1.241	8.100±0.807	8.710±1.237	8.022±0.798	**7.877±0.874**
M48 rMSE	20.05±3.697	8.693±2.028	**7.438±1.335**	8.558±1.711	7.721±1.408	7.775±1.346
Score: MMSE						
nMSE	13.63±12.01$^{\dagger *}$	2.399±0.223$^{\dagger *}$	2.034±0.189$^{\dagger *}$	2.237±0.178$^{\dagger *}$	**1.970±0.172**	1.978±0.184
wR	0.468±0.037$^{\dagger *}$	0.617±0.036$^{\dagger *}$	0.686±0.033	0.644±0.036$^{\dagger *}$	**0.687±0.033**	0.686±0.034
BL rMSE	2.630±0.256	2.151±0.200	2.132±0.230	**2.074±0.225**	2.124±0.230	2.131±0.242
M6 rMSE	3.309±0.267	2.790±0.270	**2.512±0.269**	2.678±0.259	2.529±0.275	2.538±0.278
M12 rMSE	3.698±0.366	3.122±0.323	2.765±0.294	2.965±0.326	2.735±0.286	**2.727±0.278**
M24 rMSE	4.780±0.414	3.727±0.429	3.401±0.444	3.611±0.414	**3.334±0.428**	3.358±0.427
M36 rMSE	5.701±1.003	3.136±0.627	**2.795±0.439**	3.056±0.569	2.851±0.467	2.820±0.501
M48 rMSE	29.99±1.447	3.924±0.914	3.986±1.269	4.412±1.437	**3.044±0.810**	3.154±0.850
Score: RAVLT.TOTAL						
nMSE	15.94±1.091$^{\dagger *}$	8.523±0.497$^{\dagger *}$	6.216±0.347$^{\dagger *}$	7.284±0.440$^{\dagger *}$	**6.011±0.363**	6.015±0.353
wR	0.480±0.040$^{\dagger *}$	0.575±0.033$^{\dagger *}$	0.721±0.021$^{\dagger *}$	0.648±0.029$^{\dagger *}$	**0.731±0.020**	0.730±0.020
BL rMSE	10.45±0.886	9.142±0.642	8.387±0.689	8.366±0.654	**8.281±0.728**	8.384±0.709
M6 rMSE	10.82±0.923	9.443±0.902	8.198±0.760	8.768±0.776	8.204±0.736	**8.169±0.725**
M12 rMSE	12.10±0.938	10.70±0.919	8.958±0.684	9.747±0.841	8.802±0.713	**8.757±0.705**
M24 rMSE	13.91±1.484	11.58±1.154	9.413±0.987	10.63±0.859	9.140±0.797	**9.128±0.813**
M36 rMSE	15.64±1.784	11.19±1.319	9.017±1.067	10.55±1.205	8.779±1.033	**8.692±1.035**
M48 rMSE	41.24±3.362	11.51±1.659	9.420±1.844	11.44±1.630	**8.825±1.979**	9.067±1.947
Score: RAVLT.TOT6						
nMSE	3.640±0.291$^{\dagger *}$	2.676±0.137$^{\dagger *}$	2.132±0.206$^{\dagger *}$	2.434±0.132$^{\dagger *}$	2.060±0.149	**2.058±0.156**
wR	0.517±0.045$^{\dagger *}$	0.587±0.033$^{\dagger *}$	0.696±0.036$^{\dagger *}$	0.630±0.031$^{\dagger *}$	**0.708±0.028**	0.708±0.029
BL rMSE	3.481±0.267	3.141±0.189	2.987±0.265	3.051±0.182	**2.888±0.245**	2.918±0.245
M6 rMSE	3.313±0.325	3.048±0.260	2.728±0.224	2.941±0.239	2.697±0.202	**2.672±0.195**
M12 rMSE	3.618±0.307	3.291±0.246	2.892±0.244	3.157±0.213	2.862±0.239	**2.845±0.236**
M24 rMSE	3.872±0.393	3.380±0.336	2.980±0.348	3.213±0.316	2.943±0.308	**2.943±0.299**
M36 rMSE	4.141±0.735	3.437±0.372	2.840±0.375	3.251±0.336	2.823±0.406	**2.803±0.413**
M48 rMSE	7.529±1.232	4.272±0.768	3.358±0.819	3.515±0.534	**3.319±0.802**	3.406±0.810
Score: RAVLT.T30						
nMSE	3.639±0.247$^{\dagger *}$	2.854±0.129$^{\dagger *}$	2.218±0.162	2.595±0.143$^{\dagger *}$	**2.211±0.161**	2.219±0.157
wR	0.501±0.042$^{\dagger *}$	0.573±0.029$^{\dagger *}$	0.696±0.029†	0.619±0.033$^{\dagger *}$	**0.698±0.029**	0.696±0.028
BL rMSE	3.697±0.283	3.349±0.236	3.152±0.269	3.251±0.229	**3.105±0.272**	3.143±0.267
M6 rMSE	3.353±0.304	3.084±0.317	**2.794±0.183**	2.965±0.275	2.813±0.177	2.798±0.179

(Continued)

Table 7. Continued

	Ridge	Lasso	cFSGL	F-SGL	FL-SGL1	FL-SGL2
M12 rMSE	3.843±0.410	3.592±0.332	**3.087±0.307**	3.446±0.335	3.107±0.318	3.087±0.310
M24 rMSE	3.971±0.448	3.602±0.317	**3.059±0.314**	3.424±0.335	3.073±0.317	3.077±0.308
M36 rMSE	4.056±0.752	3.423±0.396	2.881±0.394	3.291±0.361	2.844±0.430	**2.833±0.403**
M48 rMSE	7.013±1.587	4.694±0.955	**3.538±0.640**	3.789±0.662	3.561±0.636	3.666±0.663
Score: RAVLT.RECOG						
nMSE	6.024±0.668$^{\dagger *}$	3.208±0.189$^{\dagger *}$	2.717±0.200*	2.987±0.192$^{\dagger *}$	2.692±0.226	**2.672±0.213**
wR	0.371±0.036$^{\dagger *}$	0.503±0.031$^{\dagger *}$	0.606±0.034$^{\dagger *}$	0.545±0.030$^{\dagger *}$	0.615±0.033	**0.620±0.033**
BL rMSE	4.204±0.304	3.511±0.243	3.385±0.278	3.426±0.249	**3.330±0.277**	3.374±0.284
M6 rMSE	4.484±0.360	3.789±0.346	3.510±0.369	3.687±0.352	3.524±0.407	**3.477±0.397**
M12 rMSE	4.576±0.338	3.679±0.231	3.378±0.244	3.569±0.249	3.379±0.241	**3.348±0.236**
M24 rMSE	4.773±0.503	3.688±0.268	3.287±0.383	3.574±0.272	3.220±0.379	**3.190±0.358**
M36 rMSE	5.148±0.475	3.527±0.260	**3.227±0.319**	3.388±0.306	3.278±0.383	3.249±0.378
M48 rMSE	12.60±0.884	4.615±0.759	3.233±0.785	3.848±0.999	**3.199±0.501**	3.238±0.541
Score: FLU.ANIM						
nMSE	9.011±0.741$^{\dagger *}$	4.636±0.327$^{\dagger *}$	3.631±0.330	4.259±0.282$^{\dagger *}$	3.598±0.319	**3.596±0.327**
wR	0.365±0.055$^{\dagger *}$	0.482±0.042$^{\dagger *}$	0.637±0.036	0.543±0.035$^{\dagger *}$	0.640±0.034	**0.641±0.035**
BL rMSE	6.139±0.504	5.180±0.414	4.748±0.416	4.912±0.438	**4.665±0.399**	4.748±0.402
M6 rMSE	5.805±0.493	4.938±0.553	4.410±0.348	4.704±0.459	4.452±0.325	**4.408±0.329**
M12 rMSE	6.386±0.800	5.448±0.862	4.677±0.840	5.132±0.783	4.659±0.844	**4.615±0.872**
M24 rMSE	7.054±0.634	5.526±0.572	4.877±0.580	5.378±0.557	4.830±0.548	**4.822±0.542**
M36 rMSE	7.452±0.930	5.168±0.681	4.439±0.576	5.066±0.643	**4.422±0.534**	4.424±0.529
M48 rMSE	20.44±2.203	6.057±1.368	**4.793±1.059**	6.139±1.329	4.893±1.025	4.825±0.962
Score: FLU.VEG						
nMSE	6.080±0.355$^{\dagger *}$	3.164±0.200$^{\dagger *}$	2.551±0.177$^{\dagger *}$	2.904±0.214$^{\dagger *}$	**2.523±0.185**	2.526±0.177
wR	0.436±0.044$^{\dagger *}$	0.569±0.041$^{\dagger *}$	0.676±0.030$^{\dagger *}$	0.613±0.036$^{\dagger *}$	**0.681±0.030**	0.681±0.029
BL rMSE	4.280±0.316	3.510±0.314	3.387±0.279	3.432±0.297	**3.341±0.291**	3.394±0.293
M6 rMSE	4.341±0.412	3.644±0.340	3.270±0.283	3.513±0.326	3.283±0.271	**3.260±0.267**
M12 rMSE	4.583±0.488	3.802±0.281	3.324±0.341	3.622±0.315	3.319±0.372	**3.298±0.367**
M24 rMSE	4.959±0.561	4.095±0.413	3.659±0.371	3.879±0.380	**3.614±0.360**	3.615±0.349
M36 rMSE	6.240±0.548	4.008±0.318	3.602±0.270	3.862±0.314	3.609±0.283	**3.566±0.267**
M48 rMSE	13.90±1.548	5.106±1.186	3.286±0.416	4.519±0.827	**3.178±0.503**	3.259±0.461
Score: TRAILS.A						
nMSE	32.79±2.971$^{\dagger *}$	23.36±1.910$^{\dagger *}$	18.48±1.688$^{\dagger *}$	20.95±1.739$^{\dagger *}$	17.79±1.652	**17.73±1.582**
wR	0.381±0.053$^{\dagger *}$	0.407±0.044$^{\dagger *}$	0.591±0.046$^{\dagger *}$	0.498±0.048$^{\dagger *}$	**0.614±0.049**	0.613±0.046
BL rMSE	27.28±3.172	24.85±3.196	22.08±2.481	23.38±2.747	**21.58±2.689**	21.97±2.678
M6 rMSE	27.71±3.081	24.74±3.353	22.35±2.517	23.63±3.202	22.20±2.567	**21.84±2.501**
M12 rMSE	28.41±3.422	25.51±4.226	21.77±2.378	24.05±3.301	20.96±1.926	**20.63±1.868**
M24 rMSE	30.38±5.968	27.51±5.388	23.35±4.718	25.77±4.839	22.76±4.379	**22.71±4.349**
M36 rMSE	31.53±6.059	23.87±5.127	**22.32±3.171**	23.19±4.644	22.41±2.972	22.41±3.045
M48 rMSE	51.83±10.60	27.23±11.88	24.06±7.931	25.38±10.89	**22.97±7.499**	23.63±7.868

(Continued)

Table 7. Continued

	Ridge	Lasso	cFSGL	F-SGL	FL-SGL1	FL-SGL2
Score: TRAILS.B						
nMSE	85.08±7.532$^{\dagger *}$	61.12±5.448$^{\dagger *}$	**51.39±4.527**	55.15±5.024$^{\dagger *}$	51.79±4.482	52.23±4.655
wR	0.466±0.046$^{\dagger *}$	0.511±0.041$^{\dagger *}$	**0.613±0.029**	0.567±0.035$^{\dagger *}$	0.610±0.031	0.607±0.033
BL rMSE	76.34±5.687	68.03±5.317	64.46±5.445	64.57±5.307	**64.15±5.554**	64.62±5.636
M6 rMSE	76.37±8.422	68.32±6.678	**63.17±6.524**	64.54±6.543	63.26±6.299	63.20±6.250
M12 rMSE	75.16±7.410	69.11±6.308	**62.47±6.082**	65.72±6.268	63.25±6.479	63.24±6.487
M24 rMSE	87.49±12.82	74.74±9.097	**66.53±8.602**	70.20±8.755	66.66±8.952	66.59±8.895
M36 rMSE	91.11±21.42	74.34±19.97	**68.16±20.64**	71.44±20.72	68.58±19.24	69.60±19.09
M48 rMSE	131.3±21.55	66.99±19.34	**56.11±16.37**	66.80±16.98	56.80±16.16	58.18±16.72

Note that the best results are boldfaced, superscript symbols † and * indicate that FL-SGL1 and FL-SGL2, respectively, significantly outperformed that method on that score. Student's t-test at a level of 0.05 was used.

Another area identified by our method, fusiform gyrus has been primarily involved in visual perception and recognition, essential for tasks like face/object recognition and color information processing. Several studies have related AD with changes in fusiform areas. Cronin-Golomb (1995) reported visual cognition deficits accompanying the progression of AD. MCI patients had widespread changes in fusiform connectivity during the performance of a face-matching task, as reported by Bokde et al. (2006).

4.5 MRI Features and Demographic Information

We also explore the prediction models by involving two other modalities: demographic information (age, gender, and years of education) and ApoE genotyping information. We followed the same experimental procedure as described in Section 4.3. As it is shown in Table 7, the FL-SGL method with multi-modality can achieve a higher performance than the one with only MRI modality with respect to nMSE and wR. Specially, the nMSE and wR of FL-SGL2 have improved from 4.975 to 4.584 ($p < 10^{-8}$) and from 0.749 to 0.774 ($p < 10^{-10}$) for the prediction of ADAS, respectively, and from 2.152 to 1.978 ($p < 10^{-7}$) and from 0.651 to 0.686 ($p < 10^{-9}$) for the prediction of MMSE. In addition, we also witness the improvement in prediction performance at all time points.

5 CONCLUSIONS

In this article, we investigated the progression of AD by means of multiple cognitive scores. We proposed a MTL formulation with a general temporal smoothness regularization that can jointly predict the cognitive scores based on a set of MRI features extracted from imaging data. The proposed model is capable of revealing the relationship between longitudinal cognitive measures and neuroimaging markers. Two efficient ADMM) methods are presented to tackle the associated optimization problem.

We performed a longitudinal stability selection using our proposed formulation to identify and analyze the temporal patterns of the biomarkers selected by our models. An experimental study on the ADNI dataset was conducted to validated the effectiveness of the proposed method by comparing with the state-of-the-art and single-task learning methods. The proposed algorithm not only showed the highest prediction performance, but also demonstrated the ability to accurately identify imaging biomarkers that are consistent with prior knowledge.

While the current study illustrates the power of MTL, especially FL-SGL formulations, each cognitive score was considered separately with multiple tasks corresponding to the same cognitive score across multiple time points. Since the cognitive scores are different ways to measure the same

underlying medical condition, we expect that a more general MTL framework that considers all cognitive scores across all time points simultaneously may be more effective as a predictive model. Such general models will be investigated as part of our future work.

REFERENCES

L. G. Apostolova, Po H. Lu, S. Rogers, R. A. Dutton, K. M. Hayashi, A. W. Toga, J. L. Cummings, and P. M. Thompson. 2006. 3D mapping of mini-mental state examination performance in clinical and preclinical Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* 20, 4 (2006), 224–231.

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

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

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. 2012. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.* 4, 1 (2012), 1–106.

N. L. Batsch and M. S. Mittelman. 2015. World Alzheimer report 2012. *Overcoming the Stigma of Dementia. Alzheimer's Disease International (ADI)* 5 (2015).

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

A. L. W. Bokde, P. Lopez-Bayo, T. Meindl, S. Pechler, C. Born, F. Faltraco, S. J. Teipel, H.-J. Möller, and H. Hampel. 2006. Functional connectivity of the fusiform gyrus during a face-matching task in subjects with mild cognitive impairment. *Brain* 129, 5 (2006), 1113–1124.

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

H. Braak and E. Braak. 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 (1985), 325–332.

A. Burns and S. Iliffe. 2009. Alzheimers disease. *BMJ* 338, b158.

R. Caruana. 1997. Multitask learning. *Machine Learning* 28, 1 (1997), 41–75.

C. Chen, B. He, Y. Ye, and X. Yuan. 2016. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Math. Program.* 155, 1–2 (2016), 57–79.

F. Chung. 1997. *Spectral Graph Theory.* vol. 92. American Mathematical Soc.

A. Cronin-Golomb. 1995. Vision in alzheimer's disease. *Gerontol.* 35, 3 (1995), 370–376.

W. Deng, M. Lai, Z. Peng, and W. Yin. 2017. Parallel multi-block ADMM with O (1/k) convergence. *Journal of Scientific Computing* 71, 2 (2017), 712–736.

W. Deng, W. Yin, and Y. Zhang. 2013. Group sparse optimization by alternating direction method. In *SPIE Optical Engineering+Applications.* 88580R–88580R.

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

R. S. Doody, V. Pavlik, P. Massman, S. Rountree, E. Darby, and W. Chan. 2010. Predicting progression of Alzheimer's disease. *Alzheimer's Res. Ther.* 2, 1 (2010), 2.

N. R. Draper and Harry Smith. 2014. *Applied Regression Analysis.* vol. 326. John Wiley & Sons.

C. Echávarri, P. Aalten, H. B. M. Uylings, H. I. L. Jacobs, P. J. Visser, E. H. B. M. Gronenschild, F. R. J. Verhey, and S. Burgmans. 2011. Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. *Brain Struct. Funct.* 215, 3–4 (2011), 265–271.

G. Frisoni, N. Fox, C. Jack, P. Scheltens, and P. Thompson. 2010. The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 2 (2010), 67–77.

X. W. Gao, R. Hui, and Z. Tian. 2017. Classification of CT brain images based on deep learning networks. *Comput. Methods Program. Biomed.* 138 (2017), 49–56.

A. R. Gonçalves, F. J. Von Zuben, and A. Banerjee. 2016. Multi-task sparse structure learning with gaussian copula models. *J. Mach. Learn. Res.* 17, 33 (2016), 1–30.

S. J. Greene, R. J. Killiany, Alzheimer's Disease Neuroimaging Initiative, and others. 2010. Subregions of the inferior parietal lobule are affected in the progression to Alzheimer's disease. *Neurobiol. Aging* 31, 8 (2010), 1304–1311.

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

B. He, M. Tao, and X. Yuan. 2012. Alternating direction method with gaussian back substitution for separable convex programming. *SIAM J. Optim.* 22, 2 (2012), 313–340.

M. Hong, T. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z. Luo. 2014. A block successive upper bound minimization method of multipliers for linearly constrained convex optimization. arXiv:1401.7079.

M. Hong and Z. Luo. 2017. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming* 162, 1-2 (2017), 165–199.

L. Huang, Y. Jin, Y. Gao, K. Thung, and D. Shen. 2016. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiol. Aging* 46 (2016), 180–191.

K. Ito, B. Corrigan, Q. Zhao, J. French, R. Miller, H. Soares, E. Katz, T. Nicholas, B. Billing, R. Anziano, T. Fullerton, and Alzheimer's Disease Neuroimaging Initiative. 2011. Disease progression model for cognitive deterioration from Alzheimer's disease neuroimaging initiative database. *Alzheimer's Dementia* 7 (2011), 151–160.

U. A. Khan, L. Liu, F. A. Provenzano, D. E. Berman, C. P. Profaci, R. Sloan, R. Mayeux, K. E. Duff, and S. A. Small. 2014. Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical Alzheimer's disease. *Nat. Neurosci.* 17, 2 (2014), 304–311.

J. Liu and J. Ye. 2009. Efficient euclidean projections in linear time. In *International Conference on Machine Learning*. ACM, 657–664.

J. Liu, L. Yuan, and J. Ye. 2010. An efficient algorithm for a class of fused lasso problems. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 323–332.

X. Liu, P. Cao, D. Zhao, and A. Banerjee. 2016. Multi-task spare group lasso for characterizing Alzheimer's disease. In *5th Workshop on Data Mining for Medicine and Healthcare*. 49.

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

M. López, R. Bruña, S. Aurtenetxe, J. Pineda-Pardo, A. Marcos, J. Arrazola, A. I. Reinoso, P. Montejo, R. Bajo, and F. Maestú. 2014. Alpha-band hypersynchronization in progressive mild cognitive impairment: A magnetoencephalography study. *J. Neurosci.* 34, 44 (2014), 14551–14559.

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

R. Merris. 1994. Laplacian matrices of graphs: A survey. *Linear Algebra Appl.* 197 (1994), 143–176.

C. Misra, Y. Fan, and C. Davatzikos. 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *Neuroimage* 44, 4 (2009), 1415–1422.

P. Rai, A. Kumar, and H. Daume III. 2012. Simultaneously leveraging output and task structures for multiple-output regression. In *Advances in Neural Information Processing Systems*. 3185–3193.

S. L. Risacher, A. J. Saykin, J. D. West, L. Shen, H. A. Firpi, B. C. McDonald, and the Alzheimer's Disease Neuroimaging Initiative (ADNI). 2009. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr. Alzheimer Res.* 6 (2009), 347–361.

C. M. Stonningtona, C. Chub, S. Klöppelb, C. R. Jack Jr., J. Ashburner, R. S. J. Frackowiakf, and Alzheimer Disease Neuroimaging Initiative. 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *Neuroimage* 51, 4 (2010), 1405–1413.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc.: Series B (Statistical Methodology)* 67, 1 (2005), 91–108.

T. N. Tombaugh. 2005. Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Arch. Clin. Neuropsychol.* 20, 4 (2005), 485–503.

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

P. Vemuri, H. J. Wiste, S. D. Weigand, L. M. Shaw, J. Q. Trojanowski, M. W. Weiner, D. S. Knopman, R. C. Petersen, and C. R. Jack. 2009. MRI and CSF biomarkers in normal, MCI, and AD subjects predicting future clinical change. *Neurology* 73, 4 (2009), 294–301.

P. J. Visser, F. R. J. Verhey, P. A. M. Hofman, P. Scheltens, and J. Jolles. 2002. Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. *J. Neurol., Neurosurg. Psych.* 72, 4 (2002), 491–497.

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

J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen. 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*. 940–947.

H. Wang and A. Banerjee. 2014. Bregman alternating direction method of multipliers. In *Advances in Neural Information Processing Systems*. 2816–2824.

H. Wang, A. Banerjee, and Z. Luo. 2014. Parallel direction method of multipliers. In *Advances in Neural Information Processing Systems*. 181–189.

H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen. 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*.

L. Wang, F. C. Goldstein, E. Veledar, A. I. Levey, J. J. Lah, C. C. Meltzer, C. A. Holder, and H. Mao. 2009. Alterations in cortical thickness and white matter integrity in mild cognitive impairment measured by whole brain cortical thickness mapping and diffusion tensor imaging. *AJNR Am. J. Neuroradiol.* 30, 5 (2009), 893–899.

Y. Wang, J. Yang, W. Yin, and Y. Zhang. 2008. A new alternating minimization algorithm for total variation image reconstruction. *SIAM J. Imag. Sci.* 1, 3 (2008), 248–272.

L. Wasserman. 2006. *All of Nonparametric Statistics*. Springer-Verlag New York.

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

M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, and J. C. Morris. 2017. Recent publications from the Alzheimer's disease neuroimaging initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's Dementia* 13, 4 (2017), e1–e85.

L. Xu, X. Wu, K. Chen, and L. Yao. 2015. Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment. *Comput. Methods Programs Biomed.* 122, 2 (2015), 182–190.

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

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

J. Yang and Y. Zhang. 2011. Alternating direction algorithms for L1-problems in compressive sensing. *SIAM J. Sci. Comput.* 33, 1 (2011), 250–278.

J. Ye, K. Chen, T. Wu, J. Li, Z. Zhao, R. Patel, M. Bae R. Janardan, H. Liu, G. Alexander, and E. Reiman. 2008. Heterogeneous data fusion for alzheimer's disease study. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1025–1033.

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

Y. L. Yu. 2013a. Better approximation and faster algorithm using the proximal average. In *Advances in Neural Information Processing Systems*. 458–466.

Y. L. Yu. 2013b. On decomposing the proximal map. In *Advances in Neural Information Processing Systems*. 91–99.

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

D. Zhang, D. Shen, and 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 (2012), 895–907.

Y. Zhang and D. Y. Yeung. 2010. A convex formulation for learning task relationships in multi-task learning. In *Conference on Uncertainty in Artificial Intelligence*. 733–742.

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

J. Zhou, L. Yuan, J. Liu, and J. Ye. 2011. A multi-task learning formulation for predicting disease progression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 814–822.

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

Received May 2017; revised March 2018; accepted April 2018