NeuroImage 78 (2013) 233–248

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Modeling disease progression via multi-task learning

Jiayu Zhou [a,b], Jun Liu [a,b], Vaibhav A. Narayan [c], Jieping Ye [a,b,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Center for Evolutionary Medicine and Informatics, The Biodesign Institute, ASU, Tempe, AZ, USA
[b] Department of Computer Science and Engineering, ASU, Tempe, AZ, USA
[c] Johnson & Johnson Pharmaceutical Research & Development, LLC, Titusville, NJ, USA

ARTICLE INFO

Article history:
Accepted 28 March 2013
Available online 12 April 2013

Keywords:
Alzheimer's disease
Disease progression
Multi-task learning
Fused Lasso
MMSE
ADAS-Cog

ABSTRACT

Alzheimer's disease (AD), the most common type of dementia, is a severe neurodegenerative disorder. Identifying biomarkers that can track the progress of the disease has recently received increasing attentions in AD research. An accurate prediction of disease progression would facilitate optimal decision-making for clinicians and patients. A definitive diagnosis of AD requires autopsy confirmation, thus many clinical/cognitive measures including Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog) have been designed to evaluate the cognitive status of the patients and used as important criteria for clinical diagnosis of probable AD. In this paper, we consider the problem of predicting disease progression measured by the cognitive scores and selecting biomarkers predictive of the progression. Specifically, we formulate the prediction problem as a multi-task regression problem by considering the prediction at each time point as a task and propose two novel multi-task learning formulations. We have performed extensive experiments using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Specifically, we use the baseline MRI features to predict MMSE/ADAS-Cog scores in the next 4 years. Results demonstrate the effectiveness of the proposed multi-task learning formulations for disease progression in comparison with single-task learning algorithms including ridge regression and Lasso. We also perform longitudinal stability selection to identify and analyze the temporal patterns of biomarkers in disease progression. We observe that cortical thickness average of left middle temporal, cortical thickness average of left and right Entorhinal, and white matter volume of left Hippocampus play significant roles in predicting ADAS-Cog at all time points. We also observe that several MRI biomarkers provide significant information for predicting MMSE scores for the first 2 years, however very few are shown to be significant in predicting MMSE score at later stages. The lack of predictable MRI biomarkers in later stages may contribute to the lower prediction performance of MMSE than that of ADAS-Cog in our study and other related studies.

Introduction

Alzheimer's disease (AD), a severe neurodegenerative disorder, is characterized by loss of memory and reduction of cognitive function due to progressive impairment of neurons and their connections, leading directly to death (Khachaturian, 1985). AD accounts for 60–70% of age-related dementia; it currently affects about 5.3 million individuals in United States and more than 30 million worldwide and the number is projected to be over 114 million by 2050 (A. Association, 2010; Wimo et al., 2003). Alzheimer's disease has been not only the substantial financial burden to the health care system but also the psychological and emotional burdens to patients and their families. Currently there is no cure for Alzheimer's and efforts are underway to develop sensitive and consistent biomarkers for AD. In order to better understand the disease, an important area that has recently received increasing attention is to understand how the disease progresses and identify related pathological biomarkers for the progression. Realizing its importance, NIH in 2003 funded the Alzheimer's Disease Neuroimaging Initiative (ADNI). The initiative is facilitating the scientific evaluation of neuroimaging data including magnetic resonance imaging (MRI), positron emission tomography (PET), other biomarkers, and clinical and neuropsychological assessments for predicting the onset and progression of MCI (Mild Cognitive Impairment) and AD. The identification of sensitive and specific markers of very early AD progression will facilitate the diagnosis of early AD and the development, assessment, and monitoring of new treatments.

* Corresponding author at: Department of Computer Science and Engineering, Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, 699 S. Mill Ave, Tempe, AZ 85287, USA.
E-mail address: jieping.ye@asu.edu (J. Ye).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.03.073

A definitive diagnosis of AD can only be made through an analysis of brain tissue during a brain biopsy or autopsy (Jeffrey et al., 2003). Many clinical/cognitive measures such as Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog) have been designed to evaluate the cognitive status of the patients and they have been used as important criteria for clinical diagnosis of probable AD (McKhann et al., 1984). Previous studies have shown the correlation between MMSE and the underlying AD pathology and progressive deterioration of functional ability (Jeffrey et al., 2003). ADAS-Cog is the gold standard in AD drug trial for cognitive function assessment (Rosen et al., 1984). Since neurodegeneration of AD proceeds years before the onset of the disease and the therapeutic intervention is more effective in the early stage of the disease, there is thus an urgent need to (1) accurately predict the progression of the disease measured by cognitive scores, e.g., MMSE and ADAS-Cog, and (2) identify a small set of biomarkers (measurements) and risk factors most predictive of the progression. The prime candidate biomarkers and risk factors for tracking disease progression include neuroimages such as MRI, cerebrospinal fluid (CSF), and baseline clinical assessments (Dubois et al., 2007).

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 (Ito et al., 2010; Tombaugh, 2005). The relationship between cognitive scores and imaging markers based on MRI such as gray matter volumes, density and loss (Apostolova et al., 2006; Chetelat and Baron, 2003; Frisoni et al., 2002; Frisoni et al., 2010; Stonnington et al., 2010), shape of ventricles (Ferrarini et al., 2008; Thompson et al., 2004) and hippocampal (Thompson et al., 2004) has been explored by correlating these features with baseline MMSE scores. Duchesne et al. showed that the intensity and volume of medial temporal lobe altogether with other risk factors and the gray matter were correlated with the one-year MMSE score (Duchesne et al., 2009), which allowed us to predict near-future clinical scores of patients. Murphy et al. examined the relations between 6-month atrophy patterns in medial temporal region and memory reduction in terms of clinical scores (Murphy et al., 2010). To predict the longitudinal response to AD progression, Ashford and Schmitt built a model with horologic function using “time-index” to measure the rate of dementia progression (Ashford and Schmitt, 2001). In (Davatzikos et al., 2009), the so-called SPARE-AD index was proposed based on spatial patterns of brain atrophy and its linear effect against MMSE was reported. In a more recent study, Ito et al. modeled the progression rate of cognitive scores using power functions (Ito et al., 2010).

There are two types of progression models that have been commonly used in the literature: the regression model (Duchesne et al., 2009; Stonnington et al., 2010) and the survival model (Pearson et al., 2005; Vemuri et al., 2009). The correlation between the ground truth and the prediction, and the squared error between the two are commonly used to evaluate the progression models (Duchesne et al., 2009; Stonnington et al., 2010). Many existing works consider a small number of input features, and the model building involves an iterative process in which the features are added to the model sequentially (Ito et al., 2010; Walhovd et al., 2010); alternatively, univariate analysis is performed individually on all covariates and those who exceed a certain significance threshold are included in the model (Murphy et al., 2010). For high-dimensional data, such as neuroimages (i.e., MRI and/or PET), the methods of sequentially evaluating individual features are suboptimal. In such cases, dimension reduction techniques such as principle component analysis are commonly applied to project the data into a lower-dimensional space (Duchesne et al., 2009). One disadvantage of dimension reduction is that the models are no longer interpretable. A better alternative is to use feature selection in modeling the disease progression (Stonnington et al., 2010). Most existing works focus on the prediction of target at a single time point (baseline Stonnington et al., 2010, or one year Duchesne et al., 2009); however, a joint analysis of the tasks from multiple time points is expected to improve the performance especially when the number of subjects is small and the number of input features is large.

To address the aforementioned challenges, we propose to develop novel multi-task learning formulations to model disease progression. The idea of multi-task learning is to utilize the intrinsic relationships among multiple related tasks in order to improve the prediction performance; it is most effective when the number of samples for each task is small. One of the key issues in multi-task learning is to identify how the tasks are related and build learning models to capture such task relatedness. One way of modeling multi-task relationship is to assume that all tasks are related and the task models are closed to each other (Evgeniou et al., 2006), or the tasks are clustered into groups (Bakker and Heskes, 2003; Jacob et al., 2009; Thrun and O'Sullivan, 1998; Zhou et al., 2011). Alternatively, one can assume that the tasks share a common subspace (Ando and Zhang, 2005; Chen et al., 2009), or a common set of features (Argyriou et al., 2008; Obozinski et al., 2006).

In this paper, we propose novel multi-task learning formulations for predicting disease progression measured by the clinical scores (ADAS-Cog and MMSE). Specifically, we formulate the prediction of clinical scores at a sequence of time points as a multi-task regression problem, where each task concerns the prediction of a clinical score at one time point. For the disease progression considered in this paper, it is reasonable to assume that a small subset of features is predictive of the progression, and the multiple regression models from different time points satisfy the smoothness property, that is, the difference of the cognitive scores between two successive time points is small. To this end, we develop a novel multi-task learning formulation based on a temporal group Lasso regularizer (TGL). The regularizer consists of two components including an $\ell_{2,1}$-norm penalty (Yuan and Lin, 2006) on the regression weight vectors, which ensures that a small subset of features will be selected for the regression models at all time points, and a temporal smoothness term, which ensures a small deviation between two regression models at successive time points. In order to better capture the *temporal patterns* of the biomarkers in disease progression (Caroli et al., 2010; Jack et al., 2010), we further propose a convex fused sparse group Lasso (cFSGL) formulation that 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 sparse group Lasso penalty, and in the meantime incorporates the temporal smoothness using the fused Lasso penalty. The proposed formulation is challenging to solve due to the use of non-smooth penalties including the sparse group Lasso and fused penalties. We show that the proximal operator associated with the optimization problem of cFSGL exhibits a certain decomposition property and can be solved efficiently. Therefore cFSGL can be efficiently solved using the accelerated gradient method (Nemirovski, 2005; Nesterov, 2004).

We have performed extensive experiments to demonstrate the effectiveness of the proposed models using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Specifically, we use the baseline MRI features to predict MMSE/ADAS-Cog scores in the next 4 years. A set of 648 subjects including 191 cognitively normal older individuals (NL), 319 patients with mild cognitive impairment (MCI), and 138 Alzheimer's disease patients (AD), are included in our study. Our experimental results show that the proposed multi-task learning formulations outperform single-task learning algorithms including ridge regression and Lasso for predicting future MMSE/ADAS-Cog scores. We also observe that including demographic and ApoE genotyping information as additional covariates further improves the prediction performance. We apply our models on the subgroup that only consists of MCI converters and AD patients and we observe similar improved performance from the proposed models. We have also performed longitudinal stability selection using our proposed formulations to identify and analyze the temporal patterns of biomarkers selected in our models. We observe that the cortical thickness average of left middle temporal, the cortical thickness

average of left and right Entorhinal, and the white matter volume of left Hippocampus play significant roles in predicting ADAS-Cog at all time points. We also observe that several MRI biomarkers provide significant information for predicting MMSE scores for the first 2 years, however very few are shown to be significant in predicting MMSE score at later stages. The lack of predictable MRI biomarkers in later stages may contribute to the lower prediction performance of MMSE than that of ADAS-Cog in our study and other related studies. We further study the specific progression model for MCI patients and observe that in most cases the prediction performance witnesses improvement with AD and NL samples included in the training step.

Subjects and methods

Subjects

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals (NL) to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

The ADNI project is a longitudinal study, where a variety of measurements are collected repeatedly over a 6-month or 1-year interval. The date when the patient performs the screening in the hospital for the first time is called *baseline*, and the time point for the follow-up visits is denoted by the duration starting from the baseline. For instance, we use the notation "M06" to denote the time point half year after the first visit. Currently ADNI has up to 48 months' follow-up data available for some patients. However, many patients drop out from the study for many reasons.

In ADNI, all participants received 1.5 Tesla (T) structural MRI. The MRI image features in this study were based on the imaging data from the ADNI database processed by the UCSF team, who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). Details of the analysis procedure are available at http://adni.loni.ucla.edu/research/mri-post-processing/. More details on ADNI MRI imaging instrumentation and procedures (Jack et al., 2008) may be found at ADNI website (http://adni.loni.ucla.edu). We downloaded the MRI data from ADNI website and further performed the following preprocessing steps:

- remove features with more than 1000 missing entries (for all patients and all time points)[2];
- remove image records with failed quality control;
- exclude patients without baseline MRI records;
- complete the missing entries using the average value.

After the preprocessing procedure, there are a total of 648 subjects (138 AD, 319 MCI and 191 NC) and 305 MRI features. The MRI features can be grouped into 5 categories: average cortical thickness, standard deviation in cortical thickness, the volumes of cortical parcellations (based on regions of interest automatically segmented in the cortex), the volumes of specific white matter parcellations, and the total surface area of the cortex. The demographic information of subjects used in this study at different time points is given in Table 1.

Modeling disease progression via temporal group Lasso

In the longitudinal AD study, we measure the cognitive scores of selected patients repeatedly at multiple time points. By considering the prediction of cognitive scores at a single time point as a regression task, we formulate the prediction of clinical scores at multiple future time points as a multi-task regression problem. We employ multi-task regression formulations instead of solving a set of independent regression problems since the intrinsic temporal smoothness information among different tasks can be incorporated into the model as prior knowledge.

Consider a multi-task regression problem of t time points with n training samples of d features. Let $\{\mathbf{x}_1, \mathbf{x}_2 \cdots, \mathbf{x}_n\}$ be the input data at the baseline, and $\{\mathbf{y}_1, \mathbf{y}_2 \cdots, \mathbf{y}_n\}$ be the targets, where each $\mathbf{x}_i \in \mathbb{R}^d$ represents a sample (patient), and $\mathbf{y}_i \in \mathbb{R}^t$ is the corresponding target (clinical scores) at different time points. In this paper we employ linear models for the prediction. Specifically, the prediction model for the ith time point is given by $f^i(\mathbf{x}) = \mathbf{x}^T\mathbf{w}^i$, where $\mathbf{w}^i$ is the weight vector of the model. Let $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ be the data matrix, $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times t}$ be the target matrix, and $W = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^t] \in \mathbb{R}^{d \times t}$ be the weight matrix. One simple approach is to estimate W by minimizing the following objective function:

$$\min_W \|XW - Y\|_F^2 + \theta_1 \|W\|_F^2,$$

where the first term measures the empirical error on the training data, $\theta_1 > 0$ is a regularization parameter, and $\|W\|_F$ is the Frobenius

Table 1
Demographic information of subjects at different time points. There are three types of subjects included in the study: cognitively normal older individuals (NL), patients with mild cognitive impairment (MCI) and Alzheimer's disease (AD) patients. In this study, if an MCI patient converts to AD patient within the 48 months after screening, then we consider the patient to be an MCI convert patient (MCI-C), or it is an MCI stable patient (MCI-S). The methods in this paper use MRI features only from the baseline. In this table the sample size indicates the number of patients that has baseline MRI features and corresponding target cognitive scores (MMSE or ADAS-Cog) at future time points.

Time point	Attribute	MMSE	ADAS-Cog
M06	Sample size (NL, MCI-S, MCI-C, AD)	648 (191, 177, 142, 138)	648 (191, 177, 142, 138)
	ApoE-ε4 copies (0, 1, 2)	(335, 242, 71)	(335, 242, 71)
	Age	75.2 ± 6.7	75.2 ± 6.7
M12	Sample size (NL, MCI-S, MCI-C, AD)	642 (190, 173, 142, 137)	638 (188, 173, 141, 136)
	ApoE-ε4 copies (0, 1, 2)	(332, 240, 70)	(311, 238, 69)
	Age	75.2 ± 6.7	75.2 ± 6.7
M24	Sample size (NL, MCI-S, MCI-C, AD)	569 (183, 144, 125, 117)	564 (182, 144, 125, 113)
	ApoE-ε4 copies (0, 1, 2)	(290, 216, 63)	(287, 214, 63)
	Age	75.2 ± 6.6	75.2 ± 6.6
M36	Sample size (NL, MCI-S, MCI-C, AD)	389 (161, 119, 99, 10)	377 (156, 116, 95, 10)
	ApoE-ε4 copies (0, 1, 2)	(226, 131, 32)	(216, 129, 32)
	Age	75.2 ± 6.4	75.2 ± 6.3
M48	Sample size (NL, MCI-S, MCI-C, AD)	87 (47, 13, 25, 2)	85 (47, 13, 23, 2)
	ApoE-ε4 copies (0, 1, 2)	(51, 27, 9)	(49, 27, 9)
	Age	74.7 ± 5.2	74.6 ± 5.2

[2] The following features are deleted due to too many missing entries: ST100SV, ST122SV, ST126SV,ST22CV, ST22SA, ST22TA, ST22TS, ST28CV, ST33SV, ST41SV, ST63SV, ST67SV, ST81CV, ST81SA, ST81TA, ST81TS, ST87CV, ST92SV and ST8SV.

norm, defined as $\sqrt{\sum_{i=1}^{d}\sum_{j=1}^{t}W_{i,j}^2}$. The formulation is illustrated in Fig. 1. The regression method above is known as the *ridge regression* and it admits an analytical solution given by:

$$W = \left(X^T X + \theta_1 I\right)^{-1} X^T Y.$$

In building models with high dimensional features ($d \gg n$), feature selection methods are typically employed to identify a small set of relevant features. Lasso (Tibshirani, 1996), is a popular method for sparse linear regression, which simultaneously performs feature selection and regression. In the context of disease progression, the Lasso formulation solves the following optimization problem:

$$\min_W \|XW - Y\|_F^2 + \theta_1 \|W\|_1,$$

where $\|W\|_1$ is the ℓ_1 norm of W defined as $\sum_{i=1}^{d}\sum_{j=1}^{t}|W_{ij}|$.

One major limitation of the regression models above is that the tasks at different time points are assumed to be independent with each other, which is not the case in the longitudinal AD study considered in this paper.

Temporal smoothness prior

Applying single task learning methods such as ridge or Lasso regression on modeling disease progression often yields fluctuated prediction values at different time points for one patient, as shown in Fig. 2. In the course of disease progression, it is reasonable to assume that the difference of the cognitive scores between two successive time points is relatively small. During the inference of our models, for a patient i with two consecutive predictions $\hat{y}_i^{(j)}$ and $\hat{y}_i^{(j+1)}$ at time point j and $j+1$ respectively, a large difference between the predictions $\left|\hat{y}_i^{(j)} - \hat{y}_i^{(j+1)}\right|$ is discouraged. Since we use linear models ($y_i^{(j)} \approx \hat{y}_i^{(j)} = \mathbf{x}_i^T \mathbf{w}^j$), the difference between the predictions can be related to the difference between models at those time points:

$$\left|\hat{y}_i^{(j)} - \hat{y}_i^{(j+1)}\right| = \left|\mathbf{x}_i^T \mathbf{w}^j - \mathbf{x}_i^T \mathbf{w}^{j+1}\right| = \left|\mathbf{x}_i^T \left(\mathbf{w}^j - \mathbf{w}^{j+1}\right)\right|. \quad (1)$$

Inspired by Eq. (1), in order to capture the temporal smoothness of the cognitive scores at different time points, we introduce a regularization term in the regression model that penalizes large deviations between predictions at neighboring time points, resulting in the following formulation:

$$\min_W \|XW - Y\|_F^2 + \theta_1 \|W\|_F^2 + \theta_2 \sum_{i=1}^{t-1} \left\|\mathbf{w}^i - \mathbf{w}^{i+1}\right\|_2^2, \quad (2)$$

where $\theta_2 \geq 0$ is a regularization parameter controlling the temporal smoothness. This temporal smoothness term can be expressed as:

$$\sum_{i=1}^{t-1} \left\|\mathbf{w}^i - \mathbf{w}^{i+1}\right\|_F^2 = \|WH\|_F^2,$$

where $H \in \mathbb{R}^{t \times (t-1)}$ is defined as follows: $H_{ij} = 1$ if $i = j$, $H_{ij} = -1$ if $i = j+1$, and $H_{ij} = 0$ otherwise. The formulation in Eq. (2) becomes:

$$\min_W \|XW - Y\|_F^2 + \theta_1 \|W\|_F^2 + \theta_2 \|WH\|_F^2. \quad (3)$$

The optimization problem in Eq. (3) admits an analytical solution, as shown in Appendix A. We want to emphasize that the temporal smoothness is only employed during the inference of the model, and when it comes to the prediction phase only baseline features are needed to compute the predicted cognitive scores at the future time points. This is also the case for other models proposed in the paper.

Dealing with incomplete data

The clinical scores for many patients are missing at some time points, i.e., the target vector $y_i \in \mathbb{R}^t$ may not be complete. A simple strategy is to remove all patients with missing target values, which, however, significantly reduces the number of samples. We consider extending the formulation in Eq. (3) with missing target values in the training process. In this case, the analytical solution to Eq. (3) no longer exists. We show how the algorithm above can be adapted to deal with missing target values.

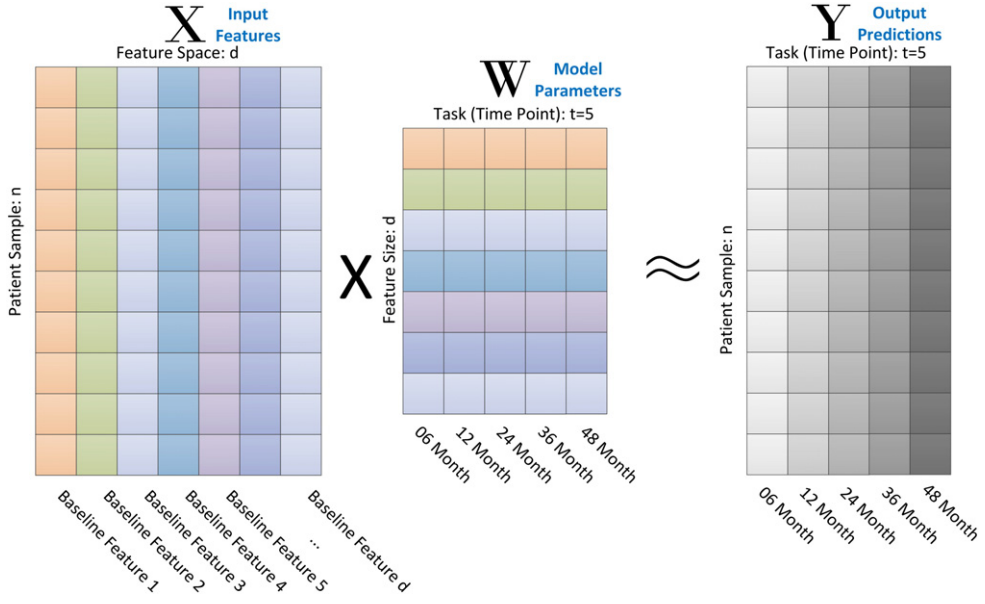

Fig. 1. Illustration of the prediction model. We denote $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ as the data matrix, $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times t}$ as the target matrix, and $W = [\mathbf{w}^1, , \mathbf{w}^2, \ldots, , \mathbf{w}^t] \in \mathbb{R}^{d \times t}$ as the weight matrix. Specifically, for the input matrix X, each row represents a patient and each column represents a feature at baseline, and for the output matrix Y, each row corresponds to a patient, and each column corresponds to the score at a future time point. In the prediction model we assume a linear relationship between input X and output Y, i.e., for the ith patient, we have $\mathbf{x}_i^T W \approx \mathbf{y}_i^T$.

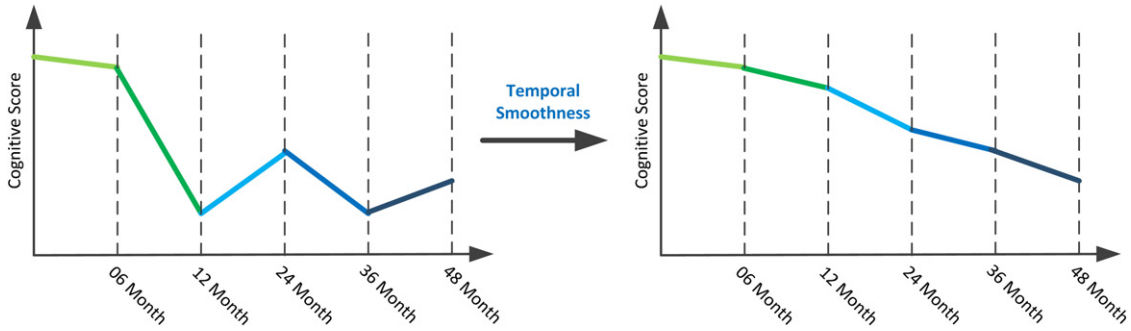

Fig. 2. Illustration of temporal smoothness. We assume that the difference of the cognitive scores between two successive time points is relatively small (right figure). Since we use linear predictive models, the difference between the predicted cognitive scores can be related to the difference between models at those time points, and therefore the temporal smoothness can be enforced by penalizing the difference between models of consecutive time points. In single task learning formulations, such as Ridge and Lasso, the predicted scores of the same patient at different time points may fluctuate as shown in the left figure.

We use a matrix $S \in \mathbb{R}^{nxt}$ to indicate missing target values, where $S_{i,j} = 0$ if the target value of sample i is missing at the jth time point, and $S_{i,j} = 1$ otherwise. We use the componentwise operator $\odot$ as follows: $Z = A \odot B$ denotes $z_{i,j} = a_{i,j}b_{i,j}$, for all i, j. The formulation in Eq. (3) can be extended to the case with missing target values as:

$$\min_W \|S\odot(XW-Y)\|_F^2 + \theta_1\|W\|_F^2 + \theta_2\|WH\|_F^2. \quad (4)$$

The optimization problem in Eq. (4) can be solved efficiently as shown in Appendix B.

Temporal group Lasso regularization

Because of the limited availability of subjects in the longitudinal AD study and a relatively large number of features (e.g., MRI features) at ADNI, the prediction model suffers from the so called "curse of dimensionality". In addition, many patients drop out from the longitudinal study after a certain period of time, which reduces the effective number of samples. One effective approach is to reduce the dimensionality of the data. However, traditional dimension reduction techniques such as PCA are not desirable since the resulting model is not interpretable, and traditional feature selection algorithms are not suitable for multi-task regression with missing target values. In the proposed formulation, we employ the group Lasso regularization based on the $\ell_{2,1}$-norm penalty for feature selection (Yuan and Lin, 2006), which assumes that a small set of features are predictive of the progression. The group Lasso regularization ensures that all regression models at different time points share a common set of features. Together with the temporal smoothness penalty, we obtain the following Temporal Group Lasso (TGL) formulation:

$$\min_W \|S\odot(XW-Y)\|_F^2 + \theta_1\|W\|_F^2 + \theta_2\|WH\|_F^2 + \delta\|W\|_{2,1} \quad (5)$$

where $\|W\|_{2,1} = \sum_{i=1}^d \sqrt{\sum_{j=1}^t W_{ij}^2}$, and δ is a regularization parameter. When there is only one task, i.e., $t = 1$, the above formulation reduces to Lasso (Tibshirani, 1996). When $t > 1$, the weights of one feature over all tasks are grouped using the ℓ_2-norm, and all features are further grouped using the ℓ_1-norm. Thus, the $\ell_{2,1}$-norm penalty tends to select features based on the strength of the feature over all t tasks.

The objective in Eq. (5) can be considered as a combination of a smooth term and a non-smooth term. The gradient descent or accelerated gradient method (AGM) (Nemirovski, 2005; Nesterov, 2004) can be applied to solve the optimization. One of the key steps in AGM is the computation of the proximal operator associated with the $\ell_{2,1}$-norm regularization. We employ the algorithm in the SLEP package (Liu et al., 2009), which computes the proximal operator associated with the general ℓ_1/ℓ_q-norm efficiently.

Modeling disease progression via fused sparse group Lasso

The TGL formulation constrains the models from all time points to share a common set of features. In order to better capture the temporal patterns of the biomarkers in disease progression (Caroli et al., 2010; Jack et al., 2010), we further propose a convex fused sparse group Lasso (cFSGL) formulation which allows simultaneous joint feature selection for multiple tasks and task-specific feature selection, and in the meantime incorporates the temporal smoothness. Mathematically, the cFSGL formulation solves the following convex optimization problem:

$$\min_W L(W) + \lambda_1\|W\|_1 + \lambda_2\left\|RW^T\right\|_1 + \lambda_3\|W\|_{2,1}, \quad (6)$$

where $\|W\|_1$ is the Lasso penalty, $\|W\|_{2,1} = \sum_{i=1}^d \sqrt{\sum_{j=1}^t W_{ij}^2}$ is the group Lasso penalty, $\|RW^T\|_1$ is the fused Lasso penalty, $R = H^T$ is a $(t-1)\times t$ sparse matrix, and λ_1, λ_2 and λ_3 are regularization parameters. The combination of Lasso and group Lasso penalties is also known as the sparse group Lasso penalty, which allows simultaneous joint feature selection for all tasks and selection of a specific set of features for each task. The fused Lasso penalty is employed to incorporate the temporal smoothness. The cFSGL formulation involves three non-smooth terms, and is thus challenging to solve. We propose to solve the optimization problem by the accelerated gradient method (AGM) (Nemirovski, 2005; Nesterov, 2004). One of the key steps in using AGM is the computation of the proximal operator associated with the composite of non-smooth penalties defined as follows:

$$\pi(V) = \arg\min_W \frac{1}{2}\|W-V\|_2^2 + \lambda_1\|W\|_1 + \lambda_2\left\|RW^T\right\|_1 + \lambda_3\|W\|_{2,1}. \quad (7)$$

It is clear that each row of W is decoupled in Eq. (7). Thus, for obtaining the ith row $\mathbf{w}_i$, we only need to solve the following optimization problem:

$$\pi(\mathbf{v}_i) = \arg\min_{\mathbf{w}_i} \frac{1}{2}\|\mathbf{w}_i-\mathbf{v}_i\|_2^2 + \lambda_1\|\mathbf{w}_i\|_1 + \lambda_2\|R\mathbf{w}_i\|_1 + \lambda_3\|\mathbf{w}_i\|_2, \quad (8)$$

where $\mathbf{v}_i$ is the ith row of V. The proximal operator in Eq. (8) is challenging to solve due to the presence of three non-smooth terms. We show that the proximal operator exhibits a certain decomposition property, based on which we can efficiently compute the proximal operator in two stages, as summarized in Appendix C.

We illustrate the models built by different approaches in Fig. 3. In the left figure we show the model built by Lasso regression. The sparsity introduced by applying Lasso has no specific patterns across tasks, as the models for different tasks are built independently. The middle figure shows the model built by TGL. Because of the use of

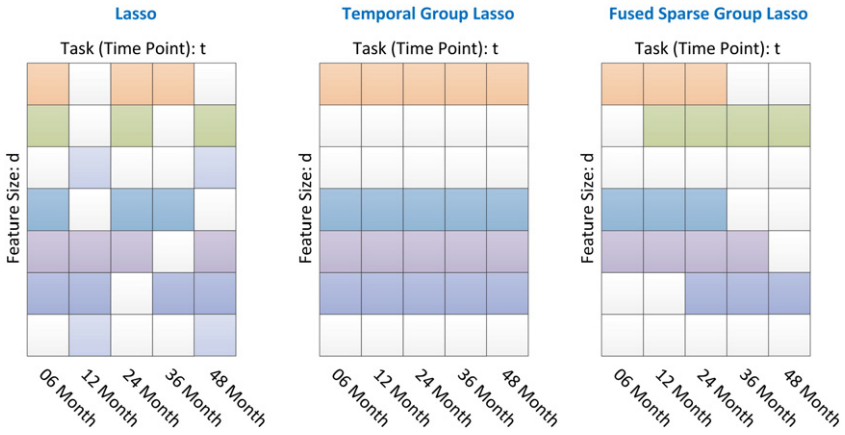

Fig. 3. A comparison of models built by different approaches. In Lasso, the models for different tasks are built independently, thus no specific sparsity patterns are observed across different tasks (left figure). The TGL formulation restricts all models from different time points to select a common set of features (middle figure). In cFSGL, the selected features across different time points are smooth due to the use of the fused Lasso penalty (right figure), that is, the selected features at nearby time points are similar to each other. For the example shown in the right figure, the models at M06 and M12 differ in one feature (the second feature); the models at M12 and M24 differ in one feature (the sixth feature); the models at M24 and M36 differ in two features (the first and fourth features); and the models at M36 and M48 differ in one feature (the fifth feature).

$\ell_{2,1}$-norm regularization to capture temporal relation, the features selected for all time points are the same. The model built by cFSGL, as shown in the right figure, has two levels of sparsity: 1) a small set of features shared across all tasks, 2) task-specific features for each time point. In addition, one key advantage of fused Lasso in cFSGL is that under the fused Lasso penalty the selected features across different time points are similar to each other, satisfying the temporal smoothness property, while the Laplacian-based penalty focuses on the smoothing of the prediction models across different time points.

Longitudinal stability selection for identifying temporal patterns of biomarkers

Stability selection (Meinshausen and Bühlmann, 2010), based on subsampling/bootstrapping, provides a general method to perform model selection using information from a set of regularization parameters. The stability ranking score gives a probability which makes it naturally interpretable. Stability selection has been successfully applied to bioinformatics applications especially in genome-related biomarker selection problems where sample size is much smaller than feature dimension ($n \ll d$) (Eleftherohorinou et al., 2011; Ryali et al., 2012; Stekhoven et al., 2011; Vounou et al., 2012).

We propose to extend the idea of stability selection to longitudinal study. The framework, called *longitudinal stability selection*, is to quantify the importance of the features selected by the proposed formulations for disease progression. Specifically, we apply stability selection to multi-task learning models for longitudinal study. The stability score (between 0 and 1) of each feature is indicative of the importance of the specific feature for disease progression. In this paper, we propose to use longitudinal stability selection with TGL and cFSGL to analyze the temporal patterns of biomarkers. The temporal pattern of stability scores of the features selected at different time points can potentially reveal how disease progresses temporally and spatially.

The longitudinal stability selection algorithm with TGL and cFSGL is given as follows. Let F be the index set of features, and let $f \in F$ denote the index of a particular feature. Let Δ be the regularization parameter space and let the stability iteration number be denoted as γ. For cFSGL an element $\delta \in \Delta$ is a triple $\langle \lambda_1, \lambda_2, \lambda_3 \rangle$. Let $B_{(i)} = \{X_{(i)}, Y_{(i)}\}$ be a random subsample from input data $\{X,Y\}$ of size $\lfloor n/2 \rfloor$ without replacement. For a given $\delta \in \Delta$, let $\hat{W}^{(i)}$ be the optimal solution of TGL or cFSGL on $B_{(i)}$. The set of features selected by the model $\hat{W}^{(i)}$ of the task at time point p is denoted by

$$U_p^{\delta}\left(B_{(i)}\right) = \left\{f : \hat{W}_{f,p}^{(i)} \neq 0\right\}.$$

We repeat this process for γ times and obtain the *selection probability* $\hat{\Pi}_{f,p}^{\delta}$ of each feature f at time point p:

$$\hat{\Pi}_{f,p}^{\delta} = \sum_{i=1}^{\gamma} I\left(f \in U_p^{\delta}\left(B_{(i)}\right)\right)/\gamma,$$

where $I(.)$ is the indicator function defined as: $I(c) = 1$ if c is true and $I(c) = 0$ otherwise. The computation of selection probability is illustrated in Fig. 4. Repeat the above procedure for all $\delta \in \Delta$, we obtain the *stability score* for each feature f at time point p:

$$\mathcal{S}_p(f) = \max_{\delta \in \Delta}\left(\hat{\Pi}_{f,p}^{\delta}\right).$$

The computation of stability score at one time point is illustrated in Fig. 5. The *stability vector* of a feature f at all t time points is given by $\mathcal{S}(f) = [\mathcal{S}_1(f) \ldots \mathcal{S}_t(f)]$, which reveals the change of the importance of feature f at different time points. We define the *stable features* at time point p as:

$$\hat{U}_p = \left\{f : S_p(f) \text{ ranks among top } \eta \text{ in } F\right\} \quad (9)$$

and choose $\eta = 20$ in our experiments. We are interested in the stable features at all time points, i.e., $f \in \hat{U} = \cup_{p=1}^{t} \hat{U}_p$. Note that $\mathcal{S}(f)$ is dependent on the progression model used.

Note that if we use TGL in longitudinal stability selection, we obtain a common list of features for all time points. If we use cFSGL in longitudinal stability selection, the features selected for different time points may differ. However, the selected features at nearby time points are similar to each other. Thus, the distribution of stability scores is expected to exhibit the temporal smoothness property, that is, for each feature the stability score is smooth across different time points.

Results

In this section we perform experimental studies to evaluate the proposed progression models and analyze the biomarkers identified using longitudinal stability selection. In Prediction performance using baseline MRI features section, we compare different modeling approaches for predicting future MMSE and ADAS-Cog scores using baseline MRI images and baseline MMSE. Note that we independently apply the models on MMSE and ADAS-Cog, and we do not assume that these cognitive scores are correlated. We further study the prediction performance with additional demographic and ApoE genotyping information included in the models. In Temporal patterns of MRI biomarkers section,

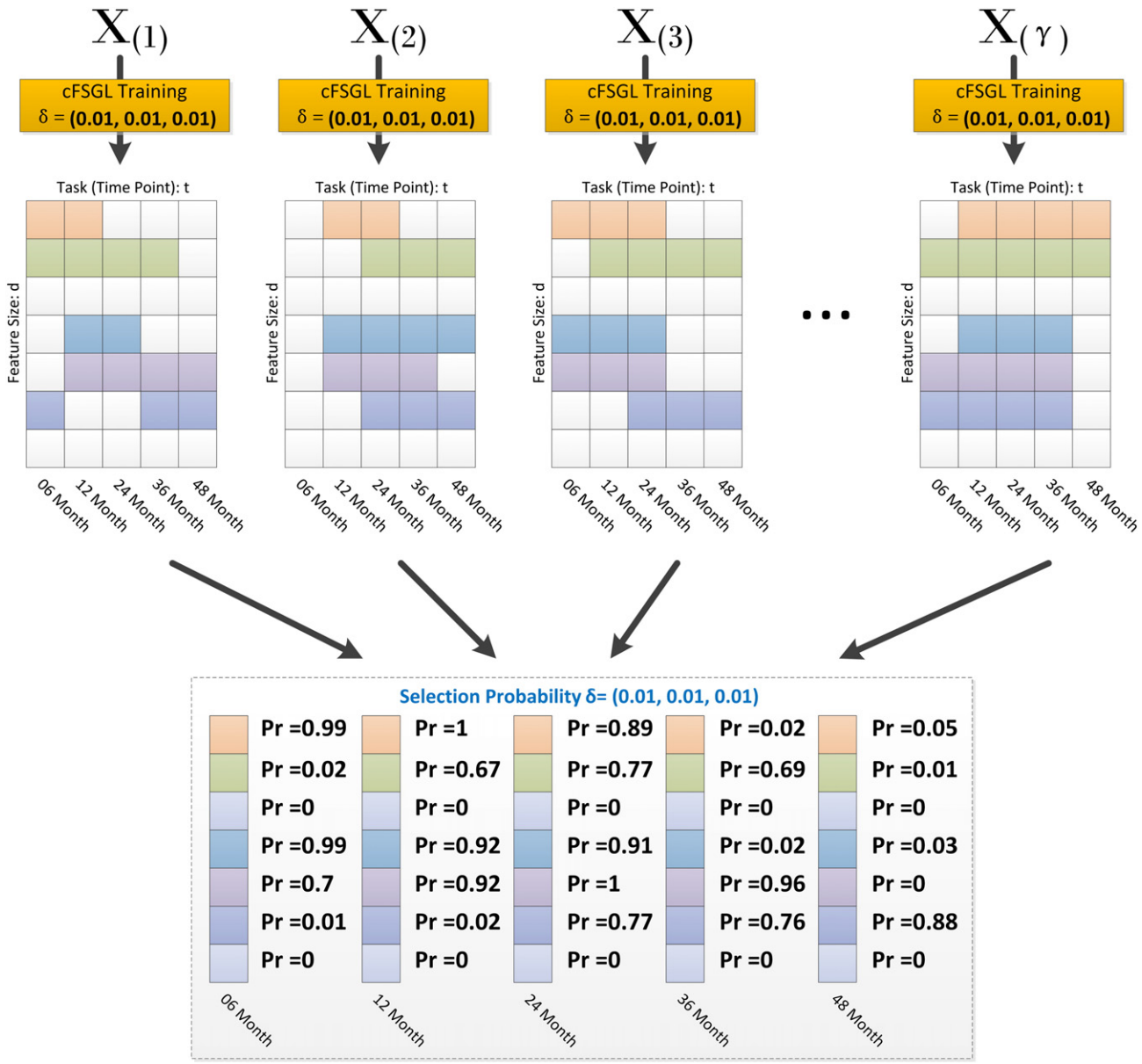

Fig. 4. Illustration of the computation of selection probabilities for all features at all time points in longitudinal stability selection. Given a fixed parameter tuple δ, the selection probabilities are estimated based on a set of γ progression models using γ bootstrapping samples. For each feature, the selection probability at a particular time point is estimated by computing the fraction of γ models at this time point that includes a nonzero coefficient for this feature. The selection probability indicates how likely a feature is selected at one particular time point by the model parameterized by δ.

we analyze the biomarkers identified via longitudinal stability selection. In Predicting the progression for MCI patients section, we study the specific progression model for MCI patients.

Prediction performance using baseline MRI features

In the first experiment, for each target we build a prediction model using baseline MRI features and baseline MMSE. We compare the proposed methods including Temporal Group Lasso (TGL) and Convex Fused Sparse Group Lasso (cFSGL) with single-task learning methods including ridge regression (Ridge) and Lasso regression (Lasso) on the prediction of MMSE and ADAS-Cog. Note that Lasso is a special case of cFSGL when both λ_2 and λ_3 are set to 0. We randomly split the data into training and testing sets using a ratio 9:1, i.e., we build models on 90% of the data and evaluate these models on the remaining 10% of the data. Since there are model parameters to be selected during the training, we use 5-fold cross validation on the training data to select these parameters. For the overall regression performance measures, we use normalized mean square error (nMSE) as used in the multi-task learning literature (Argyriou et al., 2008; Zhang and Yeung, 2010) and weighted correlation coefficient (wR) as employed in the medical literature addressing AD progression problems (Duchesne et al., 2009; Ito et al., 2010; Stonnington et al., 2010). For the task-specific regression performance measures, we use root mean square error (rMSE). The MSE, nMSE and weighted R-value are defined as follows:

$$\mathrm{rMSE}(y,\hat{y}) = \sqrt{\frac{\|y-\hat{y}\|_2^2}{n}}, \tag{10}$$

$$\mathrm{nMSE}\left(Y,\hat{Y}\right) = \frac{\sum_{i=1}^{t}\left\|Y_i-,\hat{Y}_i\right\|_2^2 \Big/ \sigma(Y_i)}{\sum_{i=1}^{t} n_i}, \tag{11}$$

$$\mathrm{wR}\left(Y,\hat{Y}\right) = \frac{\sum_{i=1}^{t}\mathrm{Corr}\left(Y_i,\hat{Y}_i\right)n_i}{\sum_{i=1}^{t} n_i}, \tag{12}$$

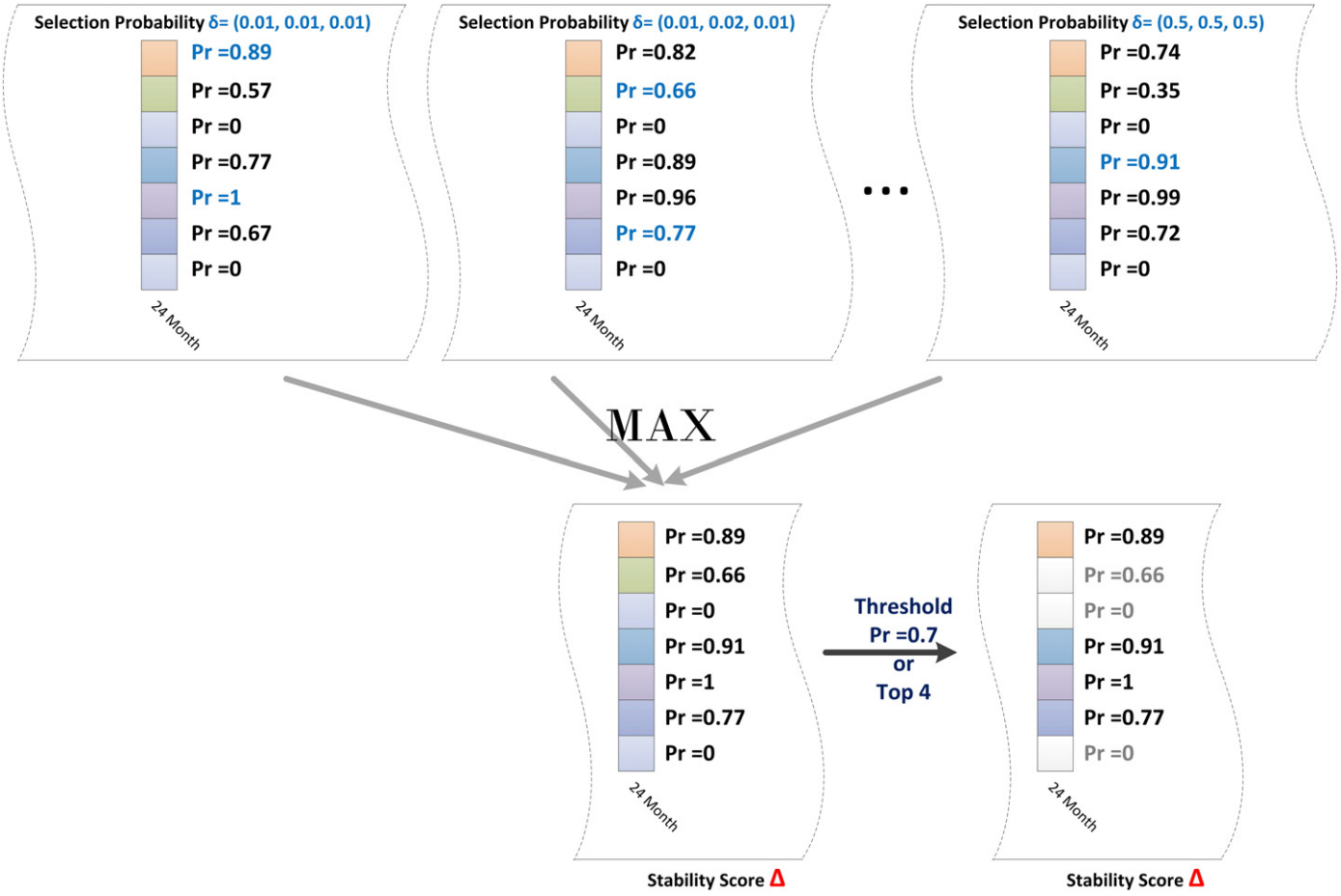

Fig. 5. Illustration of the computation of the stability score in longitudinal stability selection at a particular time point. At each time point, the stability score of a feature is the maximum selection probability it obtains at this time point over all $\delta \in \Delta$. For the example shown in the figure, the maximum selection probability for the first feature is 0.89. After the stability score is computed, we can select features at each time point by either providing a threshold on the selection probabilities or the number of features with top selection probabilities.

where for rMSE, y is the ground truth of target at a single time point and $\hat{y}$ is the corresponding prediction by a prediction model, for nMSE and wR, Y_i is the ground truth of target at time point i, $i = [1{:}t]$ and $\hat{Y}_i$ is the corresponding predicted value, and Corr is the correlation coefficient between two vectors. We report the mean and standard deviation based on 20 iterations of experiments on different splits of data. The experimental results using 90% training data are presented in Table 2.

Overall our proposed approaches outperform Ridge and Lasso, in terms of both nMSE and correlation coefficient. We have the following observations: 1) The proposed multi-task learning models (TGL and cFSGL) outperform single task learning models, which verifies the use of temporal smoothness assumption in our multi-task learning formulations. 2) cFSGL performs better than TGL. This may be due to the restrictive assumption imposed in TGL. 3) The proposed cFSGL formulation witnesses significant improvement for later time points. This may be due to the data sparseness in later time points (see Table 1), as the proposed sparsity-inducing models are expected to achieve better prediction performance in this case.

We also explore the prediction models by including baseline demographic information: age, years of education and ApoE genotyping information, and baseline ADAS-Cog scores of the patients. We follow the same experimental procedure as above. The prediction performance results are shown in Table 3. We see that the performance of predicting the two scores is improved significantly. For example, the weighted correlation coefficient between the predicted value and the true value on testing data has increased from 0.796 to 0.824 ($p < 10e{-}5$) for MMSE prediction and 0.803 to 0.854 ($p < 10e{-}5$) for ADAS-Cog prediction. We also witness the improvement in prediction performance at all time points. We show the scatter plots for the predicted values versus the actual values for MMSE and ADAS-Cog on the testing data in Figs. 6 and 7, respectively. Since there are few samples available at the last time point (M48), we only show the scatter plots for the first four time points. In

Table 2
Comparison of our proposed approaches (TGL and cFSGL) and single-task learning approaches (Ridge, Lasso) on longitudinal MMSE and ADAS-Cog prediction using MRI features (M) in terms of normalized mean square error (nMSE), weighted correlation coefficient (wR) and root mean square error (rMSE) at each time point. 90% of data is used as training data.

	Ridge	Lasso	TGL	cFSGL
Target: MMSE				
nMSE	0.548 ± 0.057	0.459 ± 0.042	0.449 ± 0.045	**0.395 ± 0.052**
wR	0.689 ± 0.030	0.746 ± 0.031	0.755 ± 0.029	**0.796 ± 0.031**
M06 rMSE	2.269 ± 0.207	2.071 ± 0.261	**2.038 ± 0.262**	2.071 ± 0.213
M12 rMSE	3.266 ± 0.556	2.973 ± 0.654	2.923 ± 0.643	**2.762 ± 0.669**
M24 rMSE	3.494 ± 0.599	3.371 ± 0.747	3.363 ± 0.733	**3.000 ± 0.642**
M36 rMSE	4.003 ± 0.853	3.786 ± 0.926	3.768 ± 0.962	**3.265 ± 0.803**
M48 rMSE	4.328 ± 1.310	3.653 ± 1.268	3.631 ± 1.226	**2.871 ± 0.884**
Target: ADAS-Cog				
nMSE	0.532 ± 0.095	0.520 ± 0.084	0.464 ± 0.067	**0.391 ± 0.059**
wR	0.705 ± 0.043	0.716 ± 0.036	0.747 ± 0.033	**0.803 ± 0.024**
M06 rMSE	5.213 ± 0.522	4.976 ± 0.518	4.820 ± 0.489	**4.451 ± 0.340**
M12 rMSE	6.079 ± 0.775	6.193 ± 0.766	5.813 ± 0.697	**5.230 ± 0.589**
M24 rMSE	7.409 ± 1.154	7.275 ± 1.099	6.835 ± 1.052	**6.249 ± 0.996**
M36 rMSE	7.143 ± 1.351	7.139 ± 1.444	6.938 ± 1.363	**5.928 ± 1.064**
M48 rMSE	6.644 ± 2.750	6.879 ± 2.465	6.000 ± 2.738	**5.980 ± 1.979**

Table 3
Comparison of our proposed approaches (TGL and cFSGL) and single-task learning approaches (Ridge, Lasso) on longitudinal MMSE and ADAS-Cog prediction using MRI, demographic, and ApoE genotyping features in terms of normalized mean square error (nMSE), weighted correlation coefficient (wR) and root mean square error (rMSE) at each time point. 90% of data is used as training data.

	Ridge	Lasso	TGL	cFSGL
Target: MMSE				
nMSE	0.477 ± 0.055	0.368 ± 0.048	0.364 ± 0.046	**0.341 ± 0.039**
wR	0.743 ± 0.022	0.809 ± 0.026	0.811 ± 0.027	**0.824 ± 0.021**
M06 rMSE	2.211 ± 0.241	1.938 ± 0.214	**1.900 ± 0.211**	1.980 ± 0.219
M12 rMSE	2.968 ± 0.685	2.679 ± 0.769	2.654 ± 0.767	**2.546 ± 0.748**
M24 rMSE	3.454 ± 0.550	3.107 ± 0.570	3.133 ± 0.579	**2.943 ± 0.582**
M36 rMSE	3.736 ± 0.792	3.311 ± 0.756	3.313 ± 0.798	**3.046 ± 0.701**
M48 rMSE	3.469 ± 1.030	2.645 ± 0.845	2.761 ± 0.883	**2.364 ± 0.792**
Target: ADAS-Cog				
nMSE	0.396 ± 0.075	0.335 ± 0.048	0.317 ± 0.044	**0.296 ± 0.048**
wR	0.791 ± 0.031	0.830 ± 0.020	0.837 ± 0.017	**0.854 ± 0.021**
M06 rMSE	4.384 ± 0.522	3.936 ± 0.430	3.858 ± 0.441	**3.863 ± 0.516**
M12 rMSE	4.906 ± 0.708	4.578 ± 0.756	4.455 ± 0.661	**4.209 ± 0.564**
M24 rMSE	6.587 ± 1.038	6.153 ± 1.145	5.945 ± 1.120	5.657 ± 1.017
M36 rMSE	6.312 ± 1.068	5.849 ± 1.028	5.613 ± 0.936	5.066 ± 0.854
M48 rMSE	5.679 ± 2.200	5.087 ± 2.082	5.181 ± 2.383	**5.182 ± 1.606**

the scatter plots, we see that the predicted values and actual clinical scores have a high correlation. The scatter plots show that the prediction performance for ADAS-Cog is better than that of MMSE.

In the study of ADNI, cognitive normal individuals and stable MCI patients are less likely to have significant changes on the cognitive scores and therefore many existing studies focus on subgroups of patients only (e.g., Duchesne et al., 2009). To this end, we apply our models on the subgroup that consists of MCI converters and AD patients only (see Table 1). At the last time point M48, there are only very few samples available and we therefore exclude the last time point from our study. We follow the same experimental setting as in the previous experiment, and the results are shown in Table 4. We observe that cFSGL achieves the best performance among all methods, with an average performance of $R = 0.671 (p < 10e-5)$ in predicting longitudinal MMSE scores and an average of $R = 0.751 (p < 10e-5)$ in predicting ADAS-Cog.

Temporal patterns of MRI biomarkers

One of the strengths of the cFSGL formulation is that it facilitates the identification of temporal patterns of biomarkers. In this experiment we study the temporal patterns of biomarkers using longitudinal stability selection. Note that because the sample size at the M48 time point is small, we perform longitudinal stability selection for M06, M12, M24, and M36 only. In all cases, the baseline MMSE score is the most important predictor and has a selection probability of 1 at all time points and we therefore do not show it in the figures.

The stability vectors using the cFSGL formulation are given in Fig. 8, where we collectively list the stable features ($\eta = 20$) at the

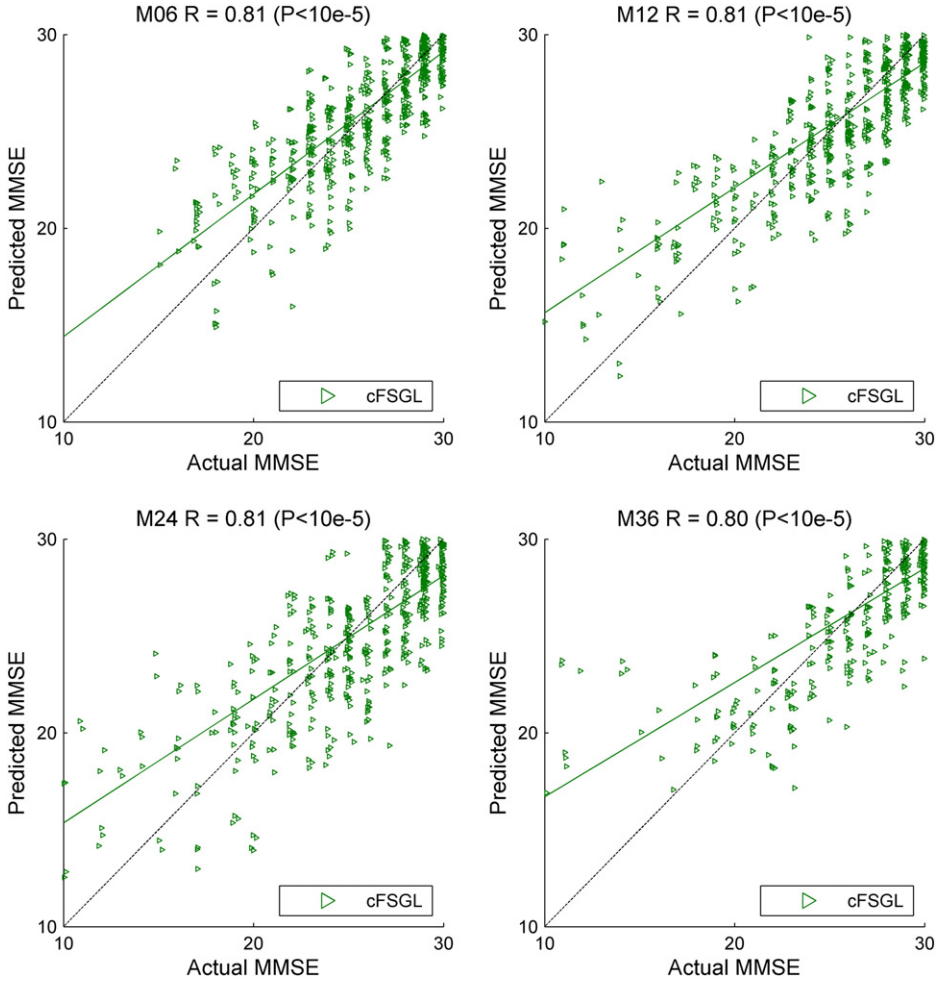

Fig. 6. Scatter plots of actual MMSE versus predicted values on testing data using cFSGL based on baseline MRI features, demographic, and ApoE genotyping features. The black dashed line in each figure is a reference of perfect correlation (predicted value exactly equals to actual value). We perform least squares regression on the points shown in the scatter plots and the green solid line is the regression line, which serves as a visual indicator of overall performance. The closer between the regression line and the reference line, the better are the prediction results. We see that the patients with low actual MMSE scores are less predictable, compared to the ones with high actual MMSE scores.

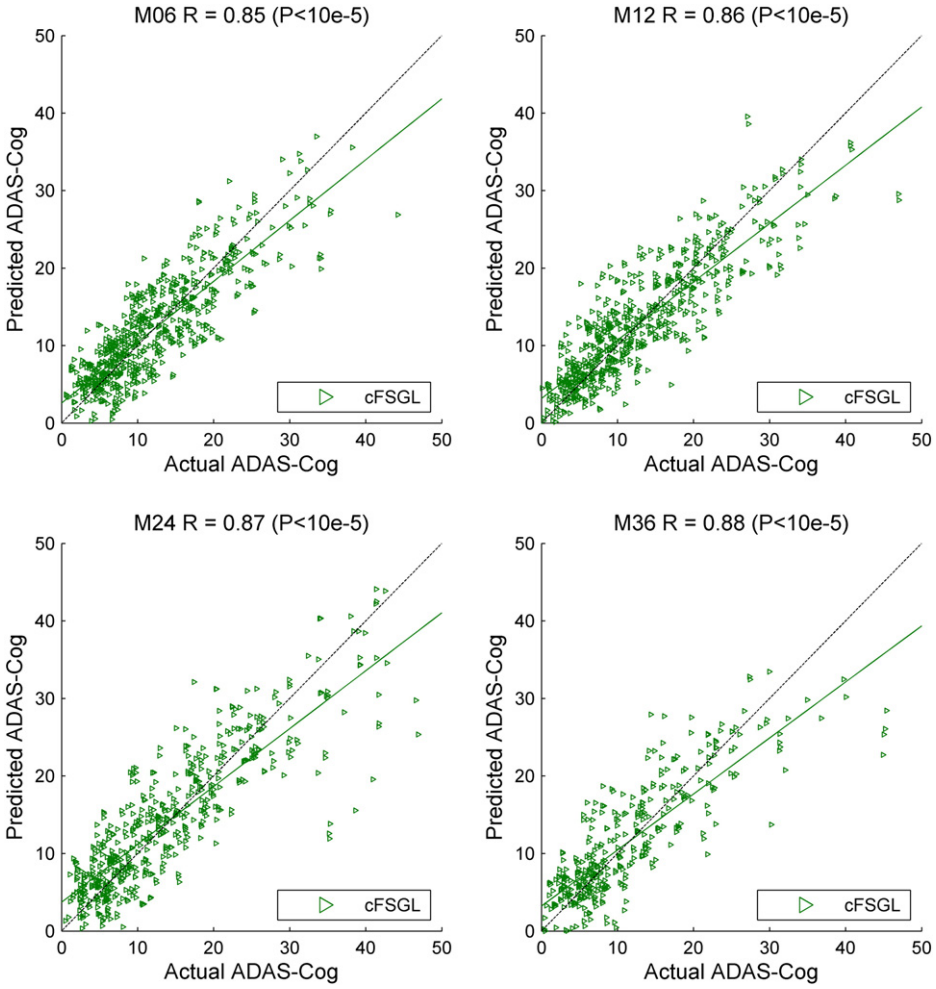

Fig. 7. Scatter plots of actual ADAS-Cog versus predicted values on testing data using cFSGL based on baseline MRI features, demographic, and ApoE genotyping features. The black dashed line in each figure is a reference of perfect correlation (predicted value exactly equals to actual value). We perform least squares regression on the points shown in the scatter plots and the green solid line is the regression line, which serves as a visual indicator of overall performance. The closer between the regression line and the reference line, the better are the prediction results. We see a high correlation between the two values. The visual prediction performance for ADAS-Cog is better than that of MMSE as shown in Fig. 6.

4 time points. The total number of features is less than 80 because one feature may be identified as stable features at multiple time points. In Fig. 8(a), we observe that cortical thickness average of left middle temporal, cortical thickness average of left and right Entorhinal, and white matter volume of left Hippocampus are important biomarkers for all time points. Cortical volume of left Entorhinal provides significant information in later stages than in the first 6 months. Several biomarkers including white matter volume of left and right Amygdala, and surface area of right posterior banks of the superior temporal sulcus (Bankssts) provide useful information only in later time points. On the contrary, some biomarkers have a large stability score during the first 2 years after baseline screening, such as cortical thickness average of left inferior temporal, left inferior parietal, and cortical thickness standard deviation of left isthmus cingulate, right lingual, left inferior parietal, and cortical volume of right precentral, right isthmus cingulate, and left middle temporal cortex.

Table 4
Comparison of our proposed approaches (TGL and cFSGL) and single-task learning approaches (Ridge, Lasso) on longitudinal MMSE and ADAS-Cog prediction for MCI converters and AD patients using MRI, demographic, and ApoE genotyping features in terms of normalized mean square error (nMSE), weighted correlation coefficient (wR) and root mean square error (rMSE) at each time point. 80% of data is used as training data.

	Ridge	Lasso	TGL	cFSGL
Target: MMSE				
nMSE	1.161 ± 0.269	0.860 ± 0.137	0.761 ± 0.143	**0.725 ± 0.128**
wR	0.526 ± 0.080	0.633 ± 0.068	0.660 ± 0.059	**0.671 ± 0.054**
M06 rMSE	3.420 ± 0.381	3.031 ± 0.280	2.881 ± 0.245	**2.862 ± 0.231**
M12 rMSE	4.025 ± 0.482	3.680 ± 0.531	3.391 ± 0.489	**3.315 ± 0.506**
M24 rMSE	5.531 ± 0.756	4.988 ± 0.924	4.636 ± 0.883	**4.551 ± 0.870**
M36 rMSE	5.971 ± 1.214	5.011 ± 1.231	4.686 ± 1.077	**4.422 ± 1.046**
Target: ADAS-Cog				
nMSE	1.031 ± 0.200	0.748 ± 0.078	0.675 ± 0.079	**0.533 ± 0.101**
wR	0.569 ± 0.059	0.695 ± 0.045	0.704 ± 0.042	**0.751 ± 0.046**
M06 rMSE	6.256 ± 0.813	5.692 ± 0.591	5.381 ± 0.583	**5.140 ± 0.800**
M12 rMSE	7.320 ± 0.988	6.334 ± 1.022	5.934 ± 0.884	**5.196 ± 0.829**
M24 rMSE	10.423 ± 1.224	9.353 ± 1.301	8.964 ± 1.331	**7.486 ± 1.249**
M36 rMSE	10.968 ± 1.833	9.319 ± 2.082	8.782 ± 1.801	**6.958 ± 1.499**

The stability vectors of stable MRI features for MMSE are shown in Fig. 8(b). We obtain very different patterns from ADAS-Cog. We find that most biomarkers provide significant information for the first 2 years and very few of them are significant in later stages. The lack of predictable MRI biomarkers in later stages is a potential factor that contributes to the lower prediction performance of MMSE than that of ADAS-Cog in our study and other related studies (Stonnington et al., 2010; Zhang and Shen, 2011). The different temporal patterns of biomarkers for these two scores also suggest that restricting the two models for predicting these two scores to share a common set of features as done in previous work may lead to sub-optimal performance.

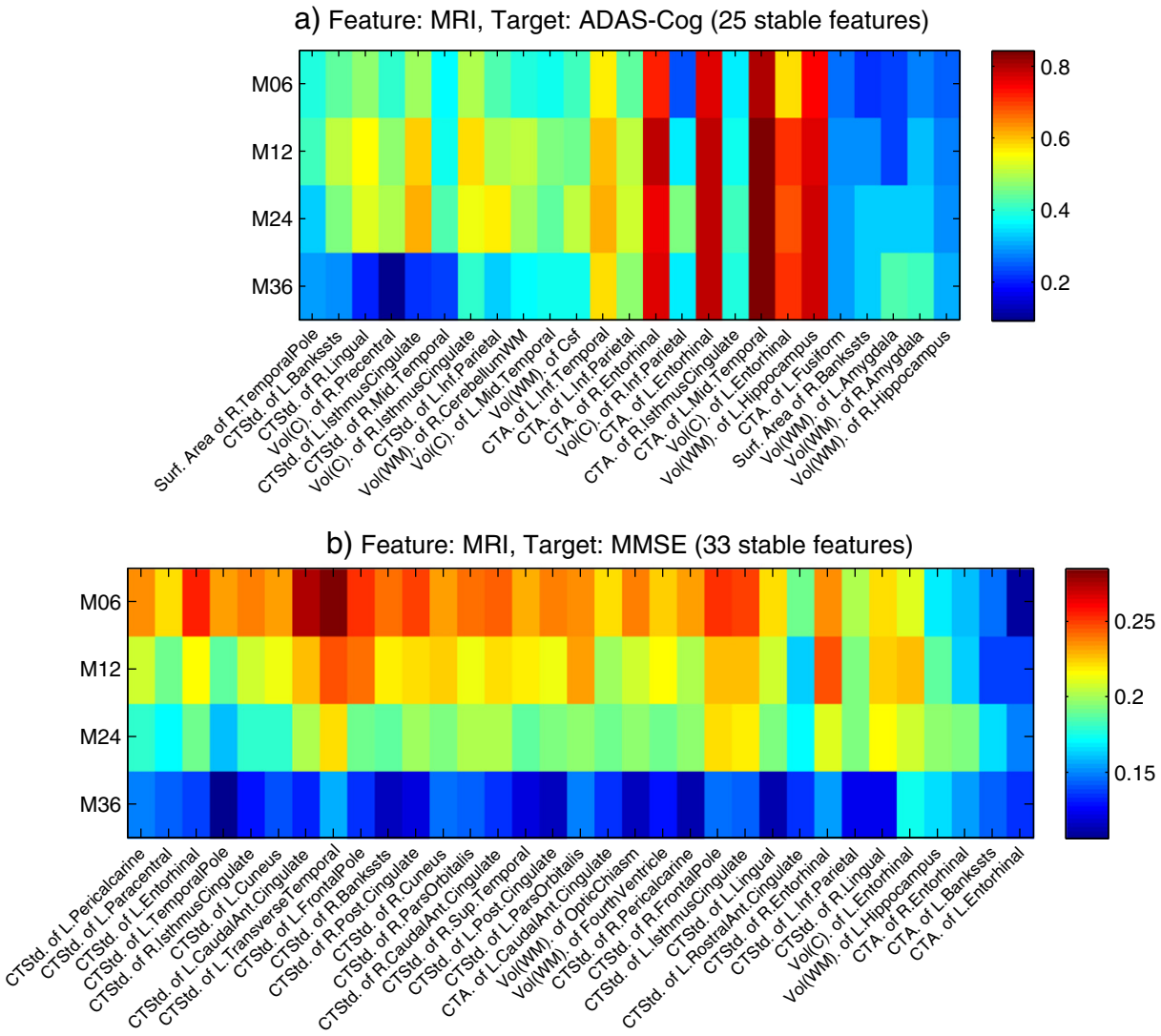

Fig. 8. The stability vector of stable MRI features using cFSGL. In (a) we see that cortical thickness average of left middle temporal, cortical thickness average of left and right Entorhinal, and white matter volume of left Hippocampus are important biomarkers for predicting ADAS-Cog scores at all time points. Cortical volume of left Entorhinal provides significant information in later stages than in the first 6 months. Several biomarkers including white matter volume of left and right Amygdala, and surface area of right Bankssts provide useful information only in later time points. In (b) we obtain very different patterns for MMSE. We find that most biomarkers provide significant information for predicting MMSE scores for the first 2 years and very few of them are significant in later stages.

Predicting the progression for MCI patients

In the study of Alzheimer's disease, the MCI patients are of particular interest. In this section we design experiments to study the prediction models for MCI patients. We study the prediction performance on MCI patients using 1) only MCI patients in the training data; and 2) MCI patients together with AD patients and normal controls in the training data.

In the first experiment we use only MCI patients in both training and testing data. We random split the MCI patients, with 90% as training data and 10% as testing data. We build prediction models using the MCI training data and test the models using MCI testing data. We permute the samples and compute the performance based on 20 partitions of the data (into training and testing sets). For other experimental settings, we follow the same practice as in our previous experiments (Prediction performance using baseline MRI features section). We use baseline MRI, demographic, and ApoE genotyping features. In the second experiment, for each partition we use the same MCI patients as in the first experiment, and further include all AD and NL samples in the training data when building the model. We use the same MCI testing data for evaluating the prediction performance. In this setting, the testing samples are the same for the two experiments and our experiments will reveal the effect of including AD and NL samples in the model building step. The performance of predicting MMSE and ADAS-Cog at all time points is given in Figs. 9 and 10, respectively.

We can observe from the figures that in most cases the prediction performance with AD and NL samples included witnesses improvement. For MMSE, the overall prediction performance in terms of nMSE improves from 0.680 ± 0.189 to 0.567 ± 0.125, and the improvement for ADAS-Cog is from 0.552 ± 0.118 to 0.544 ± 0.092. Such improvement is especially significant for sparse-learning methods (Lasso, TGL, cFSGL) at later time points. A possible explanation is that, for later time points (M36, M48), the sample size is significantly smaller, and the information from AD and NL subjects at the specific time point is useful for the prediction of MCI subjects, giving rise to performance

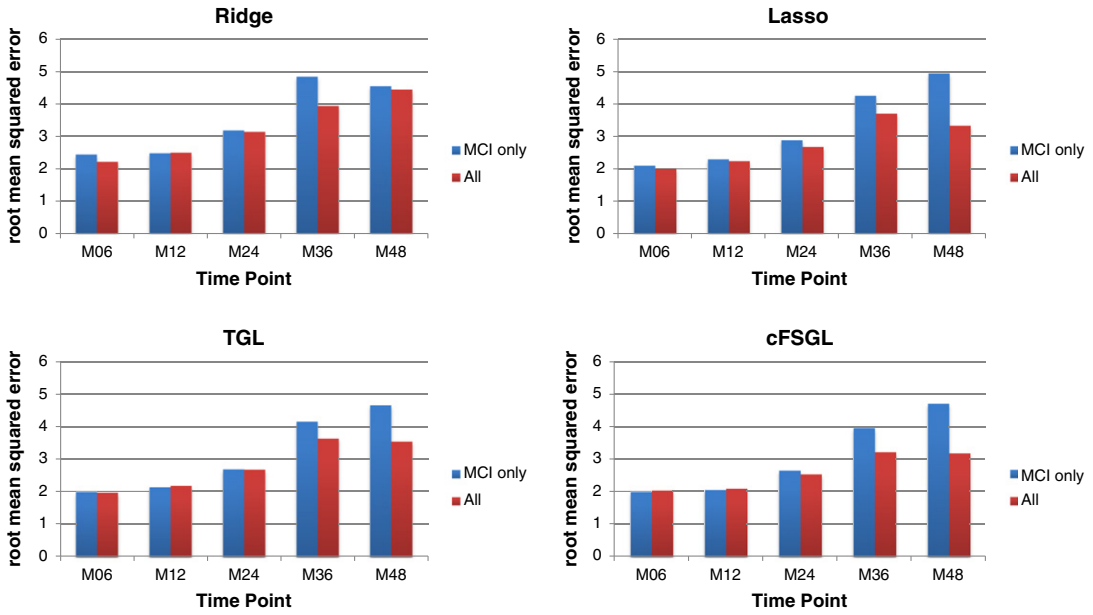

Fig. 9. Comparison of MMSE prediction models in terms of root mean square error (rMSE) on patients using only MCI patients in training (MCI only), and using MCI patients together with AD patients and normal controls (All). Lower rMSE indicates better performance. We see that in most cases the prediction performance with AD and NL samples included witnesses improvement. Such improvement is especially significant for sparse-learning methods (Lasso, TGL, cFSGL) at later time points. This may be due to the small sample size at later time points, in which the information from AD and NL subjects may be useful during the learning.

improvement. Note that for the first three time points, it does not provide much benefit to include AD and NL subjects.

Predicting progression using features from multiple time points

The goal of our predictive modeling is to predict the cognitive scores at future time points based on the information currently available. In the previous experiments we build models based on the baseline imaging information only. In the longitudinal study, the patients will be followed up for a period of time, and thus data from multiple time points may be available for building predictive models. To this end, we build models using baseline and M06 MRI features and predict the MMSE and ADAS-Cog scores at M12, M24, M36, and M48. We use the same experimental settings as in the previous experiments and report the results in Table 5. We observe improved predictive performance at these time points, as compared to the models obtained using baseline MRI features only (Table 2). When data from multiple time points are available, advanced structured sparsity techniques can be used to leverage the temporal information among features, e.g., (Wang et al., 2012; Zhang and Shen, 2012).

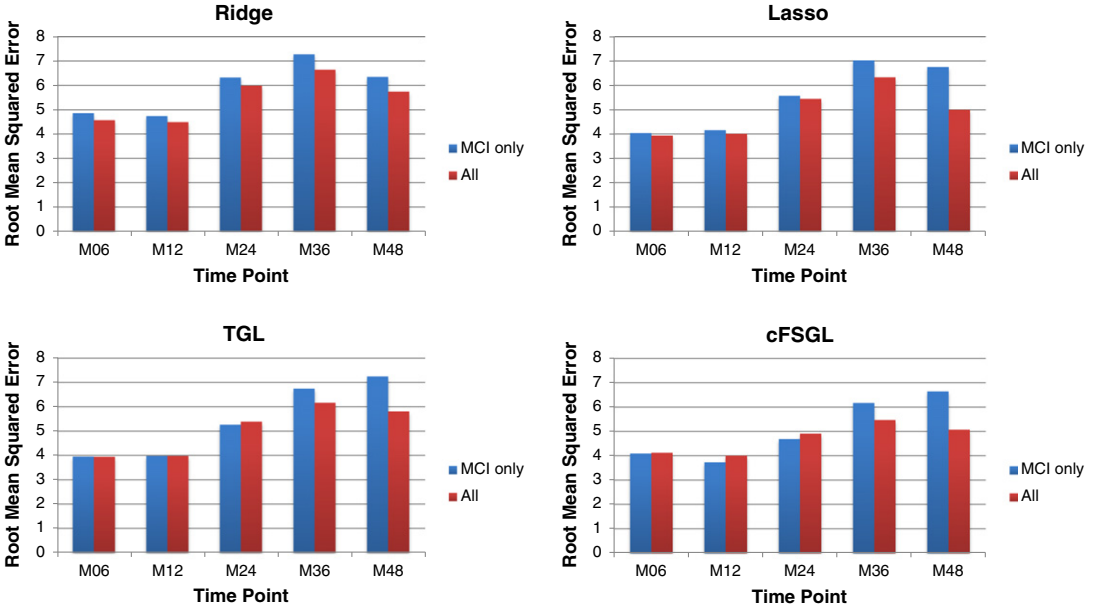

Fig. 10. Comparison of ADAS-Cog prediction models in terms of root mean square error (rMSE) on patients using only MCI patients in training (MCI only), and using MCI patients together with AD patients and normal controls (All). The results witness similar patterns as in MMSE prediction as shown in Fig. 9. The prediction performance with AD and NL samples included has improved, especially significant for sparse-learning methods (Lasso, TGL, cFSGL) at later time points.

Table 5
Comparison of our proposed approaches (TGL and cFSGL) and single-task learning approaches (Ridge, Lasso) on longitudinal MMSE and ADAS-Cog prediction using baseline and M06 MRI features in terms of normalized mean square error (nMSE), weighted correlation coefficient (wR) and root mean square error (rMSE) at each time point. 90% of data is used as training data.

	Ridge	Lasso	TGL	cFSGL
Target: MMSE				
nMSE	0.493 ± 0.083	0.368 ± 0.049	0.365 ± 0.048	**0.326 ± 0.054**
wR	0.744 ± 0.041	0.806 ± 0.038	0.813 ± 0.035	**0.836 ± 0.030**
M12 rMSE	2.458 ± 0.269	2.134 ± 0.294	**2.105 ± 0.294**	2.132 ± 0.280
M24 rMSE	3.249 ± 0.551	2.935 ± 0.571	2.920 ± 0.569	**2.677 ± 0.558**
M36 rMSE	3.682 ± 0.560	3.398 ± 0.609	3.352 ± 0.598	**3.067 ± 0.526**
M48 rMSE	3.962 ± 1.472	2.823 ± 1.346	3.084 ± 1.226	**2.428 ± 1.154**
Target: ADAS-Cog				
nMSE	0.535 ± 0.081	0.417 ± 0.047	0.391 ± 0.046	**0.345 ± 0.061**
wR	0.720 ± 0.029	0.778 ± 0.029	0.795 ± 0.028	**0.834 ± 0.024**
M12 rMSE	5.384 ± 0.588	4.933 ± 0.789	4.722 ± 0.684	**4.611 ± 0.546**
M24 rMSE	6.927 ± 0.692	6.428 ± 0.787	6.150 ± 0.748	**5.341 ± 0.672**
M36 rMSE	7.562 ± 1.072	6.508 ± 1.037	6.246 ± 1.072	**5.389 ± 0.887**
M48 rMSE	6.091 ± 2.003	**4.735 ± 1.633**	5.081 ± 1.758	5.213 ± 0.939

Table 6
Comparison of our proposed approaches (TGL and cFSGL) and single-task learning approaches (Ridge, Lasso) on longitudinal MMSE and ADAS-Cog prediction for all patients using only baseline MRI features in terms of normalized mean square error (nMSE), weighted correlation coefficient (wR) and root mean square error (rMSE) at each time point. 90% of data is used as training data.

	Ridge	Lasso	TGL	cFSGL
Target: MMSE				
nMSE	0.750 ± 0.089	0.691 ± 0.064	0.641 ± 0.055	**0.536 ± 0.061**
wR	0.554 ± 0.038	0.573 ± 0.049	0.616 ± 0.044	**0.711 ± 0.033**
M06 rMSE	2.746 ± 0.192	2.620 ± 0.229	2.538 ± 0.217	**2.486 ± 0.202**
M12 rMSE	3.812 ± 0.533	3.605 ± 0.640	3.478 ± 0.636	**3.184 ± 0.662**
M24 rMSE	4.094 ± 0.510	4.095 ± 0.644	3.978 ± 0.650	**3.423 ± 0.561**
M36 rMSE	4.539 ± 0.917	4.544 ± 1.051	4.355 ± 1.098	**3.732 ± 0.928**
M48 rMSE	4.526 ± 1.375	4.342 ± 1.378	4.068 ± 1.411	**3.380 ± 1.129**
Target: ADAS-Cog				
nMSE	0.675 ± 0.109	0.669 ± 0.094	0.605 ± 0.076	**0.468 ± 0.076**
wR	0.606 ± 0.049	0.599 ± 0.040	0.646 ± 0.036	**0.751 ± 0.030**
M06 rMSE	6.020 ± 0.620	5.967 ± 0.741	5.698 ± 0.650	**5.002 ± 0.352**
M12 rMSE	6.770 ± 0.892	6.936 ± 0.931	6.609 ± 0.882	**5.678 ± 0.724**
M24 rMSE	8.454 ± 1.372	8.361 ± 1.302	8.010 ± 1.349	**7.050 ± 1.201**
M36 rMSE	8.165 ± 1.433	8.105 ± 1.576	7.772 ± 1.537	**6.646 ± 1.148**
M48 rMSE	6.692 ± 2.830	6.709 ± 2.774	6.203 ± 2.713	**5.368 ± 2.022**

Discussion

This paper has three major contributions. First, we formulated the disease progression prediction as a multi-task learning problem. Second, we proposed two multi-task learning formulations that make use of the intrinsic temporal relationship among tasks. In our experiments on ADNI dataset, the cFSGL formulation significantly improved the prediction performance, compared to other methods. Third, we proposed longitudinal stability selection to analyze the dynamic patterns of biomarkers using our proposed formulations.

Many existing works analyzed the relationship between cognitive scores and imaging markers based on MRI such as gray matter volumes, density and loss (Apostolova et al., 2006; Chetelat and Baron, 2003; Frisoni et al., 2002, 2010; Stonnington et al., 2010), shape of ventricles (Ferrarini et al., 2008; Thompson et al., 2004) and hippocampal (Thompson et al., 2004) by correlating these features with baseline MMSE scores. Notably, the above studies related the biomarkers to only one time point. In our work we simultaneously consider the prediction models at multiple time points in order to make use of the temporal information in the longitudinal study. In studying the longitudinal progression of Alzheimer's disease, researchers developed alternative measurements for studying disease progression. An example is the pre-progression rate (Doody et al., 2001), which estimates the rate of change of cognitive status evaluated by the reduction of clinical scores. However, the computation of the pre-progression requires the estimated duration of symptoms in years. This estimation may be inaccurate due to the fact that the disease may be on set earlier than when any symptom begins (Jack et al., 2010).

To predict the longitudinal response to Alzheimer's disease progression, Ashford and Schmitt built a model with horologic function using "time-index" to measure the rate of dementia progression (Ashford and Schmitt, 2001). Doody et. al. used pre-progression rate to assign the 597 AD patients in their study into three groups (slow, intermediate, and fast) and used the data to fit the mixed-effect models (Doody et al., 2010).

Ito et al. proposed to model the progression rate of cognitive scores using power functions and used 817 patients from ADNI to fit the their model (Ito et al., 2010). In using power functions, due to the model complexity, the number of features to be included in the model is limited. Including MRI features as done in our study, for example, is prohibited.

In the longitudinal stability selection, we observe that the volume of left hippocampus, cortical thickness average of middle temporal gyri and cortical thickness average of left and right entorhinal are among the most stable features for both MMSE and ADAS-Cog scores. These findings agree with the known knowledge that in the pathological pathway of AD, medial temporal lobe (hippocampus and entorhinal cortices) is firstly affected, followed by progressive neocortical damage (Braak and Braak, 1991; Delacourte et al., 1999). Evidence of a significant atrophy of middle temporal gyri in AD patients has also been observed in previous studies (Apostolova et al., 2006; Convit et al., 2000; Julkunen et al., 2009). Besides hippocampus and middle temporal, we also find isthmus cingulate a very stable feature for MMSE. The atrophy of isthmus cingulate is considered high in AD patients (McEvoy et al., 2009). In addition, cortical thickness average of left inferior parietal and volume of right inferior parietal are also found to be stable. This agrees with evidence from the previous study that includes pathological confirmation of the diagnosis (Likeman et al., 2005), which shows that parietal atrophy contributes to predictive values for diagnosing AD. Both ADAS-Cog and MMSE are global cognitive scores that are used to evaluate the general cognitive status, and however based on the stable biomarkers found in this study, they have very different temporal patterns, which may have caused the difference in longitudinal predictability using the same set of features included in this study.

In the study of prediction models for MCI patients, our experimental results show that including AD and NL samples in training improves the performance especially at later time points. In the situation where MCI training samples are very limited, information from AD and NL samples can be used to improve the performance of MCI prediction. This result has implications for MCI targeted studies (e.g., Julkunen et al., 2009; Misra et al., 2009; Spulber et al., 2010) in that some progression information from other populations other than MCI patients can potentially benefit the studies. However, we also find that for the first three time points there is a slight performance decrease especially for the ADAS-Cog prediction (see Fig. 10) after including AD and NL samples. This suggests that, when it comes to transfer information from other population for modeling progression, simply treating all time points in the same manner may be suboptimal, which is also a common disadvantage of symmetric multi-task learning formulations. Developing disease progression models via asymmetric multi-task learning (e.g., Xue et al., 2007) may mitigate the problem. We plan to explore this in our future work.

In Duchesne et al. (2009) and Stonnington et al. (2010), no baseline MMSE and/or ADAS-Cog information is included in the model. To compare with the approaches in Duchesne et al. (2009) and Stonnington et al. (2010), we build predictive models using baseline MRI features only. We use the same experimental settings as in the previous experiments and report the results in Table 6. We can observe from the table that the performance is not as good as the one with the baseline scores included in the model (see Table 2). We also list the predictive performance reported in several existing

Table 7
Comparison of the proposed approach with related works in the literature. AD, MCI, and NL refer to Alzheimer's disease patients, mild cognitive impairment patients, and normal controls, respectively.

Method	Target	Subjects	Feature	Result (correlation)
Duchesne et al. (2009)	M12 MMSE	75 NL, 49 MCI, 75 AD	Baseline MRI, age, gender, years of education	MMSE: 0.31 ($p = 0.03$)
Stonnington et al. (2010)	Baseline MMSE and ADAS-Cog	Set1: 73 AD, 91 NL Set2 (ADNI): 113 AD, 351 MCI, 122 NL	Baseline MRI, CSF	MMSE: Set1: 0.7 ($p < 10e-5$) Set2: 0.48 ($p < 10e-5$) ADAS-Cog: Set2: 0.57 ($p < 10e-5$)
cFSGL	M06–M36 MMSE and ADAS-Cog	ADNI: 133 AD, 304 MCI, 188 NL	Baseline MRI	Avg MMSE: 0.711 ($p < 10e-5$) Avg ADAS-Cog: 0.751 ($p < 10e-5$)
cFSGL	M06–M36 MMSE and ADAS-Cog	ADNI: 133 AD, 304 MCI, 188 NL	Baseline MRI, age, ApoE4, baseline MMSE, baseline ADAS-Cog, years of education	Avg MMSE: 0.824 ($p < 10e-5$) Avg ADAS-Cog: 0.854 ($p < 10e-5$)

studies in Table 7. We find that the proposed cFSGL achieves better predictive performance than existing methods.

In the prediction of targets such as cognitive scores, one important issue is the ceiling and/or flooring effect, as the valid target lies in a closed interval. For example, the full score of MMSE is 30 and the lowest possible MMSE score is 0. The ADAS-Cog lies in the interval [0,70]. For the prediction of such targets, we can apply the idea of 'censored regression'. One of the most popular methods in censored regression is the Tobit model (Amemiya, 1973, 2010). The consistency and interpretation of the Tobit model have been well studied (Amemiya, 1973, 2010) and the Tobit model has also been used widely in many areas including economics, statistics, and bioinformatics (Epstein et al., 2003; Frone et al., 1994; McBee, 2010; McDonald and Moffitt, 1980; Ravona-Springer et al., 2012). We incorporate the Tobit model in the proposed formulation and perform experiments to compare the models with and without Tobit censoring in predicting MMSE and ADAS-Cog. The experimental setting is the same as in the previous experiments and the results are shown in Table 8. We find that the Tobit model improves the performance for both MMSE and ADAS-Cog predictions, however the improvement is minor in all cases.

Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH Grants P30 AG010129, K01 AG030514, and the Dana Foundation.

This work was funded by the US National Science Foundation (NSF) (IIS-0812551, IIS-0953662, MCB-1026710, CCF-1025177 to JY), and the National Library of Medicine (R01 LM010730 to JY).

Table 8
Comparison of our proposed cFSGL approach with and without Tobit censoring in terms of normalized mean square error (nMSE), weighted correlation coefficient (wR) and root mean square error (rMSE) at each time point. 90% of data is used as training data.

	No censoring	Tobit censoring
Target: MMSE		
nMSE	0.395 ± 0.052	**0.385 ± 0.051**
wR	0.796 ± 0.031	**0.801 ± 0.032**
M06 rMSE	2.071 ± 0.213	**2.030 ± 0.214**
M12 rMSE	2.762 ± 0.669	**2.737 ± 0.675**
M24 rMSE	3.0000.642	**2.983 ± 0.643**
M36 rMSE	3.265 ± 0.803	**3.210 ± 0.822**
M48 rMSE	2.871 ± 0.884	**2.858 ± 0.886**
Target: ADAS-Cog		
nMSE	0.391 ± 0.059	**0.390 ± 0.060**
wR	0.802 ± 0.024	**0.803 ± 0.024**
M06 rMSE	4.451 ± 0.340	**4.446 ± 0.339**
M12 rMSE	5.230 ± 0.589	**5.216 ± 0.602**
M24 rMSE	6.249 ± 0.996	**6.247 ± 0.997**
M36 rMSE	5.928 ± 1.064	**5.919 ± 1.068**
M48 rMSE	5.980 ± 1.979	**5.978 ± 1.985**

Conflict of interest

There is no conflict of interest.

Appendix A. Analytical solution of temporal smoothness regularized formulation

Denote $\mathcal{P}_r(.)$ as the row selection operator parameterized by a selection vector r. The resulting matrix of $\mathcal{P}_r(A)$ includes only A_i such that $r_i \neq 0$, where A_i is the ith row of A. Let S^i be the ith column of S. We therefore denote $X_{(i)} = \mathcal{P}_{S^i}(X) \in \mathbb{R}^{n_i \times d}$ as the input data matrix of the ith task, and $y_{(i)} = \mathcal{P}_{S^i}\left(Y^i\right) \in \mathbb{R}^{n_i \times 1}$ as the corresponding target vector, where n_i is number of samples from the ith task. First, we take the derivative of Eq. (3) with respect to W and set it to zero:

$$X^T X W - X^T Y + \theta_1 W + \theta_2 W H H^T = 0, \tag{13}$$

$$\left(X^T X + \theta_1 I_d\right) W + W\left(\theta_2 H H^T\right) = X^T Y, \tag{14}$$

where I_d is the identity matrix of size d by d. Since both matrices $(X^T X + \theta_1 I_d)$ and $\theta_2 H H^T$ are symmetric, we write the eigen-decomposition of these two matrices by $Q_1 \Lambda_1 Q_1^T$ and $Q_2 \Lambda_2 Q_2^T$, where $\Lambda_1 = \mathrm{diag}\left(\lambda_1^{(1)}, \lambda_1^{(2)}, \ldots, \lambda_1^{(d)}\right)$ and $\Lambda_2 = \mathrm{diag}\left(\lambda_2^{(1)}, \lambda_2^{(2)}, \ldots, \lambda_2^{(d)}\right)$, are their eigenvalues, and Q_1 and Q_2 are orthogonal. Plugging them into Eq. (14) we get:

$$Q_1 \Lambda_1 Q_1^T W + W Q_2 \Lambda_2 Q_2^T = X^T Y, \tag{15}$$

$$\Lambda_1 Q_1^T W Q_2 + Q_1^T W Q_2 \Lambda_2 = Q_1^T X^T Y Q_2. \tag{16}$$

Denote $\hat{W} = Q_1^T W Q_2$ and $D = Q_1^T X^T Y Q_2$. Eq. (16) becomes $\Lambda_1 \hat{W} + \hat{W}\Lambda_2 = D$. Thus $\hat{W}$ is given by:

$$\hat{W}_{ij} = \frac{D_{ij}}{\lambda_1^{(i)} + \lambda_2^{(j)}}. \tag{17}$$

The optimal weight matrix is then given by $W^* = Q_1 \hat{W} Q_2^T$.

Appendix B. Analytical solution of temporal smoothness regularized formulation with incomplete data

Similar to the case without missing target values considered in Appendix A, we take the derivative of Eq. (4) with respect to w^i $(2 \le i \le t-1)$ and set it to zero:

$$Aw^{i-1} + M_i w^i + Aw^{i+1} = T_i, \tag{18}$$

where A, M_i, and T_i are defined as follows:

$$A = -\theta_2 I_d,$$
$$M_i = X_{(i)}^T X_{(i)} + \theta_1 I_d + 2\theta_2 I_d,$$
$$T_i = X_{(i)}^T y_{(i)}.$$

For the special case $i = 1$, the term $\|w^{i-1} - w^i\|_2^2$ does not exist, nor is the term $\|w^i - w^{i+1}\|_2^2$ for $i = t$. We combine the equations for all tasks $(1 \le i \le t)$, which can be represented as a block tridiagonal linear system:

$$\begin{pmatrix} M_1 & A & & & 0 \\ A & M_2 & A & & \\ & \ddots & \ddots & \ddots & \\ & & A & M_{t-1} & A \\ 0 & & & A & M_t \end{pmatrix} \begin{pmatrix} w^1 \\ w^2 \\ \vdots \\ w^{t-1} \\ w^t \end{pmatrix} = \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_{t-1} \\ T_t \end{pmatrix}. \tag{19}$$

For a general linear system of size td, it can be solved using Gaussian elimination with a time complexity of $O((td)^3)$. For our block tridiagonal system, the complexity is reduced to $O(d^3 t)$ using block Gaussian elimination. For large-scale linear systems, the LSQR algorithm (Paige and Saunders, 1982), a popular iterative method for the solution of large linear systems of equations, can be employed with a time complexity of $O(Ntd^2)$, where N, the number of iterations, is typically small.

Appendix C. Decomposition property of cFSGL

Theorem 1. Define

$$\pi_{\mathrm{FL}}(\mathbf{v}) = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}-\mathbf{v}\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|R\mathbf{w}\|_1 \tag{20}$$

$$\pi_{\mathrm{GL}}(\mathbf{v}) = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}-\mathbf{v}\|_2^2 + \lambda_3\|\mathbf{w}\|_2. \tag{21}$$

Then the following holds:

$$\pi(\mathbf{v}) = \pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})). \tag{22}$$

Proof. The necessary and sufficient optimality conditions for Eqs. (8), (20), and (21) can be written as:

$$0 \in \pi(\mathbf{v}) - \mathbf{v} + \lambda_1 \mathrm{SGN}(\pi(\mathbf{v})) + \lambda_2 R^T \mathrm{SGN}(R\pi(\mathbf{v})) + \lambda_3 \partial g(\pi(\mathbf{v})), \tag{23}$$

$$0 \in \pi_{\mathrm{FL}}(\mathbf{v}) - \mathbf{v} + \lambda_1 \mathrm{SGN}(\pi_{\mathrm{FL}}(\mathbf{v})) + \lambda_2 R^T \mathrm{SGN}(R\pi_{\mathrm{FL}}(\mathbf{v})), \tag{24}$$

$$0 \in \pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})) - \pi_{\mathrm{FL}}(\mathbf{v}) + \lambda_3 \partial g(\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))), \tag{25}$$

where SGN(x) is a set defined in a componentwise manner as:

$$(\mathrm{SGN}(\mathbf{x}))_i = \begin{cases} [-1,1] & x_i = 0 \\ \{1\} & x_i > 0 \\ \{-1\} & x_i < 0, \end{cases} \tag{26}$$

and

$$\partial g(\mathbf{x}) = \begin{cases} \dfrac{\mathbf{x}}{\|\mathbf{x}\|_2} & \mathbf{x} \neq 0 \\ \{\mathbf{y} : \|\mathbf{y}\|_2 \le 1\} & \mathbf{x} = 0. \end{cases} \tag{27}$$

It follows from Eqs. (25) and (27) that: 1) if $\|\pi_{\mathrm{FL}}(\mathbf{v})\|_2 \le \lambda_3$, then $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})) = 0$; and 2) if $\|\pi_{\mathrm{FL}}(\mathbf{v})\|_2 > \lambda_3$, then

$$\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})) = \frac{\|\pi_{\mathrm{FL}}(\mathbf{v})\|_2 - \lambda_3}{\|\pi_{\mathrm{FL}}(\mathbf{v})\|_2} \pi_{\mathrm{FL}}(\mathbf{v}).$$

It is easy to observe that, 1) if the ith entry of $\pi_{\mathrm{FL}}(\mathrm{v})$ is zero, so is the ith entry of $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))$; 2) if the ith entry of $\pi_{\mathrm{FL}}(\mathrm{v})$ is positive (or negative), so is the ith entry of $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))$. Therefore, we have

$$\mathrm{SGN}(\pi_{\mathrm{FL}}(\mathbf{v})) \subseteq \mathrm{SGN}(\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))). \tag{28}$$

Meanwhile, 1) if the ith and the $i+1$th entries of $\pi_{\mathrm{FL}}(\mathbf{v})$ are identical, so are those of $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))$; 2) if the ith entry is larger (or smaller) than the $i+1$th entry in $\pi_{\mathrm{FL}}(\mathrm{v})$, so is in $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))$. Therefore, we have

$$\mathrm{SGN}(R\pi_{\mathrm{FL}}(\mathbf{v})) \subseteq \mathrm{SGN}(R\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))). \tag{29}$$

It follows from Eqs. (24), (25), (28), and (29) that

$$0 \in \pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})) - \mathbf{v} + \lambda_1 \mathrm{SGN}(\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))) + \lambda_2 R^T \mathrm{SGN}(R\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))) + \lambda_3 \partial g(\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))). \tag{30}$$

Since Eq. (8) has a unique solution, we can get Eq. (22) from Eqs. (23) and (30). □

Note that the fused Lasso signal approximator (Friedman et al., 2007) in Eq. (21) can be effectively solved using (Liu et al., 2010). The complete algorithm for solving the proximal operator associated with cFSGL is given in Algorithm 1.

Algorithm 1. Proximal operator associated with the convex fused sparse group Lasso (cFSGL).

Input: V, R, λ_1, λ_2, λ_3
Output: W
1: for i = 1: t do
2: $u_i = \mathrm{argmin}_w \frac{1}{2}\|w - v_i\|_2^2 + \lambda_1\|w\|_1 + \lambda_2\|Rw\|_1$
3: $w_i = \mathrm{argmin}_w \frac{1}{2}\|w - u_i\|_2^2 + \lambda_3\|w\|_2$
4: end for

References

A. Association, 2010. 2010 Alzheimer's disease facts and figures. Alzheimers Dement. 6, 158–194.
Amemiya, T., 1973. Regression analysis when the dependent variable is truncated normal. Econometrica 997–1016.
Amemiya, T., 2010. Tobit models: a survey. J. Econom. 24 (1-2), 3–61.
Ando, R., Zhang, T., 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. J. Mach. Learn. Res. 6, 1817–1853.
Apostolova, L., et al., 2006. 3D mapping of mini-mental state examination performance in clinical and preclinical Alzheimer disease. Alzheimer Dis. Assoc. Disord. 20 (4), 224.
Argyriou, A., Evgeniou, T., Pontil, M., 2008. Convex multi-task feature learning. Mach. Learn. 73 (3), 243–272.
Ashford, J., Schmitt, F., 2001. Modeling the time-course of Alzheimer dementia. Curr. Psychiatry Rep. 3 (1), 20–28.
Bakker, B., Heskes, T., 2003. Task clustering and gating for Bayesian multitask learning. J. Mach. Learn. Res. 4, 83–99.
Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82 (4), 239–259.

Caroli, A., Frisoni, G., et al., 2010. The dynamics of Alzheimer's disease biomarkers in the Alzheimer's disease neuroimaging initiative cohort. Neurobiol. Aging 31 (8), 1263–1274.
Chen, J., Tang, L., Liu, J., Ye, J., 2009. A convex formulation for learning shared structures from multiple tasks. Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp. 137–144.
Chetelat, G., Baron, J., 2003. Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. Neuroimage 18 (2), 525–541.
Convit, A., et al., 2000. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. Neurobiol. Aging 21 (1), 19–26.
Davatzikos, C., Xu, F., An, Y., Fan, Y., Resnick, S., 2009. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain 132 (8), 2026.
Delacourte, A., et al., 1999. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. Neurology 52 (6), 1158.
Doody, R., Massman, P., Dunn, J., 2001. A method for estimating progression rates in Alzheimer disease. Arch. Neurol. 58 (3), 449.
Doody, R., Pavlik, V., Massman, P., Rountree, S., Darby, E., Chan, W., et al., 2010. Predicting progression of Alzheimer's disease. Alzheimers Res. Ther. 2 (2).
Dubois, B., et al., 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol. 6 (8), 734–746.
Duchesne, S., Caroli, A., Geroldi, C., Collins, D., Frisoni, G., 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. Neuroimage 47 (4), 1363–1370.
Eleftherohorinou, H., Hoggart, C., Wright, V., Levin, M., Coin, L., 2011. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. Hum. Mol. Genet. 20 (17), 3494–3506.
Epstein, M.P., Lin, X., Boehnke, M., 2003. A Tobit variance-component method for linkage analysis of censored trait data. Am. J. Hum. Genet. 72 (3), 611.
Evgeniou, T., Micchelli, C., Pontil, M., 2006. Learning multiple tasks with kernel methods. J. Mach. Learn. Res. 6 (1), 615.
Ferrarini, L., et al., 2008. MMSE scores correlate with local ventricular enlargement in the spectrum from cognitively normal to Alzheimer disease. Neuroimage 39 (4), 1832–1838.
Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. Ann. Appl. Stat. 1 (2), 302–332.
Frisoni, G., et al., 2002. Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry. J. Neurol. Neurosurg. Psychiatry 73 (6), 657.
Frisoni, G., Fox, N., Jack, C., Scheltens, P., Thompson, P., 2010. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. 6 (2), 67–77.
Frone, M.R., Cooper, M.L., Russell, M., 1994. Stressful life events, gender, and substance use: an application of Tobit regression. Psychol. Addict. Behav. 8 (2), 59.
Ito, K., et al., 2010. Disease progression model for cognitive deterioration from Alzheimer's Disease Neuroimaging Initiative database. Alzheimers Dement. 6 (1), 39–53.
Jack Jr., C., Bernstein, M., Fox, N., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P., Whitwell, J.L., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.
Jack Jr., C., Knopman, D., Jagust, W., Shaw, L., Aisen, P., Weiner, M., Petersen, R., Trojanowski, J., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9 (1), 119–128.
Jacob, L., Bach, F., Vert, J., 2009. Clustered multi-task learning: a convex formulation. Adv. Neural Inf. Process. Syst. 21, 745–752.
Jeffrey, R.P., Coleman, R.E., Doraiswamy, P.M., 2003. Neuroimaging and early diagnosis of Alzheimer disease: a look to the future. Radiology 226, 315–336.
Julkunen, V., et al., 2009. Cortical thickness analysis to detect progressive mild cognitive impairment: a reference to Alzheimer's disease. Dement. Geriatr. Cogn. Disord. 28 (5), 404–412.
Khachaturian, Z., 1985. Diagnosis of Alzheimer's disease. Arch. Neurol. 42 (11), 1097.
Likeman, M., et al., 2005. Visual assessment of atrophy on magnetic resonance imaging in the diagnosis of pathologically confirmed young-onset dementias. Arch. Neurol. 62 (9), 1410.
Liu, J., Ji, S., Ye, J., 2009. SLEP: Sparse Learning with Efficient Projections. Arizona State University.
Liu, J., Yuan, L., Ye, J., 2010. An efficient algorithm for a class of fused lasso problems. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'10. ACM, pp. 323–332.
McBee, M., 2010. Modeling outcomes with floor or ceiling effects: an introduction to the Tobit model. Gift. Child Q. 54 (4), 314–320.
McDonald, J.F., Moffitt, R.A., 1980. The uses of Tobit analysis. Rev. Econ. Stat. 62 (2), 318–321.
McEvoy, L., et al., 2009. Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment1. Radiology 251 (1), 195.
McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E., 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34, 939–944.
Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 72 (4), 417–473.
Misra, C., Fan, Y., Davatzikos, C., 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), 1415–1422.
Murphy, E., Holland, D., Donohue, M., McEvoy, L., Hagler Jr., D., Dale, A., Brewer, J., et al., 2010. Six-month atrophy in MTL structures is associated with subsequent memory decline in elderly controls. Neuroimage 53 (4), 1310–1317.
Nemirovski, A., 2005. Efficient Methods in Convex Programming.
Nesterov, Y., 2004. Introductory Lectures on Convex Optimization: A Basic Course. Springer Netherlands.
Obozinski, G., Taskar, B., Jordan, M., 2006. Multi-task feature selection. Statistics Department, UC Berkeley, Tech. Rep.
Paige, C., Saunders, M., 1982. LSQR: an algorithm for sparse linear equations and sparse least squares. ACM Trans. Math. Softw. (TOMS) 8 (1), 43–71.
Pearson, R., Kingan, R., Hochberg, A., 2005. Disease progression modeling from historical clinical databases. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. ACM, pp. 788–793.
Ravona-Springer, R., Moshier, E., Schmeidler, J., Godbold, J., Akrivos, J., Rapp, M., Grossman, H.T., Wysocki, M., Silverman, J.M., Haroutunian, V., et al., 2012. Changes in glycemic control are associated with changes in cognition in non-diabetic elderly. J. Alzheimers Dis. 30 (2), 299–309.
Rosen, W., Mohs, R., Davis, K., 1984. A new rating scale for Alzheimer's disease. Am. J. Psychiatry 141 (11), 1356.
Ryali, Z., Chen, T., Supekar, K., Menon, V., 2012. Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. Neuroimage 59 (1), 3852–3861.
Spulber, G., Niskanen, E., MacDonald, S., Smilovici, O., Chen, K., Reiman, E., Jauhiainen, A., Hallikainen, M., Tervo, S., Wahlund, L., et al., 2010. Whole brain atrophy rate predicts progression from MCI to Alzheimer's disease. Neurobiol. Aging 31 (9), 1601–1605.
Stekhoven, D., Hennig, L., Sveinbjörnsson, G., Moraes, I., Maathuis, M., Bühlmann, P., 2011. Causal Stability Ranking.
Stonnington, C., Chu, C., Klöppel, S., Jack Jr., C., Ashburner, J., Frackowiak, R., 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 51 (4), 1405–1413.
Thompson, P., et al., 2004. Mapping hippocampal and ventricular change in Alzheimer disease. Neuroimage 22 (4), 1754–1766.
Thrun, S., O'Sullivan, J., 1998. Clustering learning tasks and the selective cross-task transfer of knowledge. Learning to learn. 181–209.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 267–288.
Tombaugh, T., 2005. Test–retest reliable coefficients and 5-year change scores for the MMSE and 3MS. Arch. Clin. Neuropsychol. 20 (4), 485–503.
Vemuri, P., et al., 2009. MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. Neurology 73 (4), 294.
Vounou, M., Janousova, E., Wolz, R., Stein, J., Thompson, P., Rueckert, D., Montana, G., 2012. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. Neuroimage 60 (1), 700–716.
Walhovd, K., et al., 2010. Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer disease. Am. J. Neuroradiol. 31 (2), 347.
Wang, H., Nie, F., Huang, H., Yan, J., Kim, S., Risacher, S., Saykin, A., Shen, L., 2012. High-order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction. In: Bartlett, P., Pereira, F., Burges, C., Bottou, L., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems, 25, pp. 1286–1294.
Wimo, A., Winblad, B., Aguero-Torres, H., von Strauss, E., 2003. The magnitude of dementia occurrence in the world. Alzheimer Dis. Assoc. Disord. 17 (2), 63.
Xue, Y., Liao, X., Carin, L., Krishnapuram, B., 2007. Multi-task learning for classification with dirichlet process priors. J. Mach. Learn. Res. 8, 35–63.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 68 (1), 49–67.
Zhang, D., Shen, D., 2011. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 59 (2), 895–907.
Zhang, D., Shen, D., 2012. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS One 7 (3), e33182.
Zhang, Y., Yeung, D.-Y., 2010. Multi-task learning using generalized *t* process. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp. 964–971.
Zhou, J., Chen, J., Ye, J., 2011. Clustered multi-task learning via alternating structure optimization. Adv. Neural Inf. Process. Syst. 24, 702–710.