Modeling Disease Progression via Fused Sparse Group Lasso

Jiayu Zhou[1,2], Jun Liu[1], Vaibhav A. Narayan[3], Jieping Ye[1,2]
[1]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, ASU, Tempe, AZ
[2]Department of Computer Science and Engineering, ASU, Tempe, AZ
[3]Johnson & Johnson Pharmaceutical Research & Development, LLC, Titusville, NJ

ABSTRACT

Alzheimer's Disease (AD) is the most common neurodegenerative disorder associated with aging. Understanding how the disease progresses and identifying related pathological biomarkers for the progression is of primary importance in the clinical diagnosis and prognosis of Alzheimer's disease. In this paper, we develop novel multi-task learning techniques to predict the disease progression measured by cognitive scores and select biomarkers predictive of the progression. In multi-task learning, the prediction of cognitive scores at each time point is considered as a task, and multiple prediction tasks at different time points are performed simultaneously to capture the temporal smoothness of the prediction models across different time points. Specifically, we propose a novel convex fused sparse group Lasso (cFSGL) formulation that allows the simultaneous selection of a common set of biomarkers for multiple time points and specific sets of biomarkers for 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 several non-smooth penalties. One of the main technical contributions of this paper is to show that the proximal operator associated with the proposed formulation exhibits a certain decomposition property and can be computed efficiently; thus cFSGL can be solved efficiently using the accelerated gradient method. To further improve the model, we propose two non-convex formulations to reduce the shrinkage bias inherent in the convex formulation. We employ the difference of convex (DC) programming technique to solve the non-convex formulations. We have performed extensive experiments using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Results demonstrate the effectiveness of the proposed progression models in comparison with existing methods for disease progression. We also perform longitudinal stability selection to identify and analyze the temporal patterns of biomarkers in disease progression.

KDD'12, August 12–16, 2012, Beijing, China.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; J.3 [**Life and Medical Sciences**]: Health, Medical information systems

General Terms

Algorithms

Keywords

Alzheimer's Disease, regression, multi-task learning, fused Lasso, sparse group Lasso, cognitive score

1. INTRODUCTION

Alzheimer's disease (AD), accounting for 60-70% of age-related dementia, is a severe neurodegenerative disorder. AD is characterized by loss of memory and declination of cognitive function due to progressive impairment of neurons and their connections, leading directly to death [21]. In 2011 there are approximately 30 million individuals afflicted with dementia and the number is projected to be over 114 million by 2050 [36]. 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. There are two types of progression models that have been commonly used in the literature: the regression model [10, 31] and the survival model [30, 34]. Many existing work consider a small number of input features, and the model building involves an iterative process in which each feature is evaluated individually by adding to the model and testing the performance of predicting the target representing the *disease status* [18, 35]. The disease status can be measured by a clinical score such as Mini Mental State Examination (MMSE) or Alzheimer's Disease Assessment Scale

cognitive subscale (ADAS-Cog) [10, 31], or the volume of a certain brain region [16], or clinically defined categories [9, 26]. When high-dimensional data, such as neuroimages (i.e., MRI and/or PET) are used as input features, 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 [10]. One disadvantage of using dimension reduction is that the models are no longer interpretable. A better alternative is to use feature selection in modeling the disease progression [31]. Most existing work focus on the prediction of target at a single time point (baseline [31], or one year [10]); however, a joint analysis of data 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, multi-task learning techniques have recently been proposed to model the disease progression [39, 43]. The idea of multi-task learning is to utilize the intrinsic relationships among multiple related tasks in order to improve the generalization 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 all tasks are related and task models are closed to each other [11], or tasks are clustered into groups [4, 20, 32, 41]. Alternatively, one can assume that the tasks share a common subspace [2, 7], or a common set of features [3, 29]. In [39], the prediction of different types of targets such as MMSE and ADAS-Cog is modeled as a multi-task learning problem and all models are constrained to share a common set of features. In [43], multi-task learning is used to model the longitudinal disease progression. Given the set of baseline features of a patient, the prediction of the patient's disease status at each time point can be considered as a regression task. Multiple prediction tasks at different time points are performed simultaneously to capture the temporal smoothness of the prediction models across different time points. However, similar to [39], the formulation in [43] constrains the models at all time points to select a common set of features, thus failing to capture the *temporal patterns* of the biomarkers in disease progression [6, 19]. It is thus desirable to develop formulations that allow the simultaneous selection of a common set of biomarkers for multiple time points and specific sets of biomarkers for different time points.

In this paper, we propose novel multi-task learning formulations for predicting the disease progression measured by the clinical scores (ADAS-Cog and MMSE). Specifically, we propose a convex fused sparse group Lasso (cFSGL) formulation that simultaneously selects a common set of biomarkers for all time points and selects a specific set of biomarkers at different time points using the sparse group Lasso penalty [14], and in the meantime incorporates the temporal smoothness using the fused Lasso penalty [33]. The proposed formulation is, however, challenging to solve due to the use of several non-smooth penalties including the sparse group Lasso and fused Lasso penalties. We show that the proximal operator associated with the optimization problem in cFSGL exhibits a certain decomposition property and can be solved efficiently. Therefore cFSGL can be efficiently solved using the accelerated gradient method [27, 28]. The convex sparsity-inducing penalties are known to introduce shrinkage bias [12]. To further improve the progression model and reduce the shrinkage bias in cFSGL, we propose two non-convex progression formulations. We employ the difference of convex (DC) programming technique to solve the non-convex formulations, which iteratively solves a sequence of convex relaxed optimization problems. We show that at each step the convex relaxed problems are equivalent to reweighted sparse learning problems [5].

We have performed extensive experiments to demonstrate the effectiveness of the proposed models using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). We have also performed longitudinal stability selection [43] using our proposed formulations to identify and analyze the temporal patterns of biomarkers in disease progression.

2. A CONVEX FORMULATION OF MODELING DISEASE PROGRESSION

In the longitudinal AD study, cognitive scores of selected patients are repeatedly measured at multiple time points. The prediction of cognitive scores at each time point can be considered as a regression problem, and the prediction of cognitive scores at multiple time points can be treated as a multi-task regression problem. By employing multi-task regression, the temporal information among different tasks can be incorporated into the model to improve the prediction performance.

Consider a multi-task regression problem of t tasks with n samples of d features. Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be the input data at the baseline, and let $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ be the targets, where each $\mathbf{x}_i \in \mathbb{R}^d$ represents a sample (patient), and $y_i \in \mathbb{R}^t$ is the corresponding targets (clinical scores) at different time points. We collectively 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 $W = \left[\mathbf{w}^1, \cdots, \mathbf{w}^t\right] \in \mathbb{R}^{d \times t}$ as the weight matrix. To consider the missing values from the target, we denote the loss function as:

$$L(W) = \|S \odot (XW - Y)\|_F^2 , \tag{1}$$

where matrix $S \in \mathbb{R}^{n \times t}$ indicates missing target values: $S_{i,j} = 0$ if the target value of sample i is missing at the jth time point, and $S_{i,j} = 1$ otherwise. The component-wise operator $\odot$ is defined as follows: $Z = A \odot B$ denotes $Z_{i,j} = A_{i,j} B_{i,j}$, for all i, j. The multi-task regression solves the following optimization problem: $\min_W L(W) + \Omega(W)$, where $\Omega(W)$ is a regularization term that captures the task relatedness.

In the multi-task setting for modeling disease progression, each task is to predict a specific target (e.g., MMSE) for a set of subjects at different time points. It is thus reasonable to assume that the difference of the predictions between immediate time points is small, i.e., the temporal smoothness [43]. It is also well believed in the literature that a small subset of biomarkers are related to the disease progression, and biomarkers involved at different stages may be different [19]. To this end, we propose a novel multi-task learning formulation for modeling disease progression which allows simultaneous joint feature selection for multiple tasks and task-specific feature selection, and in the meantime incorporates the temporal smoothness. Mathematically, the proposed formulation solves the following convex optimiza-

tion problem:

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

where $\|W\|_1$ is the Lasso penalty, the group Lasso penalty $\|W\|_{2,1}$ is given by $\sum_{i=1}^{d} \sqrt{\sum_{j=1}^{t} W_{ij}^2}$, $\left\|RW^T\right\|_1$ is the fused Lasso penalty, R is an $(t-1) \times t$ sparse matrix in which $R_{i,i} = 1$ and $R_{i,i+1} = -1$, 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. We call the formulation in Eq. (2) "convex fused sparse group Lasso" (cFSGL). 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) [27, 28]. 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:

$$\pi(V) = \arg\min_W \frac{1}{2}\|W - V\|_F^2 + \lambda_1\|W\|_1 + \lambda_2\|RW^T\|_1 + \lambda_3\|W\|_{2,1}. \quad (3)$$

It is clear that each row of W is decoupled in Eq. (3). Thus for obtaining the ith row $\mathbf{w}_i$, we only need to solve the following optimization problem:

$$\pi(\mathbf{v}_i) = \arg\min_{\mathbf{w}_i} \frac{1}{2}\|\mathbf{w}_i - \mathbf{v}_i\|_2^2 + \lambda_1\|\mathbf{w}_i\|_1 + \lambda_2\|R\mathbf{w}_i\|_1 + \lambda_3\|\mathbf{w}_i\|_2, \quad (4)$$

where $\mathbf{v}_i$ is the ith row of V. The proximal operator in Eq. (4) is challenging to compute due to the presence of three non-smooth terms. One of the key technical contributions of this paper is to 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 the following theorem:

THEOREM 1. *Define*

$$\pi_{\mathrm{FL}}(\mathbf{v}) = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|R\mathbf{w}\|_1 \quad (5)$$

$$\pi_{\mathrm{GL}}(\mathbf{v}) = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda_3\|\mathbf{w}\|_2. \quad (6)$$

Then the following holds:

$$\pi(\mathbf{v}) = \pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})). \quad (7)$$

Proof: The necessary and sufficient optimality conditions for (4), (5), and (6) can be written as:

$$\mathbf{0} \in \pi(\mathbf{v}) - \mathbf{v} + \lambda_1\mathrm{SGN}(\pi(\mathbf{v})) + \lambda_2 R^T\mathrm{SGN}(R\pi(\mathbf{v})) + \lambda_3\partial g(\pi(\mathbf{v})), \quad (8)$$

$$\mathbf{0} \in \pi_{\mathrm{FL}}(\mathbf{v}) - \mathbf{v} + \lambda_1\mathrm{SGN}(\pi_{\mathrm{FL}}(\mathbf{v})) + \lambda_2 R^T\mathrm{SGN}(R\pi_{\mathrm{FL}}(\mathbf{v})), \quad (9)$$

$$\mathbf{0} \in \pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})) - \pi_{\mathrm{FL}}(\mathbf{v}) + \lambda_3\partial g(\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))), \quad (10)$$

where $\mathrm{SGN}(\mathbf{x})$ is a set defined in a componentwise manner as:

$$(\mathrm{SGN}(\mathbf{x}))_i = \begin{cases} [-1,1] & x_i = 0 \\ \{1\} & x_i > 0 \\ \{-1\} & x_i < 0, \end{cases} \quad (11)$$

and

$$\partial g(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|_2} & \mathbf{x} \neq \mathbf{0} \\ \{\mathbf{y} : \|\mathbf{y}\|_2 \leq 1\} & \mathbf{x} = \mathbf{0}. \end{cases} \quad (12)$$

It follows from (10) and (12) that: 1) if $\|\pi_{\mathrm{FL}}(\mathbf{v})\|_2 \leq \lambda_3$, then $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})) = \mathbf{0}$; 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 i-th entry of $\pi_{\mathrm{FL}}(\mathbf{v})$ is zero, so is the i-th entry of $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))$; 2) if the i-th entry of $\pi_{\mathrm{FL}}(\mathbf{v})$ is positive (or negative), so is the i-th entry of $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))$. Therefore, we have:

$$\mathrm{SGN}(\pi_{\mathrm{FL}}(\mathbf{v})) \subseteq \mathrm{SGN}(\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))). \quad (13)$$

Meanwhile, 1) if the i-th 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 i-th entry is larger (or smaller) than the $(i+1)$-th entry in $\pi_{\mathrm{FL}}(\mathbf{v})$, so is in $\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))$. Therefore, we have:

$$\mathrm{SGN}(R\pi_{\mathrm{FL}}(\mathbf{v})) \subseteq \mathrm{SGN}(R\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))). \quad (14)$$

It follows from (9), (10), (13), and (14) that:

$$\mathbf{0} \in \pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v})) - \mathbf{v} + \lambda_1\mathrm{SGN}(\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))) + \lambda_2 R^T\mathrm{SGN}(R\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))) + \lambda_3\partial g(\pi_{\mathrm{GL}}(\pi_{\mathrm{FL}}(\mathbf{v}))). \quad (15)$$

Since (4) has a unique solution, we can get (7) from (8) and (15). □

Note that the fused Lasso signal approximator [13] in Eq.(5) can be effectively solved using [24]. The complete algorithm for computing 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: $V \in \mathbb{R}^{d\times t}$, $R \in \mathbb{R}^{t-1\times t}$, λ_1, λ_2, λ_3
Output: $W \in \mathbb{R}^{d\times t}$
1: **for** i = 1 : d **do**
2: $\quad \mathbf{u}_i = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \mathbf{v}_i\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|R\mathbf{w}\|_1$
3: $\quad \mathbf{w}_i = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \mathbf{u}_i\|_2^2 + \lambda_3\|\mathbf{w}\|_2$
4: **end for**

3. NON-CONVEX PROGRESSION MODELS

In cFSGL, we aim to select task-shared and task-specific features using the sparse group Lasso penalty. However, the decomposition property shown in Theorem 1 implies that a simple composition of the ℓ_1-norm penalty and $\ell_{2,1}$-norm penalty may be sub-optimal. Besides, the sparsity-inducing penalties are known to lead to biased estimates [12]. To this end, we propose the following non-convex multi-task regression formulation for modeling disease progression:

$$\min_W L(W) + \lambda \sum_{i=1}^{d} \sqrt{\|\mathbf{w}_i\|_1} + \gamma\|RW^T\|_1, \quad (16)$$

where the second term is the summation of the squared root of ℓ_1-norm of $\mathbf{w}_i$ ($\mathbf{w}_i$ is the ith row of W), and is called the composite $\ell_{(0.5,1)}$-norm regularization. Note that it is in fact not a valid norm due to its non-convexity. It is known that the $\ell_{0.5}$ penalty leads to a sparse solution, thus many of the rows of W will be zero, i.e., the features corresponding to the zero rows will be removed from all tasks. In addition, for the nonzero rows, due to the use of the ℓ_1 penalty for

the rows, many features within these nonzero rows will be zero, resulting in task-specific features. Thus, the use of $\ell_{(0.5,1)}$ penalty leads to a tight coupling of between-task and within-task feature selection. In addition, the $\ell_{0.5}$ penalty is expected to reduce the estimation bias associated with the convex sparsity-inducing penalties.

We also consider an alternative non-convex formulation which includes the fused Lasso term of each row within the square root, resulting in a composite $\ell_{(0.5,1)}$-like penalty:

$$\min_W L(W) + \lambda \sum_{i=1}^{d} \sqrt{\|R\mathbf{w}_i^T\|_1 + \beta\|\mathbf{w}_i\|_1}. \quad (17)$$

A good merit of using non-convex penalties is that they are closer to the optimal ℓ_0-'norm' (minimizing which is NP-hard) and give better sparsity [15]. In addition, a practical advantage of the non-convex progression models presented in Eqs. (16) and (17) is that there are only 2 regularization parameters to be estimated, compared to 3 parameters in the convex formulation in Eq. (2). However, one disadvantage of the non-convex penalties is that the associated optimization problems are non-convex and global solutions are not guaranteed. A well-known method for solving non-convex problems is to approximate the non-convex formulation by a convex relaxation via the difference of convex (DC) programming techniques [17]. Next, we show how the non-convex problems can be solved using DC programming and then relate the relaxed formulations to reweighted convex formulations.

3.1 DC Programming

The formulations in Eq. (16) and Eq. (17) can be expressed in the following general form:

$$\min_W \ell(W) + \sqrt{h(W)}, \quad (18)$$

where $\ell(W)$ and $h(W)$ are convex. Since $x - \sqrt{x}$ is convex, we decompose the objective function in Eq. (18) into the following form:

$$\min_W \ell(W) + h(W) - (h(W) - \sqrt{h(W)}).$$

Denote the two functions as $f(W) = \ell(W) + h(W)$ and $g(h(W)) = h(W) - \sqrt{h(W)}$. We can express the formulation in Eq. (18) in the form of difference of two functions:

$$\min_W f(W) - g(h(W)).$$

Using the convex-concave procedure (CCCP) algorithm [38], we can linearize $g(h(W))$ using the 1st-order Taylor expansion at the current point W' as:

$$f(W) - g(h(W)) = f(W) - g(h(W')) - \langle \nabla g(h(W')), (h(W) - h(W')) \rangle, \quad (19)$$

which is the convex upper bound of the non-convex problem. In every iteration of the CCCP algorithm, we minimize the upper bound:

$$W_{(k+1)} = \operatorname*{argmin}_W f(W) - \langle \nabla g(h(W_{(k)})), h(W) \rangle, \quad (20)$$

and the objective function is guaranteed to decrease. We obtain a local optimal W^* of Eq. (18) by iteratively solving Eq. (20). The CCCP algorithm has been applied successfully to solve many non-convex problems [8, 22, 37].

3.2 Reweighting Interpretation of Non-Convex Fused Sparse Group Lasso

We first consider the non-convex optimization problem in Eq. (16), whose convex relaxed form corresponding to Eq. (20) is given by:

$$W_{(k+1)} = \operatorname*{argmin}_W L(W) + \frac{\lambda}{2} \sum_{i=1}^{d} \mu_i \|\mathbf{w}_i\|_1 + \gamma \|RW^T\|_1, \quad (21)$$

where $\mu_i = 1/\sqrt{\|\mathbf{w}_{i(k)}\|_1 + \epsilon}$ and ϵ is a small number included to avoid singularity. It is clear that the convex relaxed problem in each iteration is a fused Lasso problem with a reweighted ℓ_1-norm term. If we omit the fused term, the general $\ell_{(1,0.5)}$-regularized optimization problem is of the following form:

$$\min_W L(W) + \lambda \sum_{i=1}^{d} \sqrt{\|\mathbf{w}_i\|_1},$$

which, under DC programming, involves solving a series of reweighted Lasso [5, 23] problems:

$$W_{(k+1)} = \operatorname*{argmin}_W L(W) + \frac{\lambda}{2} \sum_{i=1}^{d} \mu_i \|\mathbf{w}_i\|_1.$$

It is known that the reweighted Lasso reduces the estimation bias of Lasso, thus leading to a better solution. Similarly, for the non-convex optimization problem in Eq. (17), we iteratively solve the following convex problem:

$$W_{(k+1)} = \operatorname*{argmin}_W L(W) + \frac{\lambda}{2} \sum_{i=1}^{d} \nu_i \|R\mathbf{w}_i^T\|_1 + \frac{\lambda\beta}{2} \sum_{i=1}^{d} \nu_i \|\mathbf{w}_i\|_1, \quad (22)$$

where $\nu_i = 1/\sqrt{\|R\mathbf{w}_{i(k)}^T\|_1 + \beta\|\mathbf{w}_{i(k)}\|_1 + \epsilon}$. In this case, in each iteration, we solve a fused Lasso problem with a reweighted ℓ_1-term and a reweighted fused term.

The non-convex optimization problems may be sensitive to the starting point. In our algorithm in Eq. (21), for example, if all elements in row i of the model $\mathbf{w}_i$ are initialized to be close to 0, then in the next iteration μ_i will be set to a very large number. The large penalty forces the row to stay at 0 in later iterations. Therefore, in our convex relaxed algorithms in Eq. (21) and Eq. (22), we propose to use the solution of a problem similar to fused Lasso as the starting point. For example, the starting point we use in Eq. (21) is:

$$W_{(0)} = \operatorname*{argmin}_W L(W) + \lambda \sum_{i=1}^{d} \|\mathbf{w}_i\|_1 + \gamma \|RW^T\|_1. \quad (23)$$

This is equivalent to setting $\mu_i/2 = 1$. Similarly, in Eq. (22) we set $\nu_i/2 = 1$.

4. ANALYZE TEMPORAL PATTERNS OF BIOMARKERS USING LONGITUDINAL STABILITY SELECTION

We propose to employ longitudinal stability selection to quantify the importance of the features selected by the proposed formulations for disease progression. The idea of longitudinal stability selection is to apply stability selection [25]

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 cFSGL and nFSGL 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 cFSGL and nFSGL 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 $\langle \lambda_1, \lambda_2, \lambda_3 \rangle$, and for nFSGL is a tuple of the corresponding parameter pairs. 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 cFSGL or nFSGL on $B_{(i)}$. The set of features selected by the model $\hat{W}^{(i)}$ of the task at time point p is denoted by:

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

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:

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

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

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

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

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

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.

We emphasize here that unlike the previous work which gives a list of features common for all time points [43], our proposed approaches yield a different list of features at different time points. Note that in the above stability selection we use temporal information via fused Lasso. Consequently the distribution of stability scores also has temporal smoothness property: for each feature the stability scores are smooth across different time points (as shown in experimental results in Section 6.3). If simply using Lasso in stability selection, then we obtain independent probability lists at each time point, and therefore such temporal smooth pattern cannot be captured.

5. RELATION TO PREVIOUS WORK

In our previous work [43], we proposed to use the temporal group Lasso (TGL) regularization to capture task relatedness, which involves the following optimization problem:

$$\min_{W} L(W) + \theta_1 \|W\|_F^2 + \theta_2 \left\|RW^T\right\|_F^2 + \theta_3 \|W\|_{2,1}, \quad (24)$$

where θ_1, θ_2 and θ_3 are regularization parameters. The TGL formulation in Eq. (24) contains three penalty terms. The first term penalize the ℓ_2-norm of the model to prevent overfitting; the second term enforces temporal smoothness using ℓ_2-norm, which is equivalent to a Laplacian term, and the last $\ell_{2,1}$-norm introduces joint feature selection. We argue that it is more natural to incorporate the within-task feature selection and temporal smoothness using a composite penalty as in our proposed cFSGL formulation in Eq. (2).

For example, the only sparsity-inducing term in TGL formulation in Eq. (24) is the $\ell_{2,1}$-norm regularized joint feature selection. Therefore an obvious disadvantage of this formulation is that it restricts all models from different time points to select a common set of features; however, different features may be involved at different time points. In addition, one key advantage of fused Lasso compared with the Laplacian-based smoothing used in [43] is that under the fused Lasso penalty the selected features across different time points are smooth, i.e., nearby time points tend to select similar features, while the Laplacian-based penalty focuses on the smoothing of the prediction models across different time points. Thus, the fused Lasso penalty better captures the temporal smoothness of the selected features, which is closer to the real-world disease progression mechanism.

In the TGL formulation, the temporal smoothness is enforced using a smooth Laplacian term, though fused Lasso in cFSGL indeed has better properties such as sparsity continuity. We have used this restrictive model in TGL, in order to avoid the computational difficulties introduced by the composite of non-smooth terms ($\ell_{2,1}$-norm and fused Lasso). We show in this paper that the proximal operator associated with the optimization problem in cFSGL exhibits a certain decomposition property and can be computed efficiently (Theorem 1); thus cFSGL can be solved efficiently using accelerated gradient method. Another contribution of this paper is that we extend our progression model using a composite of non-convex sparsity-inducing terms, and we further propose to employ the DC programming to solve the non-convex formulations.

6. EXPERIMENTS

In this section we evaluate the proposed progression models on the data sets from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[1]. The source codes can be found in the Muli-tAsk Learning via StructurAl Regularization (MALSAR) package [42].

6.1 Experimental Setup

The ADNI project is a longitudinal study, where a variety of measurements are collected from selected subjects including Alzheimer's disease patients (AD), mild cognitive impairment patients (MCI) and normal controls (NL), repeatedly over a 6-month or 1-year interval. The measurements include MRI scans (M), PET scans (P), CSF measurements (C), and cognitive scores such as MMSE and ADAS-Cog. We denote all measurements other than the three types of

[1]www.loni.ucla.edu/ADNI

Table 1: The sample size and feature dimensionality of different data sets used in the experiments. M denotes baseline MMSE features and E denotes baseline META features.

Target	Source	M06	M12	M24	M36	M48	Dim.
MMSE	M	648	642	569	389	87	306
	M+E	648	642	569	389	87	371
ADAS	M	648	638	564	377	85	306
	M+E	648	642	569	389	87	371

Table 2: Features included in the META dataset. In META, we include baseline cognitive scores as features to predict the future cognitive scores. A detailed explanation of each cognitive score and lab test can be found at [1].

Type	Features
Demographic	age, years of education, gender
Genetic	ApoE-$\varepsilon 4$ information
Baseline cognitive scores	MMSE, ADAS-Cog, ADAS-MOD, ADAS subscores, CDR, FAQ, GDS, Hachinski, Neuropsychological Battery, WMS-R Logical Memory
Lab tests	RCT1, RCT11, RCT12, RCT13, RCT14, RCT1407, RCT1408, RCT183, RCT19, RCT20, RCT29, RCT3, RCT392, RCT4, RCT5, RCT6, RCT8

biomarkers (M, P, C) as META (E). A detailed list of the META data is given in Table 2. 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 for some patients. However, many patients drop out from the study for many reasons (e.g. deceased). In our experiments, we predict future MMSE and ADAS-Cog scores using various measurements at the baseline. For each target we build a prediction model using a data set that only contains baseline MRI features (M), and another data set that contains both MRI and META features (M+E). In the current study, CSF and PET are not used due to the small sample size. The MRI features are extracted in the same way as in [43]. There are 5 types of MRI features used: white matter parcellation volume (Vol.WM.), cortical parcellation volume (Vol.C.), surface area (Surf. Area), cortical thickness average (CTA), cortical thickness standard deviation (CTStd). The sample size and dimensionality for each time point and feature combination is given in Table 1.

6.2 Prediction Performance

In the first experiment, we compare the proposed methods including Convex Fused Sparse Group Lasso (cFSGL) and the two Non-Convex Fused Group Lasso: nFSGL1 in Eq. (16) and nFSGL2 in Eq. (17) with ridge regression (Ridge) and Temporal Group Lasso (TGL) on the prediction of MMSE and ADAS-Cog using selected types of feature combinations, namely M and M+E. Note that Lasso is a special case of cFSGL when both λ_2 and λ_3 are set to 0. For each feature combination, we randomly split the data into training and testing sets using a ratio 9 : 1. The 5-fold cross validation is used to select model parameters. For the regression performance measures, we use Normalized Mean Squared Error (nMSE) as used in the multi-task learning literature [40, 3] and weighted correlation coefficient (R-value) as employed in the medical literature addressing AD progression problems [10, 31, 18]. We report the mean and standard deviation based on 20 iterations of experiments on different splits of data. To investigate the effects of the fused Lasso term, in cFSGL we fix the value of λ_2 in Eq.(2) to be $20, 50, 100$, and perform cross validation to select λ_1 and λ_3. The three configurations are labeled as cFSGL1, cFSGL2 and cFSGL3 respectively.

The experimental results using 90% training data on MRI and MRI+META are presented in Table 3 and Table 4. Overall our proposed approaches outperform Ridge and TGL, in terms of both nMSE and correlation coefficient. We have the following observations: 1) The fused Lasso term is effective. We witness significant improvement in cFSGL when changing the parameter value for the fused Lasso term. 2) The proposed cFSGL and nFSGL formulations witness significant improvement for later time points. This may be due to the data sparseness at later time points (see Table 1), as the proposed sparsity-inducing models are expected to achieve better generalization performance in this case. 3) The non-convex nFSGL formulations are better than cFSGL in many tasks. One practical strength of the non-convex nFSGL formulations is that they have fewer parameters to be estimated (only 2 parameters).

6.3 Temporal Patterns of Biomarkers

One of the strengthens of the proposed formulations is that they facilitate the identification of temporal patterns of biomarkers. In this experiment we study the temporal patterns of biomarkers using longitudinal stability selection with cFSGL and nFSGL. Note that because the sample size at the M48 time point is too small, we perform stability selection for M06, M12, M24, and M36 only.

The stability vectors of MRI stable features using cFSGL nFSGL1 and nFSGL2 formulations are given in Figure 1, Figure 2 and Figure 3 respectively. In the figures, we collectively list the stable features ($\eta = 20$) at the 4 time points. The total number of features may be less than 80 because one feature may be identified as a stable feature at multiple time points. In Figure 1(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, which agrees with the previous findings [43]. 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. 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.

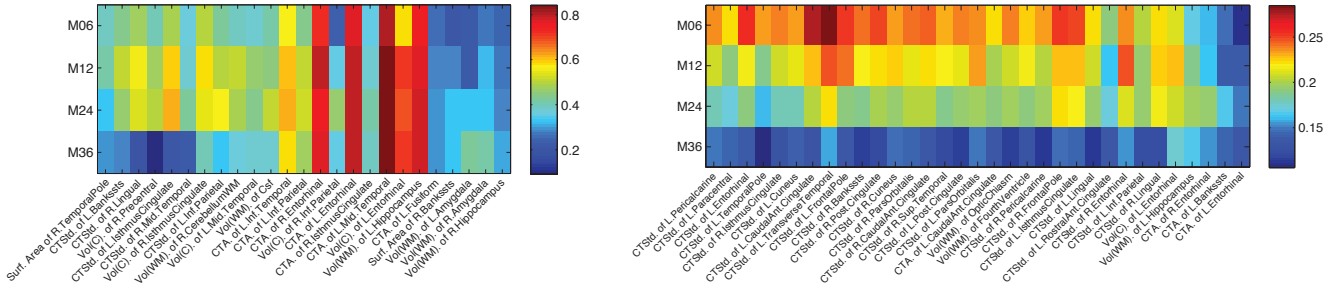
The stability vector of stable MRI features for MMSE are given in Figure 1(b). We obtain very different patterns from ADAS-Cog. We find that most biomarkers provide significant information for the first 2 years and very few of them

Table 3: Comparison of our proposed approaches (cFSGL and nFSGL) and existing approaches (Ridge and TGL) on longitudinal MMSE and ADAS-Cog prediction using MRI features (M) in terms of normalized mean squared error (nMSE), average correlation coefficient (R) and mean squared error (MSE) for each time point. 90 percent of data is used as training data.

	Ridge	TGL	cFSGL1	cFSGL2	cFSGL3	nFSGL1	nFSGL2
Target: MMSE							
nMSE	0.548 ± 0.057	0.449 ± 0.045	0.428 ± 0.052	0.400 ± 0.053	$\mathbf{0.395 \pm 0.052}$	0.412 ± 0.054	0.408 ± 0.056
R	0.689 ± 0.030	0.755 ± 0.029	0.772 ± 0.030	0.790 ± 0.032	$\mathbf{0.796 \pm 0.031}$	0.788 ± 0.031	0.792 ± 0.031
M06 MSE	2.269 ± 0.207	$\mathbf{2.038 \pm 0.262}$	2.117 ± 0.209	2.069 ± 0.209	2.071 ± 0.213	2.149 ± 0.194	2.181 ± 0.201
M12 MSE	3.266 ± 0.556	2.923 ± 0.643	2.900 ± 0.629	2.803 ± 0.662	$\mathbf{2.762 \pm 0.669}$	2.835 ± 0.662	2.793 ± 0.659
M24 MSE	3.494 ± 0.599	3.363 ± 0.733	3.125 ± 0.612	3.016 ± 0.624	3.000 ± 0.642	3.031 ± 0.604	$\mathbf{2.979 \pm 0.546}$
M36 MSE	4.003 ± 0.853	3.768 ± 0.962	3.456 ± 0.766	3.302 ± 0.781	3.265 ± 0.803	3.263 ± 0.785	$\mathbf{3.211 \pm 0.786}$
M48 MSE	4.328 ± 1.310	3.631 ± 1.226	2.857 ± 0.892	$\mathbf{2.787 \pm 0.871}$	2.871 ± 0.884	2.780 ± 0.855	2.766 ± 0.826
Target: ADAS-Cog							
nMSE	0.532 ± 0.095	0.464 ± 0.067	0.444 ± 0.059	0.404 ± 0.055	0.391 ± 0.059	0.386 ± 0.060	$\mathbf{0.381 \pm 0.057}$
R	0.705 ± 0.043	0.747 ± 0.033	0.765 ± 0.032	0.791 ± 0.026	0.803 ± 0.024	$\mathbf{0.809 \pm 0.023}$	$\mathbf{0.809 \pm 0.023}$
M06 MSE	5.213 ± 0.522	4.820 ± 0.489	4.779 ± 0.421	4.543 ± 0.374	4.451 ± 0.340	4.458 ± 0.354	$\mathbf{4.428 \pm 0.351}$
M12 MSE	6.079 ± 0.775	5.813 ± 0.697	5.605 ± 0.622	5.363 ± 0.595	5.230 ± 0.589	5.183 ± 0.597	$\mathbf{5.136 \pm 0.617}$
M24 MSE	7.409 ± 1.154	6.835 ± 1.052	6.893 ± 0.950	6.456 ± 0.974	6.249 ± 0.996	6.174 ± 0.943	$\mathbf{6.153 \pm 0.911}$
M36 MSE	7.143 ± 1.351	6.938 ± 1.363	6.475 ± 1.135	6.101 ± 1.071	5.928 ± 1.064	$\mathbf{5.819 \pm 0.945}$	5.879 ± 0.972
M48 MSE	6.644 ± 2.750	6.000 ± 2.738	5.767 ± 2.189	5.751 ± 2.081	5.980 ± 1.979	5.889 ± 1.848	$\mathbf{5.837 \pm 2.160}$

Table 4: Comparison of our proposed approaches (cFSGL and nFSGL) and existing approaches (Ridge and TGL) on longitudinal MMSE and ADAS-Cog prediction using MRI+META features (M+E) in terms of normalized mean squared error (nMSE), average correlation coefficient (R) and mean squared error (MSE) for each time point. 90 percent of data is used as training data.

	Ridge	TGL	cFSGL1	cFSGL2	cFSGL3	nFSGL1	nFSGL2
Target: MMSE							
nMSE	0.404 ± 0.056	0.320 ± 0.044	0.310 ± 0.042	0.311 ± 0.042	0.312 ± 0.043	0.308 ± 0.046	$\mathbf{0.303 \pm 0.046}$
R	0.788 ± 0.032	0.839 ± 0.027	0.842 ± 0.026	0.841 ± 0.026	0.840 ± 0.026	0.839 ± 0.027	$\mathbf{0.843 \pm 0.027}$
M06 MSE	2.188 ± 0.194	1.943 ± 0.161	1.918 ± 0.155	1.912 ± 0.153	1.907 ± 0.149	1.935 ± 0.150	$\mathbf{1.906 \pm 0.149}$
M12 MSE	2.744 ± 0.638	2.366 ± 0.722	2.355 ± 0.716	2.356 ± 0.713	2.357 ± 0.711	2.374 ± 0.696	$\mathbf{2.326 \pm 0.707}$
M24 MSE	3.113 ± 0.560	2.821 ± 0.664	2.790 ± 0.653	2.823 ± 0.656	2.875 ± 0.675	2.766 ± 0.601	$\mathbf{2.730 \pm 0.604}$
M36 MSE	3.150 ± 0.517	2.933 ± 0.657	2.851 ± 0.635	2.878 ± 0.640	2.905 ± 0.646	$\mathbf{2.755 \pm 0.550}$	2.792 ± 0.523
M48 MSE	3.639 ± 0.959	3.544 ± 1.136	3.233 ± 1.070	3.098 ± 1.013	2.956 ± 0.924	$\mathbf{2.942 \pm 0.928}$	2.961 ± 0.969
Target: ADAS-Cog							
nMSE	0.314 ± 0.036	0.278 ± 0.034	0.238 ± 0.033	$\mathbf{0.233 \pm 0.035}$	0.235 ± 0.035	0.238 ± 0.035	0.243 ± 0.035
R	0.840 ± 0.015	0.868 ± 0.016	0.882 ± 0.013	$\mathbf{0.886 \pm 0.014}$	$\mathbf{0.886 \pm 0.014}$	0.884 ± 0.015	0.880 ± 0.013
M06 MSE	3.972 ± 0.415	3.560 ± 0.469	3.566 ± 0.380	3.553 ± 0.375	3.617 ± 0.362	3.659 ± 0.356	$\mathbf{3.535 \pm 0.403}$
M12 MSE	4.365 ± 0.469	4.080 ± 0.598	3.742 ± 0.394	3.678 ± 0.389	$\mathbf{3.659 \pm 0.393}$	3.739 ± 0.367	3.742 ± 0.430
M24 MSE	6.028 ± 1.128	5.888 ± 1.641	5.226 ± 1.201	5.115 ± 1.277	5.122 ± 1.338	$\mathbf{5.111 \pm 1.222}$	5.257 ± 1.337
M36 MSE	5.824 ± 1.076	5.639 ± 1.339	4.871 ± 0.894	4.747 ± 0.957	$\mathbf{4.712 \pm 1.002}$	4.737 ± 0.917	5.055 ± 1.033
M48 MSE	6.192 ± 2.327	6.337 ± 2.487	5.133 ± 1.499	5.065 ± 1.446	5.103 ± 1.527	$\mathbf{4.968 \pm 1.339}$	5.404 ± 1.802

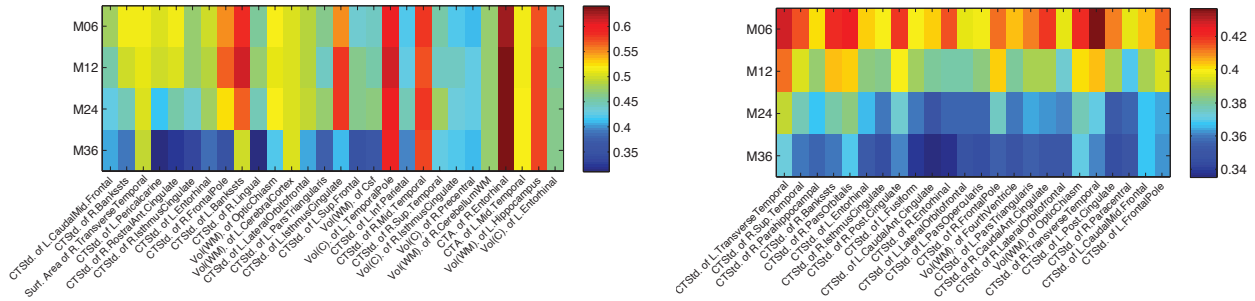


(a) Target: ADAS-Cog (25 stable features)

(b) Target: MMSE (33 stable features)

Figure 1: The stability vector of stable MRI features using Convex Fused Sparse Group Lasso (cFSGL).

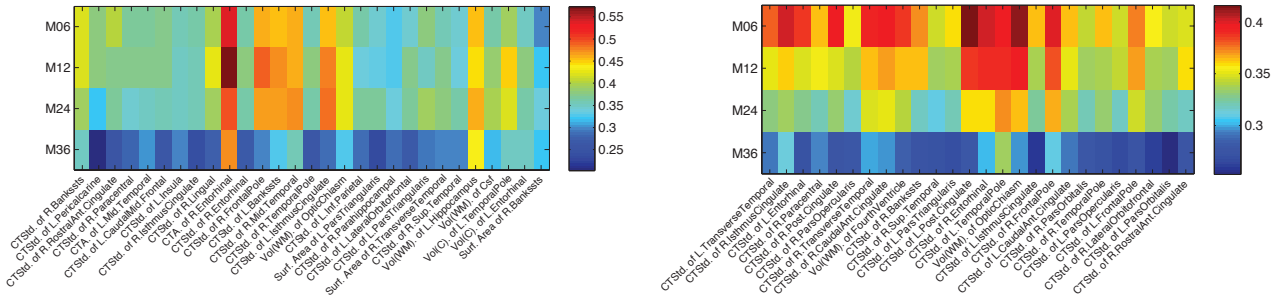
contain information about the progression in later stages. The lacking of predictable MRI biomarkers in later stages is a potential factor that contributes to the lower predictive performance of MMSE than that of ADAS-Cog in our study and other related studies [39]. These results suggest that ADAS-Cog may be a better cognitive measurement for lon-



(a) Target: ADAS-Cog (28 stable features)

(b) Target: MMSE (24 stable features)

Figure 2: The stability vector of stable MRI features using Non-Convex Fused Sparse Group Lasso (nFSGL1).



(a) Target: ADAS-Cog (29 stable features)

(b) Target: MMSE (26 stable features)

Figure 3: The stability vector of stable MRI features using Non-Convex Fused Sparse Group Lasso (nFSGL2).

gitudinal study. 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 in [39] may lead to sub-optimal performance.

We also perform stability selection of nFSGL1 and nFSGL2 using only MRI biomarkers. The results are given in Figure 2 and Figure 3. We observe that most biomarkers identified in cFSGL are also included in the top feature lists in nFSGL. This demonstrates the consistency between these two approaches. We also observe that the patterns of temporal selection stability differ from that of cFSGL in that fewer features have high probability. In nFSGL2 there is only one feature, namely cortical thickness average of right Entorhinal cortex, that has high probability at all time points, compared to 5 in cFSGL longitudinal stability selection. In nFSGL2 we observe that white matter volume of left Hippocampus also maintains a high stability vector. The higher temporal sparsity observed in nFSGL may be due to the non-convex $\ell_{(0.5,1)}$-norm penalty.

7. CONCLUSION

In this paper, we propose a convex fused sparse group Lasso (cFSGL) formulation for modeling disease progression. cFSGL allows the simultaneous selection of a common set of biomarkers for multiple time points and specific sets of biomarkers for different time points using the sparse group Lasso penalty and at the same time incorporates the temporal smoothness using the fused Lasso penalty. We show that the proximal operator associated with the optimization problem exhibits a certain decomposition property and thus can be solved effectively. To further improve the model, we propose two non-convex formulations, which are expected to reduce the shrinkage bias in the convex formulation. We employ the difference of convex (DC) programming technique to solve the non-convex formulations. The effectiveness of the proposed progression models is evaluated by extensive experimental studies on data sets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data sets. Results show that the proposed progression models are more effective than an existing multi-task learning formulation for disease progression. We also perform longitudinal stability selection to identify and analyze the temporal patterns of biomarkers for MMSE and ADAS-Cog respectively. The presented analysis can potentially provide novel insights into the AD progression.

Our proposed formulations for disease progression assume that the training data is complete, i.e., there are no missing values in the feature matrix X. We plan to extend our formulations to deal with missing data.

Acknowledgments

This work was supported in part by NIH R01 LM010730, NSF IIS-0812551, IIS-0953662, MCB-1026710, and CCF-1025177.

8. REFERENCES

[1] `www.public.asu.edu/~jye02/FSGL`.

[2] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853, 2005.

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

[4] B. Bakker and T. Heskes. Task clustering and gating for bayesian multitask learning. *The Journal of Machine Learning Research*, 4:83–99, 2003.

[5] E. Candes, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

[6] A. Caroli, G. Frisoni, et al. The dynamics of alzheimer's disease biomarkers in the alzheimer's disease neuroimaging initiative cohort. *Neurobiology of aging*, 31(8):1263–1274, 2010.

[7] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 137–144. ACM, 2009.

[8] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. In *Proceedings of the 23rd international conference on Machine learning*, pages 201–208. ACM, 2006.

[9] R. Desikan, H. Cabral, F. Settecase, C. Hess, W. Dillon, C. Glastonbury, M. Weiner, N. Schmansky, D. Salat, B. Fischl, et al. Automated mri measures predict progression to alzheimer's disease. *Neurobiology of aging*, 31(8):1364–1374, 2010.

[10] S. Duchesne, A. Caroli, C. Geroldi, D. Collins, and G. Frisoni. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *NeuroImage*, 47(4):1363–1370, 2009.

[11] T. Evgeniou, C. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(1):615, 2006.

[12] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[13] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[14] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv:1001.0736*, 2010.

[15] G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *Signal Processing, IEEE Transactions on*, 57(12):4686–4698, 2009.

[16] D. Holland, J. Brewer, D. Hagler, C. Fennema-Notestine, A. Dale, M. Weiner, L. Thal, R. Petersen, C. Jack, W. Jagust, et al. Subregional neuroanatomical change as a biomarker for alzheimer's disease. *Proceedings of the National Academy of Sciences*, 106(49):20954, 2009.

[17] R. Horst and N. Thoai. Dc programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

[18] K. Ito et al. Disease progression model for cognitive deterioration from Alzheimer's Disease Neuroimaging Initiative database. *Alzheimer's and Dementia*, 6(1):39–53, 2010.

[19] C. Jack Jr, D. Knopman, W. Jagust, L. Shaw, P. Aisen, M. Weiner, R. Petersen, and J. Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[20] L. Jacob, F. Bach, and J. Vert. Clustered multi-task learning: A convex formulation. *Advances in Neural Information Processing Systems*, 2008.

[21] Z. Khachaturian. Diagnosis of Alzheimer's disease. *Archives of Neurology*, 42(11):1097, 1985.

[22] A. Kumar and S. Zilberstein. Message-passing algorithms for quadratic programming formulations of MAP estimation. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 428–435, Barcelona, Spain, 2011.

[23] Q. Ling, Z. Wen, and W. Yin. Decentralized jointly sparse signal recovery by reweighted ℓ_q minimization. submitted (2011).

[24] J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 323–332. ACM, 2010.

[25] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[26] C. Misra, Y. Fan, and C. Davatzikos. 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, 2009.

[27] A. Nemirovski. Efficient methods in convex programming. 2005.

[28] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Netherlands, 2004.

[29] G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2006.

[30] R. Pearson, R. Kingan, and A. Hochberg. Disease progression modeling from historical clinical databases. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 788–793. ACM, 2005.

[31] C. Stonnington, C. Chu, S. Klöppel, C. Jack Jr, J. Ashburner, and R. Frackowiak. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *NeuroImage*, 51(4):1405–1413, 2010.

[32] S. Thrun and J. O'Sullivan. Clustering learning tasks and the selective cross-task transfer of knowledge. *Learning to learn*, pages 181–209, 1998.

[33] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[34] P. Vemuri et al. MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. *Neurology*, 73(4):294, 2009.

[35] K. Walhovd et al. Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer disease. *American Journal of Neuroradiology*, 31(2):347, 2010.

[36] A. Wimo, B. Winblad, H. Aguero-Torres, and E. von Strauss. The magnitude of dementia occurrence in the world. *Alzheimer Disease & Associated Disorders*, 17(2):63, 2003.

[37] C. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176. ACM, 2009.

[38] A. Yuille and A. Rangarajan. The concave-convex procedure (cccp). *Advances in Neural Information Processing Systems*, 2:1033–1040, 2002.

[39] D. Zhang and D. Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage*, 2011.

[40] Y. Zhang and D.-Y. Yeung. Multi-task learning using generalized t process. 2010.

[41] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning: A convex formulation. *Advances in Neural Information Processing Systems*, 2011.

[42] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.
`www.public.asu.edu/~jye02/Software/MALSAR`

[43] J. Zhou, L. Yuan, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822. ACM, 2011.