A Multi-Task Learning Formulation for Predicting Disease Progression

Jiayu Zhou, Lei Yuan, Jun Liu, Jieping Ye
Computer Science and Engineering, Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, AZ 85287
{Jiayu.Zhou, Lei.Yuan, Jun.Liu, Jieping.Ye}@asu.edu

ABSTRACT

Alzheimer's Disease (AD), the most common type of dementia, is a severe neurodegenerative disorder. Identifying markers that can track the progress of the disease has recently received increasing attentions in AD research. 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 propose a multi-task learning formulation for predicting the disease progression measured by the cognitive scores and selecting markers 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. We capture the intrinsic relatedness among different tasks by a temporal group Lasso regularizer. The regularizer consists of two components including an $\ell_{2,1}$-norm penalty 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. We have performed extensive evaluations using various types of data at the baseline from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database for predicting the future MMSE and ADAS-Cog scores. Our experimental studies demonstrate the effectiveness of the proposed algorithm for capturing the progression trend and the cross-sectional group differences of AD severity. Results also show that most markers selected by the proposed algorithm are consistent with findings from existing cross-sectional studies.

Categories and Subject Descriptors

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

KDD'11, August 21–24, 2011, San Diego, California, USA.

General Terms

Algorithms

Keywords

Alzheimer's Disease, regression, multi-task learning, group Lasso, stability selection, cognitive score

1. INTRODUCTION

Alzheimer's disease (AD), the most common type of dementia, is characterized by the progressive impairment of neurons and their connections resulting in loss of cognitive function and ultimately death [20]. AD currently affects about 5.3 million individuals in United States and more than 30 million worldwide with a significant increase predicted in the near future [5]. Alzheimer's disease has been not only the substantial financial burden to the health care system but also the psychological and emotional burden to patients and their families. As the research on developing promising new treatments to slow or prevent AD progressing, the need for markers that can track the progress of the disease and identify it early becomes increasingly urgent.

A definitive diagnosis of AD can only be made through an analysis of brain tissue during a brain biopsy or autopsy [18]. Many clinical/cognitive measures have been designed to evaluate the cognitive status of the patients and used as important criteria for clinical diagnosis of probable AD, such as Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog) [25]. MMSE has been shown to be correlated with the underlying AD pathology and progressive deterioration of functional ability [18]. ADAS-Cog is the gold standard in AD drug trial for cognitive function assessment [31]. Since neurodegerenation 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 address two major research questions: (1) how can we predict the progression of the disease measured by cognitive scores, e.g., MMSE and ADAS-Cog? (2) what is the smallest set of features (measurements) most predictive of the progression? The prime candidate markers for tracking disease progression include neuroimages such as magnetic resonance imaging (MRI), cerebrospinal fluid (CSF), and baseline clinical assessments [12].

The relationship between the cognitive scores and possible risk factors such as age, APOE gene, years of education and gender has been previously studied [36, 17]. Many existing works analyzed the relationship between cognitive scores and

imaging markers based on MRI such as gray matter volumes, density and loss [3, 8, 15, 16, 33], shape of ventricles [14, 34] and hippocampal [34] by correlating these features with baseline MMSE scores. In [13], the intensity and volume of medial temporal lobe altogether with other risk factors and the gray matter were shown to be correlated with the 6-month MMSE score, which allowed us to predict near-future clinical scores of patients. Relations between 6-month atrophy patterns in medial temporal region and memory declination in terms of clinical scores had also been examined in [27]. 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 [4]. In [10], 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 by Ito *et al.*, the progression rate of cognitive scores was modeled using power functions [17].

Most existing work employed either the regression model [13, 33] or the survival model [37] for modeling the disease progression. The correlation between the ground truth and the prediction is used to evaluate the model [13, 33]. When the size of covariates is small, each covariate can be individually added to the model to examine its effectiveness for predicting the target [17, 38], or univariate analysis is performed individually on all covariates and those who exceed a certain significance threshold are included in the model [27]. When the number of covariates is large and significant correlations among covariates exist, these approaches are suboptimal. To deal with the curse of dimensionality, dimension reduction techniques are commonly employed. Duchesne *et al.* used principle components analysis (PCA) to build a low dimensional feature space from image data [13]. An obvious disadvantage of dimension reduction techniques such as PCA is that the model is no longer interpretable, since all features are involved. Stonnington *et al.* used relevance vector regression (RVR), which integrated feature selection in the training stage [33]. These approaches only predict clinical scores at a single time point and their performances are far from satisfactory to be clinically useful for AD prognosis.

In this paper, we propose a multi-task learning formulation for predicting the progression of the disease measured by the clinical scores at multiple time points and simultaneously selecting markers predictive of the progression. 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. Multi-task learning aims at improving the generalization performance by learning multiple related tasks simultaneously. The key of multi-task learning is to exploit the intrinsic relatedness among the tasks. 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. The regularizer consists of two components including an $\ell_{2,1}$-norm penalty [39] 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.

We have performed extensive experimental studies to evaluate the effectiveness of the proposed algorithm. We use various types of data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database including MRI scans, CSF, and clinical scores at the baseline to predict the MMSE and ADAS-Cog scores for the next three years. Our experimental studies show that the proposed algorithm better captures the progression trend and the cross-sectional group differences of AD severity than existing methods. Results also show that most markers selected by the proposed algorithm are consistent with findings from existing cross-sectional studies.

2. PROPOSED MULTI-TASK REGRESSION FORMULATION

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 progression of clinical scores as a multi-task regression problem. We employ the multi-task regression formulation 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 $\{x_1, x_2, \cdots, x_n\}$ be the input data at the baseline, and $\{y_1, y_2, \cdots, y_n\}$ be the targets, where each $x_i \in \mathbb{R}^d$ represents a sample (patient), and $y_i \in \mathbb{R}^t$ is the corresponding targets (clinical scores) at different time points. In this paper we employ linear models for the prediction. Specifically, the prediction model for the ith time point is given by $f^i(x) = x^T w^i$, where w^i is the weight vector of the model. Let $X = [x_1, \cdots, x_n]^T \in \mathbb{R}^{n\times d}$ be the data matrix, $Y = [y_1, \cdots, y_n]^T \in \mathbb{R}^{n\times t}$ be the target matrix, and $W = \left[w^1, w^2, ..., w^t\right] \in \mathbb{R}^{d\times t}$ be the weight matrix. One simple approach is to estimate W by minimizing the following objective function:

$$\min_W \|XW - Y\|_F^2 + \theta_1 \|W\|_F^2, \tag{1}$$

where the first term measures the empirical error on the training data, the second (penalty) term controls the generalization error, $\theta_1 > 0$ is a regularization parameter, and $\|.\|_F$ is the Frobenius norm of a matrix. The regression method above is known as the *ridge regression* and it admits an analytical solution given by:

$$W = (X^T X + \theta_1 I)^{-1} X^T Y. \tag{2}$$

One major limitation of the regression model 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.

2.1 Temporal Smoothness Prior

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 of predictions at neighboring time points, resulting in the fol-

lowing formulation:

$$\min_W \|XW - Y\|_F^2 + \theta_1 \|W\|_F^2 + \theta_2 \sum_{i=1}^{t-1} \left\|w^i - w^{i+1}\right\|_2^2, \quad (3)$$

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

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

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.(3) becomes:

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

The optimization problem in Eq.(4) admits an analytical solution, as shown below. First, we take the derivative of Eq.(4) with respect to W and set it to zero:

$$X^TXW - X^TY + \theta_1 W + \theta_2 WHH^T = 0, \quad (5)$$
$$\left(X^TX + \theta_1 I_d\right) W + W\left(\theta_2 HH^T\right) = X^TY, \quad (6)$$

where I_d is the identity matrix of size d by d. Since both matrices $(X^TX+\theta_1 I_d)$ and $\theta_2 HH^T$ are symmetric, we write the eigen-decomposition of these two matrices by $Q_1\Lambda_1Q_1^T$ and $Q_2\Lambda_2Q_2^T$, where $\Lambda_1 = diag(\lambda_1^{(1)}, \lambda_1^{(2)}, \ldots, \lambda_1^{(d)})$ and $\Lambda_2 = diag(\lambda_2^{(1)}, \lambda_2^{(2)}, \ldots, \lambda_2^{(d)})$, are their eigenvalues, and Q_1 and Q_2 are orthogonal. Plugging them into Eq. (6) we get:

$$Q_1\Lambda_1Q_1^TW + WQ_2\Lambda_2Q_2^T = X^TY, \quad (7)$$
$$\Lambda_1Q_1^TWQ_2 + Q_1^TWQ_2\Lambda_2 = Q_1^TX^TYQ_2. \quad (8)$$

Denote $\hat{W} = Q_1^TWQ_2$ and $D = Q_1^TX^TYQ_2$. Eq.(8) becomes $\Lambda_1\hat{W} + \hat{W}\Lambda_2 = D$. Thus $\hat{W}$ is given by:

$$\hat{W}_{i,j} = \frac{D_{i,j}}{\lambda_1^{(i)} + \lambda_2^{(j)}}. \quad (9)$$

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

2.2 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 to extend the formulation in Eq.(4) with missing target values in the training process. In this case, the analytical solution to Eq.(4) no longer exists. We show how the algorithm above can be adapted to deal with missing target values.

We use a matrix $S \in \mathbb{R}^{n\times t}$ 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.(4) can be extended to the case with missing target values as:

$$\min_W \|S \odot (XW - Y)\|_F^2 + \theta_1 \|W\|_F^2 + \theta_2 \|WH\|_F^2. \quad (10)$$

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 $y_{(i)} = \mathcal{P}_{S^i}(Y^i) \in \mathbb{R}^{n_i\times 1}$ as the corresponding target vector, where n_i is number of samples from the ith task.

Similar to the case without missing target values considered in Section 2.1, we take the derivative of Eq.(10) with respect to w^i $(2 \leq i \leq t-1)$ and set it to zero:

$$Aw^{i-1} + M_iw^i + Aw^{i+1} = T_i, \quad (11)$$

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

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

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

$$\begin{pmatrix} M_1 & A & & & 0 \\ A & M_2 & A & & \\ & \ddots & & & \\ & & A & M_{t-1} & A \\ 0 & & & A & M_t \end{pmatrix} \begin{pmatrix} w^1 \\ w^2 \\ \vdots \\ w^{t-1} \\ w^t \end{pmatrix} = \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_{t-1} \\ T_t \end{pmatrix} \quad (12)$$

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 $O(d^3t)$ using block Gaussian elimination. For large-scale linear systems, the LSQR algorithm [30], 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.

2.3 Temporal Group Lasso Regularization

Because of the limited availability of subjects in the longitudinal AD study and a relatively large number of features at ADNI including MRI features, 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 [39], 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 formulation:

$$\min_W \|S\odot(XW - Y)\|_F^2 + \theta_1\|W\|_F^2 + \theta_2\|WH\|_F^2 + \delta\|W\|_{2,1} \quad (13)$$

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 [35]. 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.(13) can be considered as a combination of a smooth term and a non-smooth term. The gradient descent or accelerated gradient method (AGM) [29, 28] 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 [22], which computes the proximal operator associated with the general ℓ_1/ℓ_q-norm efficiently.

2.3.1 Longitudinal Stability Selection

An important issue in the practical application of the proposed formulation is the selection of an appropriate amount of regularization, known as model selection. Cross validation is commonly used for model selection, however it tends to select more features than needed [26]. In this paper, we adapt stability selection to perform model selection for the proposed multi-task regression. Stability selection is a method recently proposed to address the problem of proper regularization using subsampling/bootstrapping [26]. It should be noted that in our formulation we find in our experiments that the list of top features selected by stability selection is not sensitive to the regularization parameters θ_1 and θ_2. We thus focus on the selection of δ, which controls the sparsity of the model, in stability selection.

Let K be the index set of features, i.e., $k \in K$ denotes a feature. Given a set of regularization parameter values Δ and an iteration number γ, longitudinal stability selection works as follows. Let $B_{(j)} = \{B^X_{(j)}, B^Y_{(j)}\}$ be a random subsample from $\{X, Y\}$ of size $\lfloor n/2 \rfloor$ without replacement. For a given $\delta > 0$, let $\tilde{W}^{(j)}$ be the optimal solution of Eq.(13) on $B_{(j)}$. Denote $U^\delta(B_{(j)}) = \{k : \tilde{W}^{(j)}_k \neq 0\}$ as the set of features selected by the model $\tilde{W}^{(j)}$. This process is repeated for γ times and selection probability $\hat{\Pi}^\delta_k$ of each feature k is given by $\sum_{j=1}^{\gamma} I(k \in U^\delta(B_{(j)}))/\gamma$, where $I(.)$ is the indicator function defined as follows: $I(c) = 1$ if c is true and $I(c) = 0$ otherwise. It is clear that $\hat{\Pi}^\delta_k$ computes the fraction of bootstrap experiments for which the feature k is selected. Repeat the above procedure for all $\delta \in \Delta$, and we define the stability score for each feature k by $\mathcal{S}(k) = \max_{\delta \in \Delta}(\hat{\Pi}^\delta_k)$. To find a suitable size of stable feature set $\hat{U}^{\text{stable}}$ we can either use top η stable features:

$$\hat{U}^{\text{stable}} = \{k : \mathcal{S}(k) \text{ ranks among top } \eta \text{ in } K\},$$

or use threshold π_{thr} on the stability score:

$$\hat{U}^{\text{stable}} = \{k : \mathcal{S}(k) \geq \pi_{thr}\}.$$

In our application we choose the top η features. Indeed, cross validation can be performed to determine how many features are needed. However, our empirical results show that using top $\eta = 20$ features is sufficient in most cases of our application.

2.4 Proposed Algorithm

One undesired property of sparse learning methods such as Lasso is that the coefficients corresponding to relevant features are shrunk towards zero [40]. This shrinkage effect would lead to sub-optimal performance. To resolve this problem, existing methods apply adaptive regularization [42], multiple thresholding procedures [41], or multi-stage methods [23, 40]. In this paper, we employ a standard two-stage procedure. In the first stage the algorithm selects features using longitudinal stability selection, resulting in a subset of features $\hat{U}^{stable}$. In the second stage the algorithm performs temporal smoothness regularized regression using selected features. We summarize the proposed algorithm in Algorithm 1.

Algorithm 1 Temporal Group Lasso Multi-Task Regression (TGL)

Input: S, X, Y, θ_1, θ_2 Δ, γ, η
Output: W^*, $\hat{U}^{\text{stable}}$

Stage 1: longitudinal stability selection
Set $K = \{$the feature set in $X\}$
for $\delta \in \Delta$ **do**
  for $j = 1$ to γ **do**
    Subsample $B_{(j)} = \{B^X_{(j)}, B^Y_{(j)}\}$ from $\{X, Y\}$
    Compute $\tilde{W}^{(j)}$ by solving Eq.(13) with $\delta, B_{(j)}$
    Set $U^\delta(B_{(j)}) = \{k : \tilde{W}^{(j)} \neq 0\}$
  end for
  Calculate $\hat{\Pi}^\delta_k = \sum_{j=1}^{\gamma} I(k \in U^\delta(B_{(j)}))/\gamma$, $\forall k \in K$
end for
Calculate $\mathcal{S}(k) = \max_{\delta \in \Delta}(\hat{\Pi}^\delta_k)$, $\forall k \in K$
Set $\hat{U}^{\text{stable}} = \{k$: $\mathcal{S}(k)$ ranks among top η in $K\}$
Stage 2: temporal smoothness regularized regression
Set $\hat{X} = X$ restricted to the features from $\hat{U}^{\text{stable}}$
if $\exists p, q$ such that $S_{p,q} = 0$ **then**
  Set $t =$ number of tasks, $d = |\hat{U}^{\text{stable}}|$, $A = -\theta_2 I_d$
  for $i = 1$ to t **do**
    $X_{(i)} = \mathcal{P}^i_S(\hat{X})$, $y_{(i)} = \mathcal{P}^i_S(Y^i)$
    $M_i = X^T_{(i)} X_{(i)} + \theta_1 I_d + 2\theta_2 I_d$, $T_i = X^T_{(i)} y_{(i)}$
  end for
  Obtain W^* by solving Eq.(12) with $\{M_i\}, \{T_i\}$.
else
  Compute the analytical solution W^* as in Section 2.1.
end if

3. EXPERIMENTS

In this section we evaluate the proposed algorithm on the ADNI database[1]. The source codes are available online [1].

3.1 Experimental Setup

In the ADNI project, MRI scans, CSF measurements, and clinical scores from selected patients are obtained repeatedly over a 6-month or 1-year interval. We denote each time point by the duration starting from the baseline when the patient came to the hospital for screening. For instance, M06 indicates 6 months after the baseline. We use different combinations of MRI (M), CSF (C), and META (M) (see Table 1) at baseline to predict MMSE and ADAS-Cog scores at four time points: M06, M12, M24, and M36.

For MRI, we download 1.5T MRI data of 675 patients preprocessed by UCSF using FreeSurfer. MRI features can be grouped into 5 categories: cortical thickness average (CTA), cortical thickness standard deviation (CTStd), volume of cortical parcellation (Vol. Cort.), volume of white matter parcellation (Vol. WM.), and surface area (Surf. A.). There are 313 MRI features in total. We remove all samples which fail the MRI quality controls. For other feature

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

Table 1: Features included in the META dataset. Note that the cognitive scores at the baseline are used to predict the future cognitive scores. A detailed explanation of each cognitive score and lab test can be found at [1].

Type	Details
Demographic	age, yrs. of education, gender
Genetic	ApoE-$\varepsilon 4$ information
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

sets, we remove samples with missing entries. Each feature combination includes the intersection of available samples, that is the combined dataset has no missing values. Because changes of MMSE and ADAS-Cog are both found to be closely correlated with the baseline MMSE score [17], we include the baseline MMSE score in all feature combinations. With the pre-processing procedure described above, the number of samples available for MMSE and ADAS-Cog at different time points are described in Table 2.

Table 2: The first column of the table indicates the different combinations of three datasets. The numbers of samples for different combinations available for both MMSE and ADAS-Cog at different time points are shown in the table together with the data dimensionality (Dim). C, E, and M refer to CSF, META, and MRI, respectively.

	MMSE				ADAS-Cog				Dim
	M06	M12	M24	M36	M06	M12	M24	M36	
C	332	331	295	198	332	328	294	194	6
E	648	641	567	387	646	636	563	377	52
CE	332	331	295	198	332	328	294	194	57
M	648	641	567	387	646	636	563	377	306
EM	648	641	567	387	646	636	563	377	357
CM	332	331	295	198	332	328	294	194	311
CEM	332	331	295	198	332	328	294	194	362

3.2 Prediction Performance

In this experiment, we compare our proposed approach with ridge regression on the prediction of MMSE and ADAS-Cog using various types of data combinations. We report the results based on the leave-one-out scheme due to the small sample size. 5-fold cross validation is used to select model parameters (θ_1, θ_2 in our approach and θ_1 in ridge regression) in the training data (between 10^{-3} and 10^3). To compare with related works [33, 13], we use correlation coefficient (R-value) given by the correlation between the predicted values and the ground truth, and its correlation significance (P-value) as the evaluation criteria. P-value is given by testing the hypothesis of no correlation [2]. A good prediction model has a high R-value and low P-value (less than 0.0001 for example). The overall performance of the prediction is averaged across all time points weighted by the sample size.

The results for MMSE and ADAS-Cog are summarized in Tables 3 and 4, respectively. Overall, the proposed approach significantly outperforms ridge regression. This is especially the case when the feature space is large and the sample size is small, e.g., CM and CEM (see Table 2). Note that multi-task learning effectively increases the number of samples by learning multiple related tasks simultaneously, while ridge regression treats all tasks independently. Our results verify the effectiveness of multi-task learning for disease progression.

Table 3: Comparison of our proposed approach (TGL) and ridge regression (RidR) on longitudinal MMSE prediction using different data combinations measured by R-value and average P-value. Performance at each time point is reported in terms of R-value. Weighted averages of R-values and P-values across all time points are also given.

		M06	M12	M24	M36	AvgR	P-Val
C	TGL	0.74	0.72	0.71	0.65	0.71	3.3e-026
	RidR	0.74	0.72	0.71	0.65	0.71	3.3e-026
E	TGL	0.83	0.84	0.84	0.78	0.82	7.4e-080
	RidR	0.82	0.83	0.84	0.76	0.82	2.8e-075
CE	TGL	0.80	0.82	0.82	0.72	0.80	7.2e-034
	RidR	0.77	0.80	0.79	0.71	0.77	7.8e-033
M	TGL	0.77	0.75	0.77	0.70	0.75	2.7e-059
	RidR	0.68	0.67	0.71	0.63	0.68	1.2e-045
EM	TGL	0.83	0.84	0.85	0.80	0.83	3.8e-088
	RidR	0.71	0.71	0.74	0.54	0.69	7.1e-032
CM	TGL	0.73	0.73	0.76	0.69	0.73	1.3e-030
	RidR	0.25	0.34	0.36	0.48	0.35	1.0e-006
CEM	TGL	0.80	0.83	0.83	0.72	0.80	8.4e-034
	RidR	0.29	0.36	0.59	0.56	0.43	2.4e-008

Table 4: Comparison of our proposed approach (TGL) and ridge regression (RidR) on longitudinal ADAS-Cog prediction using different data combinations measured by R-value and average P-value.

		M06	M12	M24	M36	AvgR	P-Val
C	TGL	0.70	0.71	0.71	0.66	0.70	1.7e-026
	RidR	0.70	0.71	0.70	0.66	0.70	1.5e-026
E	TGL	0.87	0.87	0.86	0.83	0.86	2.6e-099
	RidR	0.87	0.86	0.86	0.83	0.86	4.1e-097
CE	TGL	0.86	0.85	0.86	0.80	0.85	1.7e-044
	RidR	0.86	0.84	0.85	0.74	0.83	2.7e-035
M	TGL	0.75	0.78	0.76	0.74	0.76	4.4e-066
	RidR	0.66	0.71	0.67	0.61	0.67	4.9e-041
EM	TGL	0.87	0.87	0.86	0.84	0.86	3.2e-104
	RidR	0.80	0.81	0.80	0.70	0.79	2.4e-058
CM	TGL	0.75	0.77	0.75	0.74	0.76	2.5e-035
	RidR	0.55	0.60	0.62	0.53	0.58	5.4e-016
CEM	TGL	0.87	0.86	0.87	0.81	0.86	1.9e-046
	RidR	0.69	0.66	0.63	0.40	0.62	8.6e-010

In Table 5 we compare our approach with three closely related works. In [13], PCA was employed to reduce the dimensionality of the MRI data. Since only limited samples were used, the proposed method resulted in low prediction performance. To the best of our knowledge, the work in [13] is the only one that tried to predict cognitive scores at future time points and reported R/P-Value for comparison. The other two works used the ADNI database as in our study. Stonnington's work predicted baseline cognitive scores using relevance vector regression (RVR) which enforced sparsity in the model [33]. Their regression results were better than those reported in [13]. However their prediction performance is much lower than ours. This is possibly due

to the use of different prediction models. Specifically, the model in [33] did not utilize rich temporal smoothness information as in the proposed approach. In [17], very limited risk factors were manually included and tested in their progression model. No actual prediction on the test data was performed. Instead, a visual comparison of the proposed model for different groups was provided, which was similar to our cross-sectional analysis in Section 3.3. It is hard to directly compare their approach with ours, since no prediction measures were reported in [17].

3.3 Cross-Sectional Progression

In this experiment, we study the cross-sectional progression patterns. We predict the progression of different patient groups using models built by ridge regression and our proposed approach on the CEM dataset. The predicted progressions for different groups including AD, MCI, and NL[2] are visualized in a cross-sectional fashion with the mean values of prediction by different methods and the group truth provided. The results are summarized in Figure 1 and Figure 2. For MMSE, the target progression in probable AD patients shows a significant declination pattern, with an annual declination of 3.167 ± 0.40. For MCI patients we observe a slight annual declination of 0.919 ± 0.188. For normal controls, the MMSE scores are very stable with an annual declination of 0.032 ± 0.025. A similar pattern can be observed for ADAS-Cog. Probable AD patients have an annual growth on ADAS-Cog of 7.066 ± 2.257, while for the MCI group the annual growth is 1.558 ± 0.401. In normal controls, we observe a very small annual change (0.007 ± 0.565).

We can also observe from Figure 1 and Figure 2 that the proposed approach can better capture the characteristics of the disease progression measured by clinical scores for different groups of patients. The proposed method performs significantly better than ridge regression for the probable AD group and the MCI group at M36 where the number of training samples is small. The result further verifies the effectiveness of multi-task learning for disease progression especially for small sample size problems.

3.4 Feature Evaluation

In this experiment we evaluate the relevance of features selected by longitudinal stability selection on different feature combinations. Given the range of regularization parameter values Δ, the set of selected features is a function of input data X and target Y, thus for the same data set X but different targets (clinical scores) Y, the algorithm may select different features. Indeed, although different clinical scores are used to measure the cognition status, they may emphasize different aspects of the AD pathology.

The top MRI markers selected by longitudinal stability selection are shown in Figure 3. We observe that volume of left hippocampus, CTA of middle temporal gyri and CTA 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 cortex) is firstly affected, followed by progressive neocortical damage [7, 11]. Evidence of a significant atrophy of middle temporal gyri in AD patients has also been observed in previous studies [9, 19, 3]. 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 [24]. In addition, CTA 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 [21], which shows that parietal atrophy contributes to predictive values for diagnosing AD.

[2]AD, MCI, and NL refer to Alzheimer's Disease patients, Mild Cognitive Impairment patients, and normal controls, respectively.

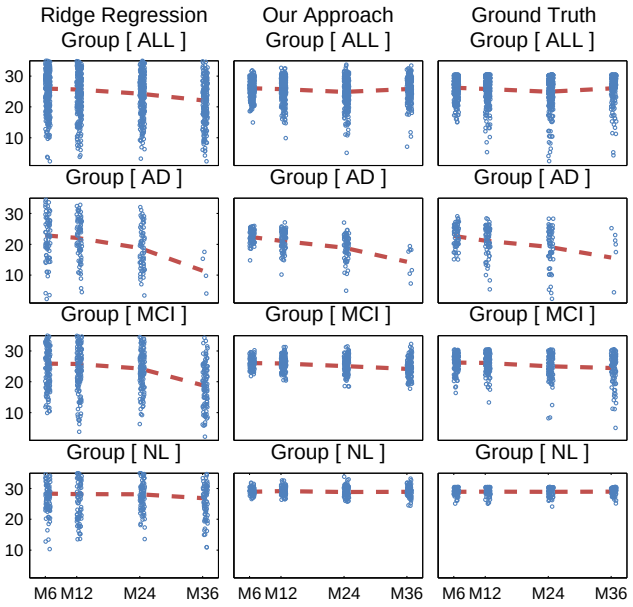

Figure 1: Visual predictive comparison for MMSE on the CEM dataset. The first, second, third, and fourth rows show the results for the complete dataset (ALL), the AD group (AD), the MCI group (MCI), and the NL group (NL), respectively. The left column shows ridge regression results. The middle column shows the results using our proposed approach. The ground truth values are plotted in the right column. The red line connects the mean values of each time point.

To access the effectiveness of CSF features, we perform longitudinal stability selection on the data combining CSF and MRI (Figure 4). In CSF, we use 3 simple features (Aβ-42, t-tau and p-tau) and 2 ratio features (t-tau/Aβ-42 and p-tau/Aβ-42). We observe that Aβ-42 and p-tau are among the top three markers, whereas t-tau is not selected in both targets.

Figure 5 shows the stability results of META features (see Table 1). Baseline clinical test scores (MMSE, ADAS-MOD, ADAS-Cog, ADAS-Cog subscores, CDR) and neuropsychological test scores (TRAASCORE, GATVEGESC) are found to be important features for both scores. Among ADAS-Cog subscores, *Word Recall* (subscore 1) and *Orientation* (subscore 7) are found to be important. These findings agree with recent itemized ADAS-Cog analysis, where orientation was found to be the most sensitive item in cross-sectional study of cognitive impairment [32]. The results of stability selection after adding META to MRI are shown in Figure 6. Many MRI features have low stability scores. A possible explanation is that some clinical tests included in META

Table 5: Comparison of the proposed approach with three 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
Duchesne *et al.* [13]	M06 MMSE	75 NL, 49 MCI, 75 AD	Baseline MRI, age, gender, years of education	MMSE: 0.31 (p=0.03)
Stonnington *et al.* [33]	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)
Ito *et al.* [17]	M06-M36 ADAS-Cog	ADNI: 186 AD, 402 MCI, 229 NL	Age, APOE4, gender, family history, years of education	Only visual check. R/P value not report
Our approach	M06-M36 MMSE and ADAS-Cog	ADNI: 133 AD, 304 MCI, 188 NL	Baseline MRI, CSF, and META (see Table 1 for details)	Avg MMSE: 0.83 (p=3.8e-88) Avg ADAS-Cog: 0.86 (p=3.2e-104)

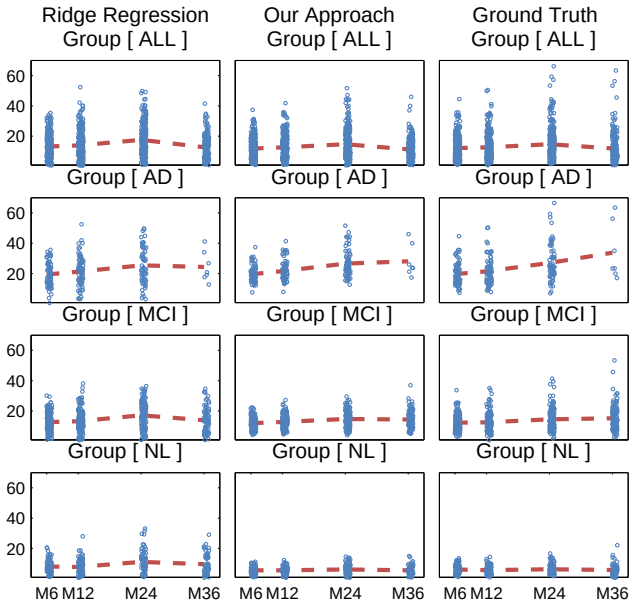

Figure 2: Visual predictive comparison for ADAS-Cog on the CEM dataset.

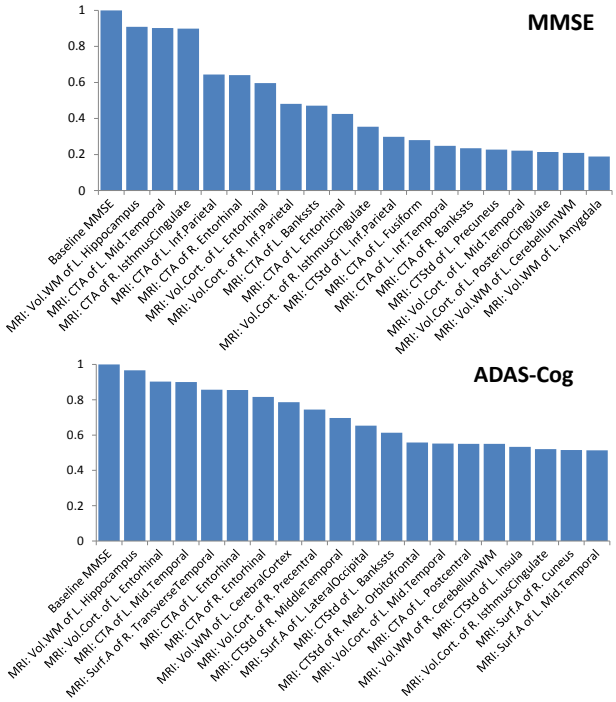

Figure 3: Longitudinal stability selection results for MMSE and ADAS-Cog on MRI data.

features reflect white matter and/or gray matter changes, as discussed in [6].

Next we perform stability selection by adding all three types of features together. Results are shown in Figure 7. We can observe from the figure that there are many features in common for both targets: baseline ADAS-Cog, ADAS-MOD and MMSE; baseline ADAS-Cog subscores 1 and 7; Logic LIMMTOTAL; APOE and DIGITSCOR.

4. CONCLUSION

In this paper, we study the feasibility of predicting AD progression measured by cognitive scores based on baseline measurements. Specifically, we formulate the progression prediction as a multi-task regression problem by considering the prediction of cognitive scores at each time point as a task. To capture the intrinsic relatedness among different tasks at different time points, we propose a temporal group Lasso regularizer, which ensures a small deviation between two regression models at successive time points based on a small subset of features. The effectiveness of the proposed approach in predicting disease progression and feature selection is evaluated by extensive experimental studies on the ADNI database. Results show that the proposed approach can better capture the progression trends than existing methods. Results also show that most features selected by the proposed approach are consistent with findings from existing cross-sectional studies. Our experimental studies demonstrate the promise of multi-task learning for predicting disease progression.

Our proposed formulation for AD progression prediction and feature selection is able to deal with missing values in the target matrix. Missing values in the input matrix, however, cannot be handled. Recent advances on matrix completion have allowed us to complete matrices with many missing values under certain conditions. We will explore such techniques for dealing with missing values in the input data. In addition, in this study we only focus on linear models; we plan to explore non-linear models in the future.

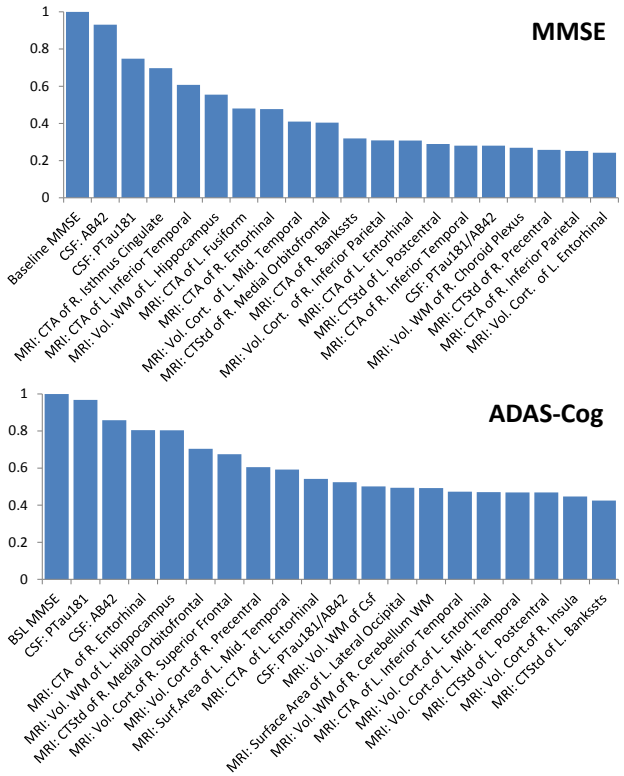

Figure 4: Longitudinal stability selection results for MMSE and ADAS-Cog on MRI+CSF data.

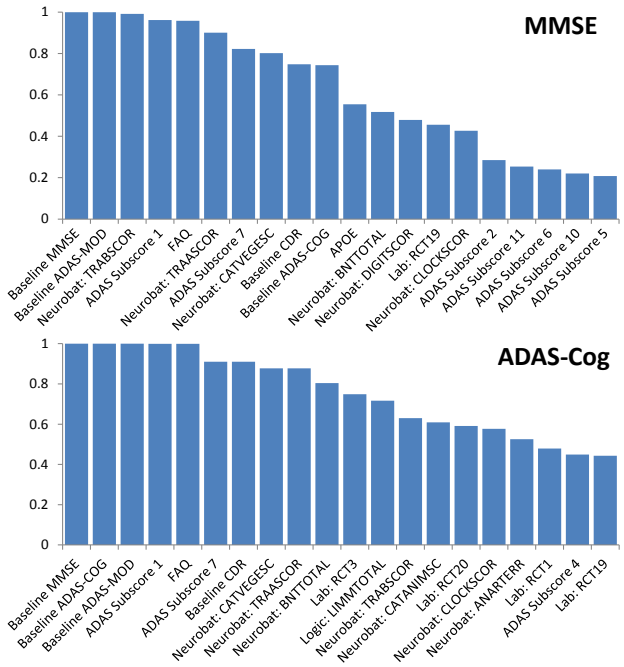

Figure 5: Longitudinal stability selection results for MMSE and ADAS-Cog on META data.

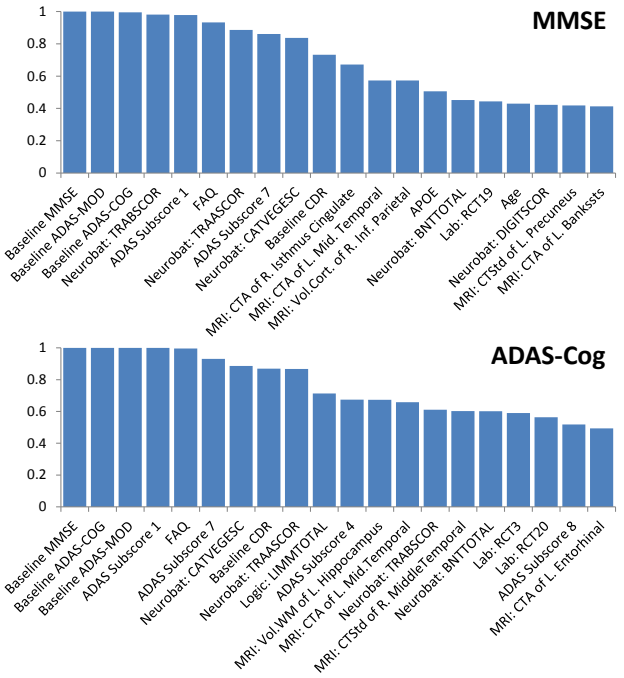

Figure 6: Longitudinal stability selection results for MMSE and ADAS-Cog on MRI+META data.

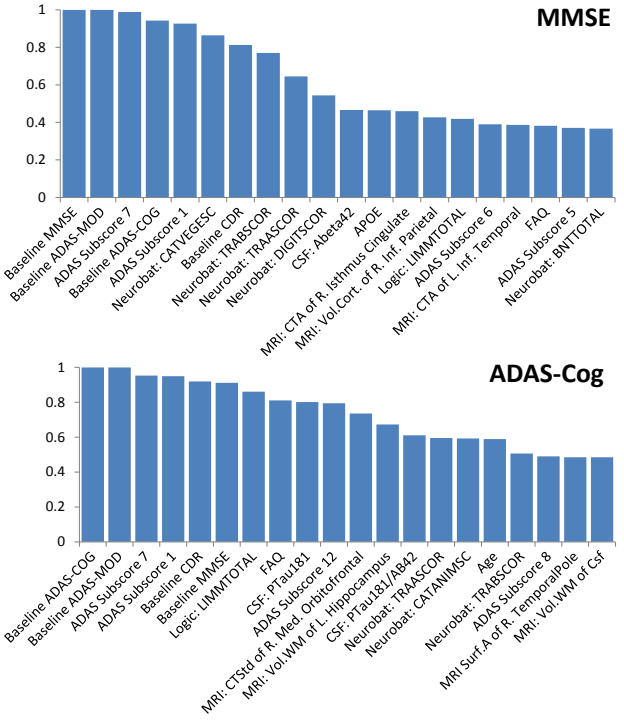

Figure 7: Longitudinal stability selection results for MMSE and ADAS-Cog on MRI+META+CSF data.

Acknowledgments

This research is sponsored in part by NSF IIS-0953662 and NSF CCF-1025177.

5. REFERENCES

[1] www.public.asu.edu/~jye02/AD-Progression.

[2] T. Anderson. *An introduction to multivariate statistical analysis*, volume 374. Wiley New York, 1958.

[3] L. Apostolova et al. 3D mapping of mini-mental state examination performance in clinical and preclinical Alzheimer disease. *Alzheimer Disease & Associated Disorders*, 20(4):224, 2006.

[4] J. Ashford and F. Schmitt. Modeling the time-course of Alzheimer dementia. *Current Psychiatry Reports*, 3(1):20–28, 2001.

[5] A. Association. 2010 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 6:158–194, 2010.

[6] L. Baxter et al. Relationship of cognitive measures and gray and white matter in Alzheimer's disease. *Journal of Alzheimer's Disease*, 9(3):253–260, 2006.

[7] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.

[8] G. Chetelat and J. Baron. Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. *Neuroimage*, 18(2):525–541, 2003.

[9] A. Convit et al. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. *Neurobiology of aging*, 21(1):19–26, 2000.

[10] C. Davatzikos, F. Xu, Y. An, Y. Fan, and S. Resnick. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain*, 132(8):2026, 2009.

[11] A. Delacourte et al. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. *Neurology*, 52(6):1158, 1999.

[12] B. Dubois et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *The Lancet Neurology*, 6(8):734–746, 2007.

[13] 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.

[14] L. Ferrarini et al. MMSE scores correlate with local ventricular enlargement in the spectrum from cognitively normal to Alzheimer disease. *Neuroimage*, 39(4):1832–1838, 2008.

[15] G. Frisoni et al. Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry. *Journal of Neurology, Neurosurgery & Psychiatry*, 73(6):657, 2002.

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

[17] 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.

[18] R. P. Jeffrey, R. E. Coleman, and P. M. Doraiswamy. Neuroimaging and early diagnosis of alzheimer disease : A look to the future. *Radiology*, 226:315–336, 2003.

[19] V. Julkunen et al. Cortical Thickness Analysis to Detect Progressive Mild Cognitive Impairment: A Reference to Alzheimer's Disease. *Dementia and geriatric cognitive disorders*, 28(5):404–412, 2009.

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

[21] M. Likeman et al. Visual assessment of atrophy on magnetic resonance imaging in the diagnosis of pathologically confirmed young-onset dementias. *Archives of neurology*, 62(9):1410, 2005.

[22] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[23] J. Liu, P. Wonka, and J. Ye. Multi-stage dantzig selector. In *Advances in Neural Information Processing Systems 23*, pages 1450–1458. 2010.

[24] L. McEvoy et al. Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Structural Changes in Mild Cognitive Impairment1. *Radiology*, 251(1):195, 2009.

[25] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. Stadlan. 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*, 34:939–44, 1984.

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

[27] E. Murphy et al. Six-month atrophy in MTL structures is associated with subsequent memory decline in elderly controls. *NeuroImage*, 2010.

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

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

[30] C. Paige and M. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.

[31] W. Rosen, R. Mohs, and K. Davis. A new rating scale for Alzheimer's disease. *American Journal of Psychiatry*, 141(11):1356, 1984.

[32] J. Sevigny, Y. Peng, L. Liu, and C. Lines. Item analysis of ADAS-Cog: effect of baseline cognitive impairment in a clinical AD trial. *Am J Alzheimers Dis Other Demen*, 25(2):119–24, 2010.

[33] 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.

[34] P. Thompson et al. Mapping hippocampal and ventricular change in Alzheimer disease. *Neuroimage*, 22(4):1754–1766, 2004.

[35] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[36] T. Tombaugh. Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Archives of clinical neuropsychology*, 20(4):485–503, 2005.

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

[38] 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.

[39] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[40] T. Zhang. Some sharp performance bounds for least squares regression with L1 regularization. *Annals of Statistics*, 37(5A):2109–2144, 2009.

[41] S. Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. *Advances in Neural Information Processing Systems*, 22, 2009.

[42] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.