Predicting Alzheimer's Disease Cognitive Assessment via Robust Low-Rank Structured Sparse Model

Jie Xu[1,3], **Cheng Deng**[1], **Xinbo Gao**[1], **Dinggang Shen**[2], **Heng Huang**[3,1]
[1]Xidian University, Xi'an 710071, China
[2]Department of Radiology and BRIC, UNC-Chapel Hill, USA
[3]University of Texas at Arlington, USA

Abstract

Alzheimer's disease (AD) is a neurodegenerative disorder with slow onset, which could result in the deterioration of the duration of persistent neurological dysfunction. How to identify the informative longitudinal phenotypic neuroimaging markers and predict cognitive measures are crucial to recognize AD at early stage. Many existing models related imaging measures to cognitive status using regression models, but they did not take full consideration of the interaction between cognitive scores. In this paper, we propose a robust low-rank structured sparse regression method (RLSR) to address this issue. The proposed model simultaneously selects effective features and learns the underlying structure between cognitive scores by utilizing novel mixed structured sparsity inducing norms and low-rank approximation. In addition, an efficient algorithm is derived to solve the proposed non-smooth objective function with proved convergence. Empirical studies on cognitive data of the ADNI cohort demonstrate the superior performance of the proposed method.

1 Introduction

Alzheimer's disease (AD), a common form of dementia, affects nerve cells in areas of the brain responsible for memory, cognition, language, and motor activity [Dailey, 2017]. By linear extrapolation of estimates from 2006, the population worldwide who have AD will increase to over 100 million by 2050 [Thompson *et al.*, 2003; Moradi *et al.*, 2015]. In fact, researchers believe that early detection will be key to preventing, slowing and stopping Alzheimer's disease. Neuroimaging as a powerful tool for accurate identification and understanding informative feature is necessary for early Alzheimer's disease prognosis and diagnosis [Liu *et al.*, 2015; Nie *et al.*, 2016]. Therefore, many machine learning methods have been proposed to study neuroimaging measures to detect pathology associated with AD and to predict cognitive scores [Wang *et al.*, 2011b; Huo *et al.*, 2016; Wang *et al.*, 2016]. Among them, structural magnetic resonance imaging (MRI) scans are one of the most extensively used imaging modality in tracking AD progression.

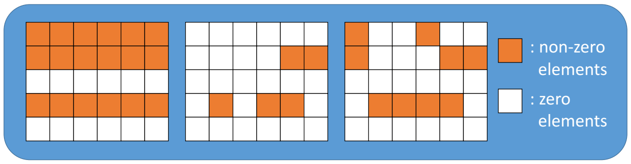


Figure 1: The sparse shrinkage patterns of matrix B imposed different structured sparsity-inducing norms: (a) $l_{2,1}$-norm, (b) $l_{2,1}$-norm + $l_{1,1}$-norm, (c) $l_{2,1}$-norm + $l_{1,2}$-norm. Blue points represent the non-zero weights and white points represent the zero weights. In (b), the $l_{1,1}$-norm suppresses the first feature selected by $l_{2,1}$-norm. In (c), $l_{1,2}$-norm will keep at least one non-zero weight for this feature, leading to the stable feature selection results.

In the association study of selecting effective longitudinal phenotypic markers to predict cognitive scores from imaging features, the input usually consists of two matrices: the imaging feature matrix $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ and the corresponding cognitive score matrix $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times m}$, where n is the number of samples, d is the number of features and m is the number of different measures of a certain cognitive performance.

A forthright method to identify informative imaging markers is to perform feature selection [Chang and Yang, 2016], which has been demonstrated as a useful way to reflect the correlation between cognitive measures after removing the indistinctive neuroimaging markers. More recently, sparse regularization model has been extensively utilized to learn the structure of data and obtain effective feature in different applications. The sparsity-inducing norm based feature selection methods solve the convex optimization problems of the form:

$$\min_{B} \mathcal{L}(B; X) + \lambda \Omega(B) \tag{1}$$

where $\mathcal{L}$ is a convex function and Ω can include one or more non-smooth sparsity-inducing norms.

When Ω is the l_1-norm, the l_1 shrinkage methods such as LASSO identify informative longitudinal phenotypic markers in the brain that are related to pathological changes of AD by imposing flat sparsity [Liu *et al.*, 2014]. However, the selected features distribute randomly among the whole brain that can not be well explained. In predicting cognitive scores task, we expect to select the most informative markers, which are important to all participants including AD, mild cognitive impairment (MCI) and heathy control (HC). To address this issue, group LASSO with a $l_{2,1}$-norm is used to impose the structured sparsity on parameter matrix for feature selection [Obozinski *et al.*, 2010; Jie *et al.*, 2015; Yang *et al.*, 2017; Chang *et al.*, 2017]. It enforces the important features to have non-zero weights cross all participants, however many important features often are only discriminant to partial classes, i.e. having large weights on these participants. Thus, such important features may be ignored by the above methods.

On this account, Lee *et al.* and Wang *et al.* proposed to add one more $l_{1,1}$-norm regularization term to achieve both structured and flat sparsity [Lee *et al.*, 2010; Wang *et al.*, 2011a]. However, because the $l_{1,1}$-norm regularization term enforces the flat sparsity and is prone to shrink the non-large values to zeros, the non-zero weights of some important features may also be forced to be zeros, i.e. some features selected by the $l_{2,1}$-norm regularization term can be totally suppressed by the $l_{1,1}$-norm regularization term. As a result, many informative longitudinal phenotypic markers are neglected during the feature selection procedure. Thus, more properly designed structured sparsity-inducing norms are desired in feature selection research.

In this paper, we propose a robust low-rank structured sparse regression method (RLSR) to simultaneously select the important neuroimaging markers and learn the underlying structure between cognitive measures. Our main contributions are three-fold: (1) The new mixed structured sparsity-inducing norms are introduced to overcome the above over-shrinkage drawback in the existing sparse learning based feature selection models. (2) The explicit rank-k low-rank matrix fitting approach is used to extract the underlying inter-relation structures between cognitive measures. (3) Because our method leads to a highly non-smooth objective, we derive an efficient algorithm to solve the new objective with proved convergence. We validate our method on cognitive data of the ADNI cohort and obtain promising results.

Notations. We summarize the notations used in this paper. Matrices are written as uppercase letters and vectors are written as bold lowercase letters. For matrix $W = \{w_{ij}\}$, its i-th row, j-th column are denoted as $\mathbf{w}^i$, $\mathbf{w}_j$ respectively. The l_p-norm of the vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $||\mathbf{v}||_p = (\sum_{i=1}^n |v_i|^p)^{\frac{1}{p}}$ for $p > 0$. The $l_{2,1}$-norm of matrix W is defined as $||W||_{2,1} = \sum_{i=1}^d ||\mathbf{w}^i||_2$ (in some related papers, people also used the notation l_1/l_2-norm). $l_{1,2}$-norm of matrix W is defined as $||W||_{1,2} = \sqrt{\sum_{i=1}^d ||\mathbf{w}^i||_1^2}$ and $l_{1,1}$-norm of matrix W is defined as $||W||_{1,1} = \sum_{i=1}^d \sum_{j=1}^m |w_{ij}|$.

2 Robust Low-Rank Structured Sparse Learning

The $l_{2,1}$-norm based objectives select the informative imaging markers across all the cognitive scores with joint sparsity, i.e. each imaging marker has either small score or large score for all the cognitive measures. However, for accurate identification of effective imaging markers, we utilize participants including AD, MCI and HC during deferent period. Consequently, many features may be irrelevant to each other, which could deteriorate the feature selection performance if we consider all the imaging markers as one group to do feature selection. On the other hand, as we discussed in the introduction section, an extra $l_{1,1}$-norm regularizer will suppress too many non-zero values and lead unstable feature selection results. For example, when we target to select top 20 features and adjust the trade-off parameter to let $l_{2,1}$-norm makes about 20 features with relatively large weights, the added $l_{1,1}$-norm will dramatically shrink the weights such that only few important features (much less than 20) can be selected. To tackle this over-shrinkage problem, we add a convex squared $l_{1,2}$-norm regularizer instead of $l_{1,1}$-norm and solve:

$$\min_B ||Y - X^T B||_F^2 + \lambda_1 ||B||_{2,1} + \lambda_2 ||B||_{1,2}^2. \quad (2)$$

The typical loss functions are the least square loss and logistic loss. To improve the computational efficiency, we utilize the least square loss in this paper to select informative markers for ADNI data. Thus, our method can be applied to both classification tasks (e.g. AD/MCI versus Normal Controls (NC)) and regression tasks (e.g., estimation of clinical cognitive scores). We perform the latter in this paper.

In Eq. (2), we proposed the novel mixed structured sparsity norms. The standard $l_{2,1}$-norm enforces the joint sparsity across all cognitive measures to select imaging markers. The new $l_{1,2}$-norm uses l_2-norm between markers, such that at least one non-zero element in the rows of B selected by $l_{2,1}$-norm regularizer will be kept. Thus, we won't lose the discriminative imaging markers selected by the $l_{2,1}$-norm regularization. At the same time, the $l_{1,2}$-norm imposes the l_1-norm between cognitive score weights of each marker to shrink the weight values of uncorrelated or irrelevant cognitive measures. For illustration purpose, in Fig. 1, we plot the sparse shrinkage patterns of the matrix B using: (a) $l_{2,1}$-norm regularizer only, (b) $l_{2,1}$- norm and $l_{1,1}$-norm regularizers, (c) $l_{2,1}$-norm and $l_{1,2}$-norm regularizers. The $l_{1,1}$-norm overshrinks the weights and removes the first feature selected by $l_{2,1}$-norm. The $l_{1,2}$-norm suppress some weights with supporting the results of $l_{2,1}$-norm, e.g. the first feature is still kept in the list.

More important, with regard to this specific task that predicting AD progression, we hope to extract and utilize the underlying interrelations between cognitive measures to enhance the accuracy of feature selection. In recent research [Ji and Ye, 2009; Deng *et al.*, 2015], the trace norm regularization has been used to seek the low-rank structured shared common representations:

$$\min_B ||Y - X^T B||_F^2 + \lambda ||B||_*. \quad (3)$$

However, there are two deficiencies: 1) the optimal low-rank approximation is resulted by tediously tuning the parameter λ, which has no direct connection to the rank value; 2) the feature selection isn't enforced in this model.

To address the above problems, we consider the following low-rank regression:

$$\min_B ||Y - X^T B||_F^2 \quad s.t. \quad rank(B) = s \leq \min(m, d), \quad (4)$$

where m is the number of classes and d is the dimension of features. Compared to the parameter λ in Eq. (3), the parameter s is more feasible to be decided by users. Moreover, in order to select features and simultaneously keep the low-rank matrix fitting, we propose the following model,

$$\begin{aligned}&\min_{W,P} ||Y - X^T WP||_F^2 + \lambda_1 ||W||_{2,1} + \lambda_2 ||W||_{1,2}^2 \\ &s.t.\ PP^T = I\end{aligned} \tag{5}$$

where $W \in \mathbb{R}^{d \times s}$, $P \in \mathbb{R}^{s \times m}$ and $s \leq \min(m, d)$. The product $B = WP$ is a low-rank matrix with $rank(B) \leq s$. Our new objective simultaneously learns the underlying interrelation between cognitive measures by low-rank matrix fitting and selects the informative neuroimaging markers by mixed structured sparsity-inducing norms. We also consider the real world data often have outliers and hence replace the least square loss by the $l_{2,1}$-norm based loss function, which imposes the l_1-norm between data points to reduce the effect of outliers. Our final objective is to solve:

$$\begin{aligned}&\min_{W,P} ||Y - X^T WP||_{2,1} + \lambda_1 ||W||_{2,1} + \lambda_2 ||W||_{1,2}^2 \\ &s.t.\ PP^T = I\end{aligned} \tag{6}$$

The resulted objective has three non-smooth terms, such that the optimization becomes difficult. To solve our new objective, we will derive an efficient algorithm in next section with proved convergence.

3 Optimization

In this section, an efficient algorithm is proposed to tackle Eq. (6), followed by the proof of its convergence.

3.1 Algorithm Derivation

We will alternatively and iteratively solve Eq. (6). To begin with, we rewrite Eq. (6) as:

$$\begin{aligned}\min_{W,P,H,\hat{D},D_k} \quad & Tr((Y - X^T WP)^T H(Y - X^T WP)) \\ & + \lambda_1 Tr(W^T \hat{D} W) + \lambda_2 \sum_{k=1}^{s} \mathbf{w}_k^T D_k \mathbf{w}_k, \quad s.t. \quad PP^T = I\end{aligned} \tag{7}$$

where $W = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_s] \in \mathbb{R}^{d \times s}$. Denote

$$E = Y - X^T WP, \tag{8}$$

then $H \in \mathbb{R}^{n \times n}$ is a diagonal matrix and defined as:

$$H(i, i) = \frac{1}{2||\mathbf{e}^i||_2}, \tag{9}$$

where $\mathbf{e}^i (\forall i = 1, 2, ..., n)$ is the i-th row of matrix E in Eq. (8). $\hat{D} \in \mathbb{R}^{d \times d}$ is a diagonal matrix and defined as

$$\hat{D}(j, j) = \frac{1}{2||\mathbf{w}^j||_2}, \forall j = 1, 2, ..., d \tag{10}$$

and $D_k \in \mathbb{R}^{d \times d}$ is also a diagonal matrix and defined as:

$$D_k(j, j) = \frac{||\mathbf{w}^j||_1}{|w_{jk}|}, \forall k = 1, 2, ..., s, \forall j = 1, 2, ..., d. \tag{11}$$

The first step is to fix $P, H, \hat{D}, D_k$, and to solve W. Thus, we need solve the following subproblem:

$$\begin{aligned}\min_{\mathbf{w}_k} \quad & \sum_{k=1}^{s} -2\mathbf{z}_k^T \mathbf{w}_k + \mathbf{w}_k^T (XHX^T)\mathbf{w}_k + \lambda_1 \mathbf{w}_k^T \hat{D} \mathbf{w}_k \\ & + \lambda_2 \mathbf{w}_k^T D_k \mathbf{w}_k,\end{aligned} \tag{12}$$

where $\mathbf{z}_k$ is the k-th column of the matrix $XHYP^T, \forall k = 1, 2, ..., s$. Then taking derivative of Eq. (12) w.r.t. $\mathbf{w}_k$ and setting it to zero, we get,

$$-2\mathbf{z}_k + 2(XHX^T)\mathbf{w}_k + 2\lambda_1 \hat{D}\mathbf{w}_k + 2\lambda_2 D_k \mathbf{w}_k = 0 \tag{13}$$

$$\mathbf{w}_k = (XHX^T + \lambda_1 \hat{D} + \lambda_2 D_k)^{-1} \mathbf{z}_k \tag{14}$$

The second step is to fix $W, H, \hat{D}, D_k$, and to solve P. Because

$$\begin{aligned}& ||Y - X^T WP||_{2,1} \\ = & Tr((Y - X^T WP)^T H(Y - X^T WP)) \\ = & Tr(Y^T HY) - 2Tr(Y^T HX^T WP) + Tr(W^T XHX^T W).\end{aligned} \tag{15}$$

Then the subproblem becomes:

$$\max_P Tr(PY^T HX^T W) \quad s.t.\ PP^T = I \tag{16}$$

The solution to Eq. (16) can be obtained by the Theorem 1. **The third step** is to fix W, P, and to solve H by Eq. (8) and Eq. (9), solve $\hat{D}$ by Eq. (10), and solve D_k by Eq. (11).

We repeat the above three steps iteratively, until the predefined stopping criterion is satisfied. We summarize the whole algorithm in Alg. 1.

The Step 2 in the iteration of Alg. 1 can be calculated by linear equation system, which can be efficiently solved. Thus, our algorithm can be applied in large-scale datasets.

Algorithm 1 The algorithm to solve Eq. (6)

```
Input:
1. The training data $X \in \mathbb{R}^{d \times n}$ with label matrix $Y \in \mathbb{R}^{n \times m}$
2. The regularization parameters $\lambda_1$ and $\lambda_2$, and the rank $s$.
Output:
1. The matrices $W \in \mathbb{R}^{d \times s}$ and $P \in \mathbb{R}^{s \times m}$.
Initialization:
1. Set $t = 0$, initialize $H^{(t)} = I_{n \times n}$, $\hat{D}^{(t)} = I_{d \times d}$, $D_k^{(t)} = I_{d \times d}, \forall k = 1, ..., s$, randomly initialize $P^{(t)} \in \mathbb{R}^{s \times m}$ with $P^{(t)} P^{(t)^T} = I \in \mathbb{R}^{s \times s}$.
Repeat:
    1. Calculate $\mathbf{z}_k^{(t)}$, which is the $k$-th column of the matrix $XH^{(t)} YP^{(t)^T}$.
    2. Calculate $W^{(t)}$ column by column by $\mathbf{w}_k^{(t)} = (XH^{(t)} X^T + \lambda_1 \hat{D}^{(t)} + \lambda_2 D_k^{(t)})^{-1} \mathbf{z}_k^{(t)}$.
    3. Calculate $M^{(t)} = Y^T H^{(t)} X^T W^{(t)}$.
    4. Do SVD of $M^{(t)}$, $M^{(t)} = U^{(t)} \Sigma^{(t)} V^{(t)^T}$.
    5. Update $P^{(t+1)} = V^{(t)} [I, 0] U^{(t)^T}$.
    6. Update $E^{(t+1)} = Y - X^T W^{(t)} P^{(t+1)}$.
    7. Update $H^{(t+1)}(i, i) = \frac{1}{2||\mathbf{e}^{(t+1)i}||_2}, \forall i = 1, 2, ..., n$.
    8. Update $\hat{D}^{(t+1)}(j, j) = \frac{1}{2||\mathbf{w}^{(t)j}||}, \forall j = 1, 2, ..., d$.
    9. Update $D_k(j, j) = \frac{||\mathbf{w}^{(t)j}||_1}{|w_{jk}^{(t)}|}, \forall j = 1, 2, ..., d.$.
    10. Update $t = t + 1$.
Until Converge
```

3.2 Convergence Analysis

Theorem 1. *The solution to the optimization problem* $\max_P Tr(PM)$ *s.t.* $PP^T = I$*, where* $p \in \mathbb{R}^{s \times m}$*,* $M \in$

$\mathbb{R}^{m\times s}$ and $s < m$ is

$$P^* = V[I,0]U^T \tag{17}$$

where U and V are the SVD of M, $M = U\Lambda V^T$ and I is the identify matrix, $I \in \mathbb{R}^{s\times s}$. $0 \in \mathbb{R}^{s\times(m-s)}$ is the matrix with all zeros entries. And $[\Phi, \Psi]$ is the matrix operation to horizontally concatenate two matrices Φ and Ψ who have the same number of rows.

Proof. We do SVD of M, $M = U\Lambda V^T$, where $U \in \mathbb{R}^{m\times m}$, $\Lambda \in \mathbb{R}^{m\times s}$, $V \in \mathbb{R}^{s\times s}$. Then, we have

$$\begin{aligned} Tr(PM) &= Tr(PU\Lambda V^T) = Tr(\Lambda V^T PU) \\ &= Tr(\Lambda Q) = \sum_{k=1}^{s} \lambda_{kk} q_{kk} \end{aligned} \tag{18}$$

where $Q = V^T PU$, $Q \in \mathbb{R}^{s\times m}$. Note that $s < m$. λ_{kk} and q_{kk} are the k-th element in the diagonal of matrix Λ and Q respectively. Moreover,

$$QQ^T = V^T PUU^T P^T V = I, \tag{19}$$

where $I \in \mathbb{R}^{s\times s}$ is the identity matrix. Thus, $q_{kk} \leq 1$, $\forall k, k = 1, 2, ..., s$. Therefore,

$$Tr(PM) = \sum_{k=1}^{s} \lambda_{kk} q_{kk} \leq \sum_{k}^{s} \lambda_{kk}, \tag{20}$$

and when $q_{kk} = 1, \forall k, k = 1, 2, ..., s$, the equality holds. In other words, $Tr(PM)$ reaches the maximum when $Q = [I, 0]$. Recall that $Q = V^T PU$, thus the optimal solution to Eq. (16) is Eq. (17). □

The convergence of the Alg. 1 is summarized in the following theorem:

Theorem 2. *The Alg. 1 will monotonically decrease the objective of the problem in Eq. (6) in each iteration and converge to the local optimum solution to the problem.*

Proof. On one hand, denote the updated P by $\tilde{P}$. Because of Theorem 1, when we fix W, H, $\hat{D}$ and D_k, we get:

$$\begin{aligned} &Tr((Y - X^T W\tilde{P})^T H(Y - X^T W\tilde{P})) \\ &\quad + \lambda_1 Tr(W^T \hat{D} W) + \lambda_2 \sum_{k=1}^{s} \mathbf{w}_k^T D_k \mathbf{w}_k \\ \leq &Tr((Y - X^T WP)^T H(Y - X^T WP)) \\ &\quad + \lambda_1 Tr(W^T \hat{D} W) + \lambda_2 \sum_{k=1}^{s} \mathbf{w}_k^T D_k \mathbf{w}_k \end{aligned} \tag{21}$$

After plugging the definition of H by Eq. (9), we can obtain,

$$\begin{aligned} &\sum_{i=1}^{n} \frac{||\tilde{\mathbf{e}}^i||_2^2}{2||\mathbf{e}^i||_2} + \lambda_1 \sum_{k=1}^{s} \mathbf{w}_k^T \hat{D} \mathbf{w}_k + \lambda_2 \sum_{k=1}^{s} \mathbf{w}_k^T D_k \mathbf{w}_k \\ \leq &\sum_{i=1}^{n} \frac{||\mathbf{e}^i||_2^2}{2||\mathbf{e}^i||_2} + \lambda_1 \sum_{k=1}^{s} \mathbf{w}_k^T \hat{D} \mathbf{w}_k + \lambda_2 \sum_{k=1}^{s} \mathbf{w}_k^T D_k \mathbf{w}_k \end{aligned} \tag{22}$$

At the same time, beginning with $(||\tilde{\mathbf{e}}^i||_2 - ||\mathbf{e}^i||_2)^2 \geq 0$, we can get

$$\sum_{i=1}^{n}(||\tilde{\mathbf{e}}^i||_2 - \frac{||\tilde{\mathbf{e}}^i||_2^2}{2||\mathbf{e}^i||_2}) \leq \sum_{i=1}^{n}(||\mathbf{e}^i||_2 - \frac{||\mathbf{e}^i||_2^2}{2||\mathbf{e}^i||_2}) \tag{23}$$

Therefore, adding Eq. (22) and Eq. (23) together, we get

$$\begin{aligned} &||Y - X^T W\tilde{P}||_{2,1} + \lambda_1 \sum_{k=1}^{s} \mathbf{w}_k^T \hat{D} \mathbf{w}_k + \lambda_2 \sum_{k=1}^{s} \mathbf{w}_k^T D_k \mathbf{w}_k \\ \leq &||Y - X^T WP||_{2,1} + \lambda_1 \sum_{k=1}^{s} \mathbf{w}_k^T \hat{D} \mathbf{w}_k + \lambda_2 \sum_{k=1}^{s} \mathbf{w}_k^T D_k \mathbf{w}_k \end{aligned} \tag{24}$$

On the other hand, denote the updated W by $\tilde{W}$, when we fix P, H. According to the step 2 in the **Repeat** part in Alg. 1, we have:

$$\begin{aligned} &Tr(Y^T HY) + \sum_{k=1}^{s}(-2\mathbf{z}_k^T \tilde{\mathbf{w}}_k + \tilde{\mathbf{w}}_k^T (XHX^T)\tilde{\mathbf{w}}_k \\ &\quad + \lambda_1 \tilde{\mathbf{w}}_k^T \hat{D} \tilde{\mathbf{w}}_k + \lambda_2 \tilde{\mathbf{w}}_k^T D_k \tilde{\mathbf{w}}_k) \\ \leq &Tr(Y^T HY) + \sum_{k=1}^{s}(-2\mathbf{z}_k^T \mathbf{w}_k + \mathbf{w}_k^T (XHX^T)\mathbf{w}_k \\ &\quad + \lambda_1 \mathbf{w}_k^T \hat{D} \mathbf{w}_k + \lambda_2 \mathbf{w}_k^T D_k \mathbf{w}_k) \end{aligned} \tag{25}$$

where $\mathbf{z}_k$ is the k-th column of $XHYP^T$.

We plug in the definition of $\hat{D}$ and D_k by Eq. (10) and Eq. (11) respectively into Eq. (25), and have:

$$\begin{aligned} &Tr(Y^T HY) + \sum_{k=1}^{s}(-2\mathbf{z}_k^T \tilde{\mathbf{w}}_k + \tilde{\mathbf{w}}_k^T (XHX^T)\tilde{\mathbf{w}}_k) \\ &\quad + \lambda_1 \sum_{k=1}^{s}\sum_{j=1}^{d} \frac{\tilde{w}_{jk}^2}{2||\mathbf{w}^j||_2} + \lambda_2 \sum_{k=1}^{s}\sum_{j=1}^{d} \frac{||\mathbf{w}^j||_1}{|w^{jk}|}\tilde{w}_{jk}^2 \\ \leq &Tr(Y^T HY) + \sum_{k=1}^{s}(-2\mathbf{z}_k^T \mathbf{w}_k + \mathbf{w}_k^T (XHX^T)\mathbf{w}_k \\ &\quad + \lambda_1 \sum_{k=1}^{s}\sum_{j=1}^{d} \frac{w_{jk}^2}{2||\mathbf{w}^j||_2} + \lambda_2 \sum_{k=1}^{s}\sum_{j=1}^{d} \frac{||\mathbf{w}^j||_1}{|w^{jk}|} w_{jk}^2 \end{aligned} \tag{26}$$

Similarly, beginning with $(||\tilde{\mathbf{w}}^i||_2 - ||\tilde{\mathbf{w}}^i||_2)^2 \geq 0$, we get

$$||\tilde{\mathbf{w}}^j||_2 - \frac{||\tilde{\mathbf{w}}^j||_2^2}{2||\mathbf{w}^j||_2} \leq ||\mathbf{w}^j||_2 - \frac{||\mathbf{w}^j||_2^2}{2||\mathbf{w}^j||_2} \tag{27}$$

Therefore,

$$\sum_{j=1}^{d}||\tilde{\mathbf{w}}^j||_2 - \sum_{j=1}^{d}\sum_{k=1}^{s} \frac{\tilde{w}_{jk}^2}{2||\mathbf{w}^j||_2} \leq \sum_{j=1}^{d}||\mathbf{w}^j||_2 - \sum_{j=1}^{d}\sum_{k=1}^{s} \frac{w_{jk}^2}{2||\mathbf{w}^j||_2}. \tag{28}$$

Meanwhile, according to am-gm inequality, we have

$$\sum_{k=1}^{s} \frac{||\mathbf{w}^j||_1}{|w^{jk}|}\tilde{w}_{jk}^2 \geq \left(||\tilde{\mathbf{w}}^j||_1\right)^2 \tag{29}$$

Thus,

$$
\begin{aligned}
&\Big(||\tilde{\mathbf{w}}^j||_1\Big)^2 - \sum_{k=1}^{s} \frac{||\mathbf{w}^j||_1}{|w^{jk}|}\tilde{w}_{jk}^2 \le 0 = \Big(||\tilde{\mathbf{w}}^j||_1\Big)^2 - \sum_{k=1}^{s} \frac{||\mathbf{w}^j||_1}{|w^{jk}|} w_{jk}^2 \\
&\Rightarrow \sum_{j=1}^{d}\Big(||\tilde{\mathbf{w}}^j||_1\Big)^2 - \sum_{k=1}^{s}\sum_{j=1}^{d} \frac{||\mathbf{w}^j||_1}{|w^{jk}|}\tilde{w}_{jk}^2 \\
&\le 0 = \sum_{j=1}^{d}\Big(||\tilde{\mathbf{w}}^j||_1\Big)^2 - \sum_{k=1}^{s}\sum_{j=1}^{d} \frac{||\mathbf{w}^j||_1}{|w^{jk}|} w_{jk}^2
\end{aligned}
\tag{30}
$$

Sum up Eq. (26), $\lambda_1 \times$Eq. (28), $\lambda_2 \times$Eq. (30), we can arrive at the following conclusion:

$$
\begin{aligned}
&Tr((Y - X^T\tilde{W}P)^T H(Y - X^T\tilde{W}P)) \\
&+ \lambda_1||\tilde{W}||_{2,1} + \lambda_2||\tilde{W}||_{1,2}^2 \\
\le& Tr((Y - X^T W P)^T H(Y - X^T W P)) \\
&+ \lambda_1||W||_{2,1} + \lambda_2||W||_{1,2}^2
\end{aligned}
\tag{31}
$$

We update W and P alternatively, then we arrive at our goal:

$$
\begin{aligned}
&||Y - X^T\tilde{W}\tilde{P}||_{2,1} + \lambda_1||\tilde{W}||_{2,1} + \lambda_2||\tilde{W}||_{1,2}^2 \\
\le& ||Y - X^T W P||_{2,1} + \lambda_1||W||_{2,1} + \lambda_2||W||_{1,2}^2 \\
&s.t.\ \ \tilde{P}\tilde{P}^T = I, PP^T = I
\end{aligned}
\tag{32}
$$

□

In other words, using Alg. 1, we can monotonically decrease the objective function Eq. (6) in each iteration and finally it will converge.

4 Experimental Results

In this section, we evaluate prediction performance of the proposed method by applying it to Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (adni.loni.usc.edu), where a wide range of imaging markers measured over a period of 2 years are examined and associated to cognitive scores that are relevant to AD.

4.1 Data Descriptions

We apply the proposed method to the ADNI cohort to predict the cognitive scores of the participants from each of their two types of imaging phenotypes, i.e. FreeSurfer markers and voxel-based morphometry (VBM) markers. The detailed information are shown in Table 1. Mean modulated gray matter measures obtained from 90 target regions of interest, normalized by the total intracranial volume, were extracted as features.

Table 1: Numbers of participants in the experiments using two different types of imaging markers

Imaging phenotypes	#Total	#AD	#MCI	#HC
FreeSurfer	496	99	225	172
VBM	440	85	203	152

4.2 Performance Comparison on the ADNI Cohort

First, we intend to identify a certain set of informative markers that are closely relate to pathological change due to AD. We compared our method against three most related algorithms including multivariate ridge regression (RR), joint $l_{2,1}$-norm minimization ($l_{2,1}$) on both loss function and regularization [Nie *et al.*, 2010], linear regression with trace norm. These comparing methods are all widely used in statistical learning and brain image analysis.

In all experiments, we automatically tune the regularization parameters by selecting among the values $\{10^r : r \in \{-5, ..., 5\}\}$ with standard 5-fold cross-validation strategy.After the algorithm converges, we sort the row index of matrix W by the summation of the absolute values in each row, and features are selected by the top ranked indices. To measure prediction performance, we compute the root mean square error (RMSE) between the predicted score and the ground truth.

The prediction experiment evaluated by ridge regression is repeated for 100 times and average results are reported in Fig. 2. As shown in Fig. 2, we can clearly see that the prediction results of our method consistently outperforms other competing methods in nearly all the test cases for all the cognitive tasks except some outlier part in Fig. 2f. But it doesn't really matter since the proposed method catches up soon. The reasons why the proposed method performs best go as follows: RR assumed the cognitive measures to be independent at different time point which neglects the correlations along the time. And for $l_{2,1}$, since the pathological change of brain structures due to AD usually do not occur in the pre-identified regions with certain shapes, thus it is difficult to define meaningful feature groups. This makes $l_{2,1}$ perform worse. Trace norm is a good way to seek the underlying interrelations between cognitive scores, but it ignores the fact that the informative markers relate to AD among all the imaging measures only occupy a small part. As for the proposed method, we not only detect group structure within longitudinal phenotypic neuroimaging markers, but also capture the correlations among cognitive measures. In addition, for ease of comparison, we also list the RMSE using top 10 and top 20 selected features evaluated by ridge regression in Table 2.

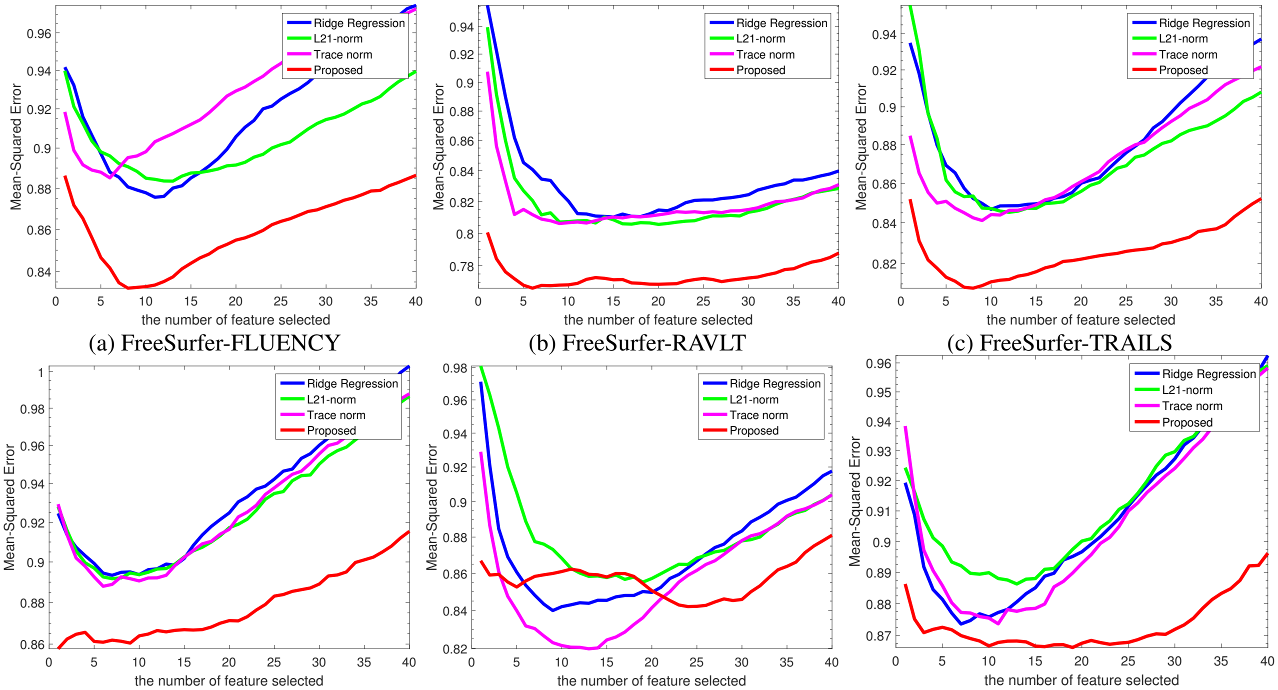
4.3 Identification of Informative Markers

The primary goal of the proposed method is to identify the informative markers which is important for AD diagnosis and prediction. Therefore, we examine the imaging markers selected by our method and show it in Fig. 3. Due to the limit size of display, we only provide one tenth of feature names for both FreeSurfer and VBM markers. As shown in Fig. 3, we observe that hippocampal measures (LHippocampus, RHippocampus, LHippVol and RHippVol) are among the top selected features. These findings are in accordance with the known knowledge that in the pathological pathway of AD, hippocampal is one of sections that can recognize Alzheimer-related changes [Braak and Braak, 1991; Delacourte *et al.*, 1999].

In summary, the identified neuroimaging markers are highly suggestive and effective for tracking the progression of AD, since it strongly agrees with the existing research findings. It also illustrates the necessity and correctness of the selected imaging cognitive associations to reveal the relationships between MRI measures and cognitive scores.

Table 2: Prediction performance measured by RMSE

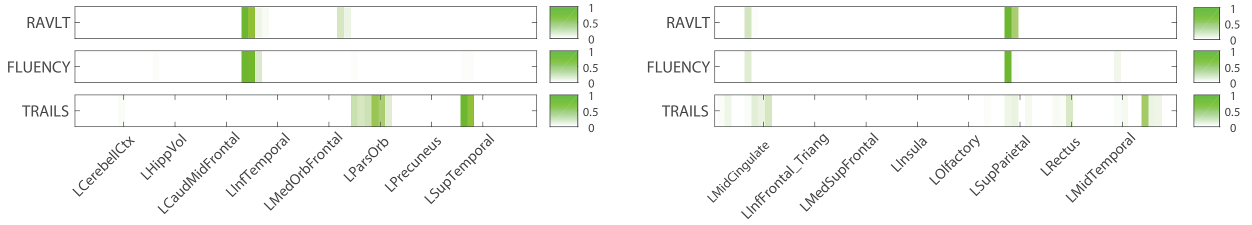
		RMSE of top 10 features				RMSE of top 30 features			
		RR	$l_{2,1}$	Trace	Proposed	RR	$l_{2,1}$	Trace	Proposed
FreeSurfer	FLUENCY	0.8777	0.8849	0.8982	**0.8328**	0.9411	0.9145	0.9560	**0.8710**
	RAVLT	0.8202	0.8073	0.8066	**0.7685**	0.8245	0.8132	0.8150	**0.7726**
	TRAILS	0.8467	0.8471	0.8441	**0.8110**	0.8970	0.8820	0.8923	**0.8302**
VBM	FLUENCY	0.8937	0.8937	0.8906	**0.8639**	0.9601	0.9501	0.9555	**0.8891**
	RAVLT	0.8420	0.8682	**0.8215**	0.8610	0.8834	0.8779	0.8781	**0.8459**
	TRAILS	0.8758	0.8899	0.8754	**0.8667**	0.9273	0.9297	0.9241	**0.8719**



(a) FreeSurfer-FLUENCY (b) FreeSurfer-RAVLT (c) FreeSurfer-TRAILS

(d) VBM-FLUENCY (e) VBM-RAVLT (f) VBM-TRAILS

Figure 2: RMSE of four feature selection algorithms on different cognitive assessment scores.



(a) FreeSurfer (b) VBM

Figure 3: Heat maps of our learned weight matrices on different cognitive assessment scores.

5 Conclusion

To reveal the relationship between cognitive measures and neuroimaging markers, we proposed a novel robust low-rank structured sparse regression model, which selects the most informative imaging markers to predict the cognitive scores for complex brain disorders. Using the new mixed structured sparsity inducing norms and the low-rank approximation function, the proposed method can efficiently identify the effective neuroimaging markers with utilizing the underlying interrelation structures between different cognitive measures. In addition, we provide an efficient algorithm with proved convergence. Validation experiments conducted on multiple data demonstrate the promise of the proposed method.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China 61572388, and U.S. NIH R01 AG049371, NSF IIS 1302675, IIS 1344152, DBI 1356628, IIS 1619308, IIS 1633753.

References

[Braak and Braak, 1991] Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.

[Chang and Yang, 2016] Xiaojun Chang and Yi Yang. Semi-supervised feature analysis by mining correlations among multiple tasks. *IEEE transactions on neural networks and learning systems*, 2016.

[Chang *et al.*, 2017] Xiaojun Chang, Zhigang Ma, Yi Yang, Zhiqiang Zeng, and Alexander G Hauptmann. Bi-level semantic representation analysis for multimedia event detection. *IEEE transactions on cybernetics*, 47(5):1180–1197, 2017.

[Dailey, 2017] Christina Dailey. The impact of alzheimer's disease-the silent killer. *JCCC Honors Journal*, 7(2):1, 2017.

[Delacourte *et al.*, 1999] A Delacourte, JP David, N Sergeant, L Buee, A Wattez, P Vermersch, F Ghozali, C Fallet-Bianco, F Pasquier, F Lebert, et al. The biochemical pathway of neurofibrillary degeneration in aging and alzheimer's disease. *Neurology*, 52(6):1158–1158, 1999.

[Deng *et al.*, 2015] Cheng Deng, Zongting Lv, Wei Liu, Junzhou Huang, Dacheng Tao, and Xinbo Gao. Multi-view matrix decomposition: A new scheme for exploring discriminative information. In *IJCAI*, pages 3438–3444, 2015.

[Huo *et al.*, 2016] Zhouyuan Huo, Dinggang Shen, and Heng Huang. New multi-task learning model to predict alzheimer's disease cognitive assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 317–325. Springer, 2016.

[Ji and Ye, 2009] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning*, pages 457–464. ACM, 2009.

[Jie *et al.*, 2015] Biao Jie, Daoqiang Zhang, Bo Cheng, and Dinggang Shen. Manifold regularized multitask feature learning for multimodality disease classification. *Human brain mapping*, 36(2):489–507, 2015.

[Lee *et al.*, 2010] Seunghak Lee, Jun Zhu, and Eric P Xing. Adaptive multi-task lasso: with application to eQTL detection. In *Advances in neural information processing systems*, pages 1306–1314, 2010.

[Liu *et al.*, 2014] Manhua Liu, Daoqiang Zhang, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Identifying informative imaging biomarkers via tree structured sparse learning for ad diagnosis. *Neuroinformatics*, 12(3):381–394, 2014.

[Liu *et al.*, 2015] Mingxia Liu, Daoqiang Zhang, and Dinggang Shen. View-centralized multi-atlas classification for alzheimer's disease diagnosis. *Human brain mapping*, 36(5):1847–1865, 2015.

[Moradi *et al.*, 2015] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Alzheimer's Disease Neuroimaging Initiative, et al. Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. *Neuroimage*, 104:398–412, 2015.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[Nie *et al.*, 2016] Liqiang Nie, Luming Zhang, Lei Meng, Xuemeng Song, Xiaojun Chang, and Xuelong Li. Modeling disease progression via multisource multitask learners: A case study with alzheimer's disease. *IEEE transactions on neural networks and learning systems*, 2016.

[Obozinski *et al.*, 2010] Guillaume Obozinski, Ben Taskar, and Michael I Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

[Thompson *et al.*, 2003] Paul M Thompson, Kiralee M Hayashi, Greig De Zubicaray, Andrew L Janke, Stephen E Rose, James Semple, David Herman, Michael S Hong, Stephanie S Dittmer, David M Doddrell, et al. Dynamics of gray matter loss in alzheimer's disease. *Journal of neuroscience*, 23(3):994–1005, 2003.

[Wang *et al.*, 2011a] Hua Wang, Feiping Nie, Heng Huang, Shannon Risacher, Chris Ding, Andrew J Saykin, Li Shen, et al. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *International Conference on Computer Vision (ICCV)*, pages 557–562. IEEE, 2011.

[Wang *et al.*, 2011b] Hua Wang, Feiping Nie, Heng Huang, Shannon Risacher, Andrew J Saykin, Li Shen, et al. Identifying ad-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 115–123. Springer, 2011.

[Wang *et al.*, 2016] Xiaoqian Wang, Dinggang Shen, and Heng Huang. Prediction of memory impairment with mri data: A longitudinal study of alzheimer's disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–281. Springer, 2016.

[Yang *et al.*, 2017] Yanhua Yang, Cheng Deng, Shangqian Gao, Wei Liu, Dapeng Tao, and Xinbo Gao. Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Transactions on Multimedia*, 19(3):519–529, 2017.