FROM PHENOTYPE TO GENOTYPE: A STRUCTURED SPARSE LEARNING FRAMEWORK FOR IMAGING GENETICS STUDIES

by

HUA WANG

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
June 2012

To my wife, Yanzi, whose constant loyalty and support made it possible.

To my two newborn daughters, Seraphina and Angelina, whose comings in this summer have provided me with the power and creativity.

ACKNOWLEDGEMENTS

First and foremost I want to thank my supervising professor Dr. Heng Huang. It has been an honor to be his first Ph.D. student. He has taught me, both consciously and unconsciously, how good computer science research is done. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example he has provided as a successful computer scientist and professor.

I would also like to extend my appreciation to my academic advisors Dr. Chris Ding, Dr. Jeff Lei, Dr. Nan Zhang for their interest in my research and for taking time to serve in my dissertation committee.

The members of the Computational Science Lab (CSL) at the Department of Computer Science and Engineering (CSE) have contributed immensely to my personal and professional time at the University of Texas at Arlington. The group has been a source of friendships as well as good advice and collaboration. We worked together on many fascinating research topics in machine learning and data mining, as well as their applications in bioinformatics, computer vision and medical image computing. I very much appreciated the lab members, whom I have had the pleasure to work with or alongside of, for their enthusiasm, intensity, and amazing ability.

I am very grateful to all the teachers who taught me during the years I spent in school, for encouraging and inspiring me to pursue graduate studies.

Finally, I would like to express my deep gratitude to my wife, who has encouraged and inspired me. I am extremely fortunate to be so blessed. I am also extremely grateful to my parents for their sacrifice, encouragement and patience. I also thank several of my friends who have helped me throughout my career.

June 11, 2012

ABSTRACT

FROM PHENOTYPE TO GENOTYPE: A STRUCTURED SPARSE LEARNING FRAMEWORK FOR IMAGING GENETICS STUDIES

HUA WANG, Ph.D.

The University of Texas at Arlington, 2012

Supervising Professor: Heng Huang

Sparsity is one of the intrinsic properties of real-world data, thus sparse representation based learning models have been widely used to simplify data modeling and discover predictive patterns. By enforcing properly designed structured sparsity, one can unify specific data structures with the learning model. We proposed several novel structured sparsity learning models for multi-modal data fusion, heterogeneous tasks integration, and group structured feature selection.

We applied our new structured sparse learning methods to the emerging imaging genetics studies by integrating phenotypes and genotypes to discover new biomarkers which are able to characterize neurodegenerative process in the progression of Alzheimer's disease and other brain disorders. Different to traditional association studies, our new structured sparse learning models can elegantly take advantage of the useful information contained in biomarkers, cognitive measures, and disease status, where, crucially, the interrelated structures within and between both genetic/imaging data and clinical outcomes are gracefully exploited by our newly designed convex sparse regularization models.

We empirically evaluate our new methods on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort to identify Alzheimer's disease (AD) risky biomarkers, where we have achieved not only clearly improved prediction performance for cognitive measurements and diagnosis status, but also a compact set of highly suggestive biomarkers relevant to AD.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Backgrounds and Introduction of Imaging Genetics

Alzheimer's disease (AD) is the most common age related neurodegenerative disease affecting nearly 25 million people worldwide, a number expected to triple in the next 50 years. Patients with AD show significant impairment in multiple cognitive domains, including deficits in memory and executive functioning. Progress in the early clinical diagnosis of AD has led to the characterization of a prodromal syndrome featuring relatively isolated memory deficits termed "amnestic mild cognitive impairment" (mild cognitive impairment; MCI). Amnestic MCI is conceptualized as a preliminary stage of AD-associated neurodegeneration with the majority of patients eventually progressing to AD at a rate of 10%—15% per year.

The increasing recognition that early diagnosis and therapeutic intervention will be necessary to prevent the development of AD underscores the need to develop sensitive and specific biomarkers for detecting and monitoring MCI and AD. Structural magnetic resonance imaging (MRI) has shown significant promise as a biomarker to detect early MCI and AD-associated changes, as well as to predict the rate of disease progression. Cross-sectional studies evaluating the utility of structural MRI in detecting neurodegeneration have identified significant brain atrophy in patients with MCI and AD, particularly in regions of the medial temporal lobe (MTL) using regional volumetric extraction tools such as manual tracing of regions of interest, and more recently, automated segmentation and parcellation of target regions. Other semiautomated tools which provide 3-dimensional mapping of brain morphology, including voxel-based morphometry (VBM), tensor-based morphometry (TBM) and related techniques have also identified significant global and local tissue changes in patients with MCI and AD, including decreased whole brain, hippocampal, and temporal lobar gray matter (GM) density. Structural MRI techniques have also been shown to provide sensitive prediction of disease progression. Hippocampal volume and GM density, as well as measures of MTL volume and cortical thickness, have been identified as sensitive biomarkers for predicting conversion from MCI to probable AD.

Imaging genetics is an emergent transdisciplinary research field where the association between genetic variation and imaging measures as Quantitative Traits (QTs) or continuous phenotypes is evaluated. Imaging genetic studies have certain advantages over traditional case control studies. First, QT association studies have been shown to have increased statistical power and thus decreased sample size requirements. Second, imaging phenotypes may be closer to the underlying biological etiology of the disease, making it easier to identify underlying genes. Therefore, my graduate study focuses on imaging studies to associate phenotypes to genotypes using structured sparse learning method.

Under the framework of imaging genetics, my research focus on association studies for AD, where we emphasize the identification of AD relevant biomarkers.

1.2 Structured Sparsity and Its Applications in Machine Learning

The concept of parsimony is central in many scientific domains. In the context of statistics, signal processing or machine learning, it takes the form of variable or feature selection problems, and is commonly used in two situations: First, to make the model or the prediction more interpretable or cheaper to use, *i.e.*, even if the underlying problem does not admit sparse solutions, one looks for the best sparse

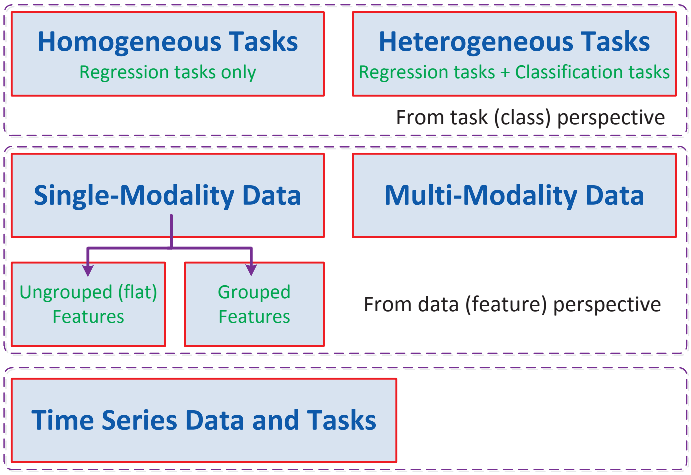

Homogeneous Tasks
Regression tasks only
Heterogeneous Tasks
Regression tasks + Classification tasks
From task (class) perspective
Single-Modality Data
Multi-Modality Data
Ungrouped (flat) Features
Grouped Features
From data (feature) perspective
Time Series Data and Tasks

Figure 1.1.Road map of my graduate research..

approximation. Second, sparsity can also be used given prior knowledge that the model should be sparse. In these two situations, reducing parsimony to finding models with low cardinality turns out to be limiting, and structured parsimony has emerged as a fruitful practical extension, with applications to image processing, text processing or bioinformatics. For example, structured sparsity is used to encode prior knowledge regarding network relationship between genes; it is also used as an alternative to structured nonparametric Bayesian process based priors for topic models.

1.3 Association Studies via Structured Sparse Learning

My graduate studies focused on identifying strong associations between regional imaging phenotypes as QTs and SNP genotypes as QTLs and aims to provide guidance for refined statistical modeling and follow-up studies of candidate genes or loci. My research can be roughly divided into five parts as illustrated in Fig. 1.1 and described as follows.

First, by placing the imaging biomarker identification problem under the framework of feature selection using structured sparse learning methods, we developed a joint Alzheimer's Disease (AD) classification and cognitive measurement regression model [2], which is able to identify the imaging biomarkers that are both AD-sensitive and cognition-relevant and can help reveal complex relationships among brain structure, cognition and disease status.

Second, by recognizing that existing memory performance prediction methods via regression usually do not take into account either the interconnected structures within human brains or interrelatedness among memory scores, we proposed a novel Sparse Multi-tAsk Regression and feaTure selection (SMART) method [3] to jointly analyze all the imaging and clinical data under a single regression framework and with shared underlying sparse representations, by which we achieved improved cognitive measurement prediction performance.

Third, We also developed a new imaging genetics study model [4] to capture the group structures of genetic data (*e.g.*, structures due to gene locations, genetic links, *etc.*) by proposing a novel matrix norm as regularization. Because the learning objectives of all above methods involve multiple terms non-smooth matrix norms, the formulated optimization problems are hard to solve in general. Thus we developed a series of new algorithms to solve them with rigorously proved correctness and convergence.

Fourth, because traditional association studies typically perform independent and pairwise analysis among neuroimaging measures, cognitive scores, and disease status, and ignore the important underlying interacting relationships between these units, we propose a new sparse multi-modal multi-task learning method to reveal complex relationships from gene to brain to symptom. Our main contributions are three-fold: 1) introducing combined structured sparsity regularizations into multi-modal multi-task learning to integrate multi-dimensional heterogeneous imaging genetics data and identify multi-modal biomarkers; 2) utilizing a joint classification and regression learning model to identify disease-sensitive and cognition-relevant biomarkers; 3) deriving a new efficient optimization algorithm to solve our non-smooth objective function and providing rigorous theoretical analysis on the global optimum convergency. Using the imaging genetics data from the Alzheimer's Disease Neuroimaging Initiative database, the effectiveness of the proposed method is demonstrated by clearly improved performance on predicting both cognitive scores and disease status. The identified multi-modal biomarkers could predict not only disease status but also cognitive function to help elucidate the biological pathway from gene to brain structure and function, and to cognition and disease.

Finally, we study how the SNP values change when phenotypic measures are varied. This alternative approach may have a potential to help us discover important imaging genetic associations from a different perspective. Taking into account the temporal structure of the longitudinal imaging data and the interrelatedness among the SNPs, we propose a novel task-correlated longitudinal sparse regression model to study the association between the phenotypic imaging markers and the genotypes encoded by SNPs. In our new association model, we extend the widely used $\ell_{2,1}$-norm for matrices to tensors to jointly select imaging markers that have common effects across all the regression tasks and time points, and meanwhile impose the trace-norm regularization onto the unfolded coefficient tensor to achieve low rank such that the interrelationship among SNPs can be addressed. The effectiveness of our method is demonstrated by both clearly improved prediction performance in empirical evaluations and a compact set of selected imaging predictors relevant to disease sensitive SNPs.

CHAPTER 2

SPARSE MULTI-TASK REGRESSION AND FEATURE SELECTION TO IDENTIFY IMAGING PREDICTORS FOR MEMORY PERFORMANCE

2.1 Introduction

Through employing pattern classification methods, neuroimaging has demonstrated its effectiveness in predicting Alzheimer's disease (AD) status based on individual magnetic resonance imaging (MRI) and/or positron emission tomography (PET) scans [5–7]. Because AD is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions, it is important to understand how structural and functional changes in brain can influence the performance of neuropsychological tests. As a result, regression models have been used to study whether neuroimaging measures can help predict clinical scores and track AD progression [8,9]. For example, in [8], stepwise regression was performed in a pairwise fashion to relate each MRI and FDG-PET measures of the eight candidate regions to each of the four Rey's Auditory Verbal Learning Test (RAVLT) memory scores. This approach was univariate and thereby overlooked the interrelated structures within both imaging data and clinical data. In [9], using relevance vector regression, the voxel-based morphometry (VBM) features extracted from the entire brain were jointly analyzed to predict each selected clinical score, while the investigations of different clinical scores are independent from each other.

In this chapter, we embrace, rather than ignore, the complexity of the mapping between interconnected imaging measures and interrelated clinical scores; and propose a novel Sparse Multi-tAsk Regression and feaTure selection (SMART) method to jointly analyze all the imaging and clinical data within a single regression model and common subspace. Our research focuses on investigating the relationships between MRI measures and RAVLT memory scores using the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort [10]. Instead of including all possible imaging measures to predict memory performance, the proposed SMART method is designed to select the most prominent imaging features that are able to predict memory performance with improved prediction accuracy. Different from LASSO [11] and other related methods that perform feature selection separately for each individual memory score, the proposed sparse multi-task learning model treats each memory score as a cognition task and selects imaging features that can jointly influence multiple scores/tasks. We propose to use the combined $\ell_{2,1}$-norm and ℓ_1-norm regularizations to select features with high correlations to a subset of memory scores. To demonstrate the effectiveness of the proposed SMART method, we apply it to identify relevant MRI markers that can predict multiple RAVLT memory scores. Our empirical results yield not only clearly improved prediction rates in all the test cases, but also a compact set of RAVLT-relevant MRI predictors that are in accordance with prior studies.

2.2 Sparse Multi-Task Regression and Feature Selection

Recently sparse regularizations have been applied to classification based feature selection studies. LASSO [11] was shown to efficiently select useful features for a single task. However, in our work, we expect to estimate predictive models for several related memory performance scores together, not an individual one. The multi-task feature learning [12, 13] used the $\ell_{2,1}$-norm regularization to couple feature selection across tasks using a strict assumption - all tasks share a common underlying representation. However, in many cases, the common pattern is shared by many tasks, but not all.

To address this issue, we propose a new Sparse Multi-tAsk Regression and feaTure selection (SMART) model to include both $\ell_{2,1}$-norm and ℓ_1-norm regularizations for selecting imaging features, *i.e.*, morphometric variables, and predicting memory performance. The combined convex norms help us pick up the features with high correlations to a subset of tasks. The new objective leads to a more difficult optimization problem. To address this problem, we derive a new efficient algorithm with proved global convergency. In this chapter, given a matrix M, we denoted its i-th row and j-th column as m^i and m_j, respectively.

2.2.1 Joint Sparse Regularizations Using Mixed Non-Smooth Norms

To identify the predictable correlations between memory performance scores and morphometric variables, the linear (least square) regression method is a standard way in medical image analysis research. Given the morphometric variables of n training samples $\left\{x_i \in \mathbb{R}^d\right\}_{i=1}^n$ and the associated memory scores $\{y_i \in \mathbb{R}^c\}_{i=1}^n$, traditional least square regression solves the following optimization problem to obtain the projection matrix $W \in \mathbb{R}^{d\times c}$ (the bias $\mathbf{b}$ is absorbed into W when the constant value 1 is added as an additional dimension for each data):

$$\min_W \sum_{i=1}^n \left\|W^T x_i - y_i\right\|_2^2 = \left\|X^T W - Y\right\|_F^2 , \tag{2.1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$ and $Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n\times c}$.

In the regular linear regression, the weight matrix W is not sparse. All morphometric variables are involved to the memory scores prediction. However, some of them are irrelevant to memory performance prediction. Therefore, it is desirable to select the important morphometric variables for more accurate scores prediction. To this end, instead of imposing the squared ℓ_2-norm regularization as in traditional ridge regression, we impose the $\ell_{2,1}$-norm regularization. Because the $\ell_{2,1}$-norm regularization penalizes each row of W as a whole and enforce sparsity among the rows, it is able to select the most prominent morphometric variables [14]. Specifically, we solve the following convex optimization problem:

$$\min_{W} \left\|X^T W - Y\right\|_F^2 + \gamma \left\|W\right\|_{2,1}, \tag{2.2}$$

where $\|\cdot\|_{2,1}$ denotes the $\ell_{2,1}$-norm of a matrix.

We further consider some important morphometric variables are only correlated to a subset of tasks. The $\ell_{2,1}$-norm cannot handle them properly. Thus, we add an ℓ_1-norm regularizer to impose the sparsity among all elements in W and propose our new Sparse Multi-tAsk Regression and feaTure selection (SMART) model as:

$$\min_{W} \left\|X^T W - Y\right\|_F^2 + \gamma_1 \left\|W\right\|_1 + \gamma_2 \left\|W\right\|_{2,1}. \tag{2.3}$$

Although our objective function is convex, it is difficult to be solved, because the both regularization terms are non-smooth. Here, we propose an efficient algorithm to solve our objective function in Eq. (2.3).

Taking the derivative with respect to $w_i (1 \le i \le c)$, and setting it to zero, we have

$$XX^T w_i - Xy_i + \gamma_1 D_i w_i + \gamma_2 \tilde{D} w_i = 0, \tag{2.4}$$

where $D_i (1 \le i \le c)$ is a diagonal matrix with the k-th diagonal element as $\frac{1}{2|w_{ki}|}$, $\tilde{D}$ is a diagonal matrix with the k-th diagonal element as $\frac{1}{2\|w^k\|_2}$. Thus,

$$w_i = (XX^T + \gamma_1 D_i + \gamma_2 \tilde{D})^{-1} X y_i. \tag{2.5}$$

Note that D_i and $\tilde{D}$ depend on W and thus is also unknown variables. We propose an iterative algorithm to solve this problem, which is as listed in Algorithm 3.

Algorithm 1: Algorithm to solve the proposed objective.

```
Input: $X$, $Y$
Initialize $W^1 \in \mathbb{R}^{d\times c}$, $t = 1$ ;
while not converge do
    1. Calculate the diagonal matrices $D_i^{(t)}(1 \le i \le c)$ and $\tilde{D}^{(t)}$, where the $k$-th diagonal element of $D_i^{(t)}$ is $\frac{1}{2|w_{ki}^{(t)}|}$, the $k$-th diagonal element of $\tilde{D}^{(t)}$ is $\frac{1}{2\|(w^{(t)})^k\|_2}$ ;
    2. For each $i(1 \le i \le c)$, $w_i^{(t+1)} = (XX^T + \gamma_1 D_i^{(t)} + \gamma_2 \tilde{D}^{(t)})^{-1} X y_i$ ;
    3. $t = t + 1$ ;
Output: $W^{(t)} \in \mathbb{R}^{d\times c}$.
```

2.2.2 Algorithm Analysis

Theorem 1 *Algorithm 3 decreases the objective value in each iteration.*

Proof: According to Step 2 in the algorithm, we have

$$\begin{aligned} W^{(t+1)} = \min_W & Tr(X^T W - Y)^T (X^T W - Y) \\ & + \gamma_1 \sum_{i=1}^{c} w_i^T D_i^{(t)} w_i + \gamma_2 Tr W^T \tilde{D}^{(t)} W, \end{aligned} \tag{2.6}$$

therefore we have

$$
\begin{aligned}
& Tr(X^TW^{(t+1)}-Y)^T(X^TW^{(t+1)}-Y) \\
& + \gamma_1 \sum_{i=1}^{c} (w_i^{(t+1)})^T D_i^{(t)} w_i^{(t+1)} + \gamma_2 Tr(W^{(t+1)})^T \tilde{D}^t W^{(t+1)} \\
& \leq Tr(X^TW^{(t)}-Y)^T(X^TW^{(t)}-Y) \\
& + \gamma_1 \sum_{i=1}^{c} (w_i^{(t)})^T D_i^{(t)} w_i^{(t)} + \gamma_2 Tr(W^{(t)})^T \tilde{D}^{(t)} W^{(t)} \\
\Rightarrow \quad & Tr(X^TW^{(t+1)}-Y)^T(X^TW^{(t+1)}-Y) \\
& + \gamma_1 \sum_{i=1}^{d}\sum_{j=1}^{c} \left(\frac{(w_{ij}^{(t+1)})^2}{2||w_{ij}^{(t)}||} - ||w_{ij}^{(t+1)}|| + ||w_{ij}^{(t+1)}|| \right) \\
& + \gamma_2 \sum_{k=1}^{d} \left(\frac{||(w^{(t+1)})^k||_2^2}{2||(w^{(t)})^k||_2} - ||(w^{(t+1)})^k||_2 + ||(w^{(t+1)})^k||_2 \right) \\
& \leq Tr(X^TW^{(t)}-Y)^T(X^TW^{(t)}-Y) \\
& + \gamma_1 \sum_{i=1}^{d}\sum_{j=1}^{c} \left(||w_{ij}^{(t)}|| + \frac{(w_{ij}^{(t)})^2}{2||w_{ij}^{(t+1)}||} - ||w_{ij}^{(t)}|| \right) \\
& + \gamma_2 \sum_{k=1}^{d} \left(||(w^{(t)})^k||_2 + \frac{||(w^{(t)})^k||_2^2}{2||(w^{(t)})^k||_2} - ||(w^{(t)})^k||_2 \right) \\
\Rightarrow \quad & Tr(X^TW^{(t+1)}-Y)^T(X^TW^{(t+1)}-Y) \\
& + \gamma_1 \sum_{i=1}^{d}\sum_{j=1}^{c} ||w_{ij}^{(t+1)}|| + \gamma_2 \sum_{k=1}^{d} ||(w^{(t+1)})^k||_2 \\
& \leq Tr(X^TW^{(t)}-Y)^T(X^TW^{(t)}-Y) \\
& + \gamma_1 \sum_{i=1}^{d}\sum_{j=1}^{c} ||w_{ij}^{(t)}|| + \gamma_2 \sum_{k=1}^{d} ||(w^{(t)})^k||_2
\end{aligned}
$$

The last step holds, because [14] for any vector w and w_0, we have $\|w\|_2 - \frac{\|w\|_2^2}{2\|w_0\|_2} \leq \|w_0\|_2 - \frac{\|w_0\|_2^2}{2\|w_0\|_2}$. Thus, the algorithm decreases the objective value in each iteration. □

At the convergence, $W^{(t)}$, $D_i^{(t)}(1 \leq i \leq c)$ and $\tilde{D}^{(t)}$ will satisfy the Eq. (4.9). As the problem (2.3) is a convex problem, satisfying the Eq. (4.9) indicates that W is a global optimum solution to the problem (2.3). Therefore, Algorithm 3 will converge to the global optimum of the problem (2.3). Because we have closed form solution in each iteration, our algorithm converges very fast.

2.3 Imaging and Memory Data

Both MRI and memory data used in this study were obtained from the ADNI database[1]. ADNI is a landmark investigation sponsored by the NIH and industrial partners designed to collect longitudinal neuroimaging, biological and clinical information from 822 participants that will track the neural correlates of memory loss from an early stage. Further information can be found in [15] and at www.adni-info.org. Following a previous imaging genetics study [16], 708 out of 733 non-Hispanic Caucasian participants with no missing MRI morphometric and RAVLT information were included in this study. The 708 participants are categorized by three baseline diagnostic groups: healthy control (HC, $n = 199$), mild cognitive impairment (MCI, $n = 346$) (thought to be a preclinical stage of AD), and AD ($n = 163$).

Two widely employed automated MRI analysis techniques were used to process and extract imaging measures across the entire brain from all baseline scans of ADNI participants as previously described [16]. First, voxel-based morphometry (VBM) [8] was performed to define global gray matter (GM) density maps and extract local GM density values for 86 target regions. Second, automated parcellation via FreeSurfer V4 [17] was conducted to define 56 volumetric and cortical thickness values and to extract total intracranial volume (ICV). The full descriptions about these measures are available in [16]. All these measures were adjusted for the baseline age, gender,

[1] `http://www.loni.ucla.edu/ADNI`

Table 2.1.Descriptions of RAVLT cognitive measures.

Task ID	Description
TOTAL	Total score of the first 5 learning trials
TOT6	Trial 6 total number of words recalled
TOTB	List B total number of words recalled
T30	30 minute delay total number of words recalled
RECOG	30 minute delay recognition score

education, handedness, and baseline ICV using the regression weights derived from the healthy control participants.

The cognitive measures we use to test the proposed SMART method are the baseline RAVLT memory scores from all ADNI participants [18]. The standard RAVLT format starts with a list of 15 unrelated words (List A) repeated over five different trials and participants are asked to repeat. Then the examiner presents a second list of 15 words (List B), and the participant is asked to remember as many words as possible from List A. Trial 6, termed as 5 minute recall, requests the participant again to recall as many words as possible from List A, without reading it again. Trial 7, termed as 30 minute recall, is administrated in the same way as Trial 6, but after a 30 minute delay. Finally, a recognition test with 30 words read aloud, requesting the participant to indicate whether or not each word is on List A. The RAVLT has proven useful in evaluating verbal learning and memory. The five RAVLT scores are summarized in Table 5.1.

2.4 Experimental Results and Discussions

In this section, we evaluate the proposed SMART method by applying it to the ADNI cohort, where a wide range of MRI morphometric features are examined and selected to predict memory performance measured by five RAVLT scores shown in

Table 5.1. The goal is to select a compact set of RAVLT-relevant MRI features while maintaining high predictive power.

2.4.1 Improved Memory Performance Prediction

In our experiments, we examine three different sets of morphometric variables $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^d$ for each participant, where $d = 86$ for VBM morphometric variables, $d = 56$ for FreeSurfer morphometric variables, and $d = 144$ for the combined set of VBM and Freesurfer variables. Evaluating the memory performance prediction on the three baseline diagnosis groups (HC, MCI, AD) and the group with all participants (HC + MCI + AD) using the three types of morphometric variables, we end up with a total of twelve test cases as in Table 2.2, where, *e.g.*, "FreeSurfer_HC" denotes the test case conducted on the participants of MCI group using FreeSurfer morphometric variables, and "VBM+FreeSurfer_all" denotes the test case conducted on all the participants using the combined morphometric variables by VBM and FreeSurfer.

We compare SMART against multivariate regression (MRV) in memory performance prediction. For each test case, we randomly pick 80% participants and use their morphometric variables and memory scores as training data, and perform the prediction for the remaining participants. The prediction performances assessed by root mean square error (RMSE), a widely used measurement for statistical regression analysis, are reported in Table 2.2.

A first observation on the results in Table 2.2 shows that the proposed SMART method consistently outperforms the conventional multivariate regression method in all the test cases for all the cognitive tasks. The FreeSurfer measures, VBM measures, and combined measures have similar predictive powers.

A more careful analysis shows that, using our method, it is easier to predict memory performance for AD than HC, while MCI shows an intermediate pattern.

This partially agrees with the findings in [8], which claims that MR morphormetry is not related to memory in HC, but positively related to memory functions in MCI and AD. Using multivariate regression, the above trend holds only for FreeSurfer measures. For VBM and combined cases, it is far more difficult to predict memory performance in MCI than HC and AD (11.495 *vs*. 8.651 and 7.233 for VBM, and 68.22 *vs*. 12.265 and 14.552 for VBM + FreeSurfer).

Finally, we can see that the most predictable outcome is T30 for AD group with RMSE of 1.050 for FreeSurfer, 0.904 for VBM, and 0.858 for the combined measures. Considering TOTAL is the sum of the 5 scores, the performance for AD group is decent with RMSE of 5.042 for FreeSurfer, 5.120 for VBM, and 4.805 for the combined measures. The least predictable outcome is RECOG, whose RMSEs are generally greater than 2.7.

2.4.2 Feature Selection Capabilities

The main advantage of the proposed SMART method lies in its capability to simultaneously perform regression analysis and feature selection. Besides reducing the computational complexity of the learning model as in other applications, feature selection is of significant importance in the study of neuroimaging, because it has a potential to identify the relevant imaging predictors and explain the effects of morphometric changes in relation to memory performance.

The heat map of the regression coefficients of each FreeSurfer measure w.r.t. each cognitive task (W in Eq. (2.3)) learned by SMART is shown in Fig. 2.1. The bigger the magnitude of an coefficient is, the more important the feature is in predicting the corresponding memory score. For example, “HippVol” (hippocampal volume) plays the most important role in memory performance prediction when testing on all participants, while “LatVent” (volume of lateral ventricle) is the most effective predictor when the test is conducted on AD group. The selected features by our method are marked with “x”. The heat map of regression coefficients of VBM measures are shown in Fig. 2.2. Fig. 2.3 visualizes the cortical map of selected features for prediction of TOTAL score using FreeSurfer measures in the total sample (left) and the AD sample (right).

Fig. 2.1 shows that “HippVol” is consistently selected in all the groups except AD, implicating that it is an important indicator for cognitive decline and has a potential for early detection of AD. This perfectly accords with many evidences in existing literatures [19–24]. In addition, “EntCtx” (thickness of entorhinal cortex), “Parahipp” (thickness of parahippocampal gyrus), “Precuneus” (thickness of precuneus) and “InfParietal” (thickness of inferior parietal gyrus) are also selected in different test conditions. These areas are important components of the brain’s episodic memory network [8], which has been proved to be normally engaged during episodic recall and heavily impact the memory performance [8,25,26]. Similar observations that our selections match literature evidences can also be found in both Fig. 2.1 and Fig. 2.2, which concretely confirm the effectiveness of the proposed method from neurobiological perspective.

Moreover, besides selecting common prominent features across all cognitive tasks through imposing $\ell_{2,1}$ regularization as in Eq. (2.3), we also enforce sparsity on W through ℓ_1 regularization, such that the relative importance of the selected features are properly weighted. For example, as in Fig. 2.2, the “Hippocampus” (GM density) is only selected in MCI and AD groups, but not selected by HC group. This observation, again, is extensively supported in literature. It has been shown that, in normal aging, memory, including listing learning measures with clinically applied retention intervals ($<$ 1h), appears weakly related to medial temporal lobe (MTL) [23], whereas memory has consistently been related to MLT volumes in MCI and AD [23]. This provides one more evidence showing the ability of SMART for properly identifying relevant features.

2.5 Conclusions

In this chapter, we proposed a new SMART model to perform both regression analysis for memory performance prediction and morphometric variables selection in an MCI/AD study. Different from related existing methods that ignore the interrelated structures within imaging data or those within clinical data, SMART analyzes all the imaging and clinical data within a single regression framework and common subspace, such that the predictive performance can be improved by these correlations. Our experiments using the MRI and RAVLT data of the ADNI cohort yielded promising results: (1) the prediction performance of SMART was consistently better than conventional multi-variate regression, (2) a compact set of imaging predictors were identified in each test case and were in accordance with prior findings, and (3) these selected imaging features could predict multiple memory scores at the same time and had a potential to play an important role in determine cognitive functions and characterizing AD progression. These promising results were consistent with our theoretical foundation and prior studies, which demonstrated the effectiveness of the proposed method.

Table 2.2.Prediction performance measured by RMSE.

Test cases		TOTAL	TOT6	TOTB	T30	RECOG
FreeSurfer_HC	MVR	8.762	4.362	3.281	4.305	4.021
	SMART	6.645	2.940	2.235	2.806	3.621
FreeSurfer_MCI	MVR	6.998	2.765	2.399	2.480	3.427
	SMART	5.600	1.990	1.953	1.709	3.181
FreeSurfer_AD	MVR	5.897	1.768	2.058	1.382	3.390
	SMART	5.042	1.452	1.716	1.050	2.830
FreeSurfer_all	MVR	5.926	2.238	2.036	2.090	3.342
	SMART	5.736	2.139	1.961	1.966	3.196
VBM_HC	MVR	8.651	3.772	2.885	3.496	4.776
	SMART	6.705	2.844	2.139	2.656	3.584
VBM_MCI	MVR	11.495	4.256	4.621	4.032	5.598
	SMART	5.584	1.832	1.931	1.669	3.017
VBM_AD	MVR	7.223	2.162	2.622	1.479	4.163
	SMART	5.120	1.518	1.826	0.904	2.781
VBM_all	MVR	6.090	2.290	2.140	2.141	3.396
	SMART	5.718	2.103	1.993	1.921	3.182
VBM+FreeSurfer_HC	MVR	12.265	5.416	4.349	5.089	6.703
	SMART	6.664	2.829	2.230	2.683	3.577
VBM+FreeSurfer_MCI	MVR	68.222	26.146	23.489	30.033	34.306
	SMART	5.533	1.901	1.869	1.606	3.114
VBM+FreeSurfer_AD	MVR	14.552	4.307	5.141	4.297	8.430
	SMART	4.805	1.218	1.731	0.858	2.865
VBM+FreeSurfer_all	MVR	6.505	2.596	2.258	2.540	3.582
	SMART	5.809	2.208	2.000	2.051	3.214

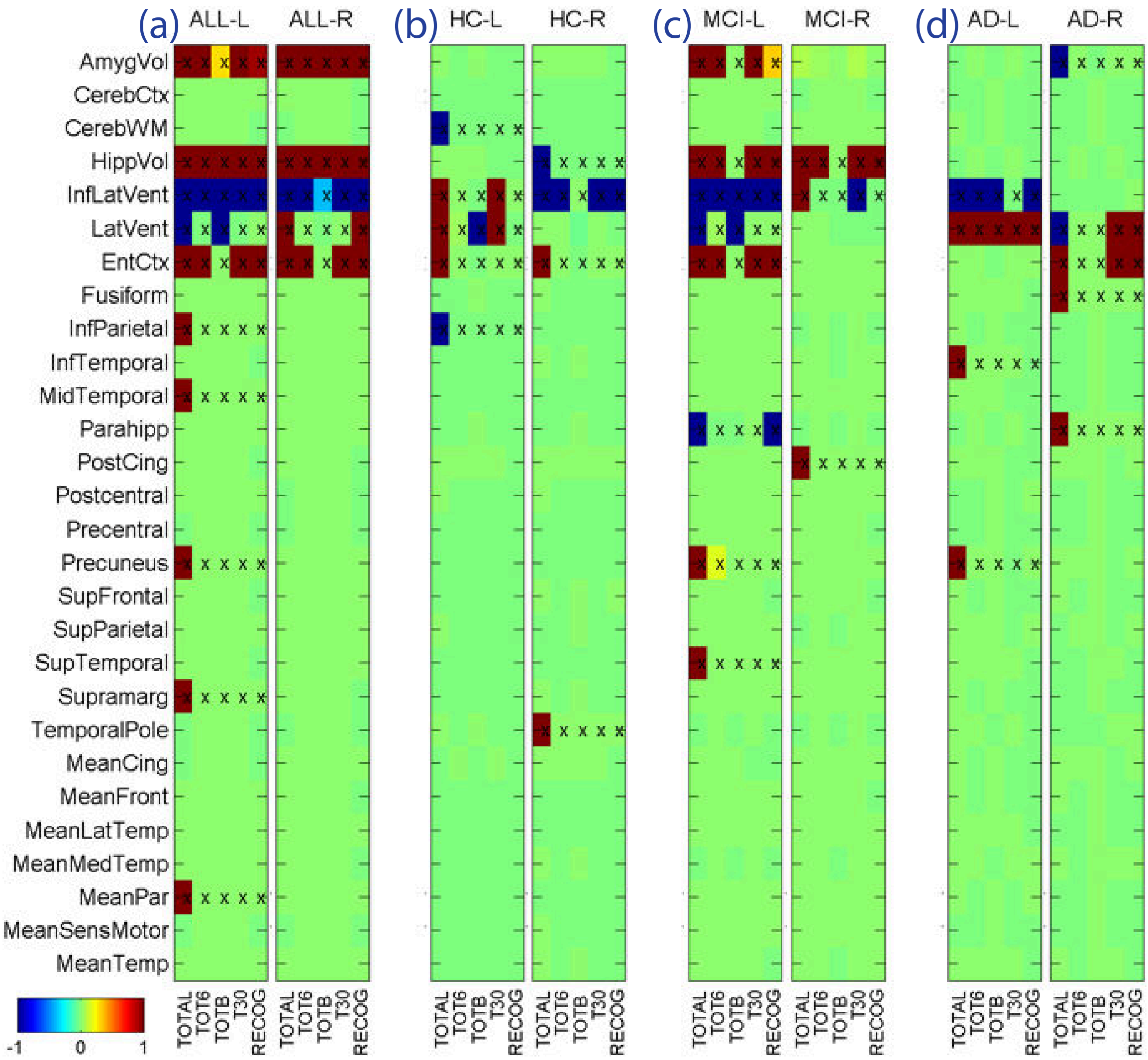


Figure 2.1. Heat map of selected features for prediction using FreeSurfer measures in (a) the total sample, (b) HC, (c) MCI, and (d) AD. In each of (a-d), regression weights (*i.e.*, coefficients) for left and right measures are visualized as two separate panels, where columns in each panel correspond to different memory scores. Since our method selects features with absolute values ≥ 1, the range of the color map is limited to [-1,1] for a more effective visualization. All selected features are marked with “x”. .

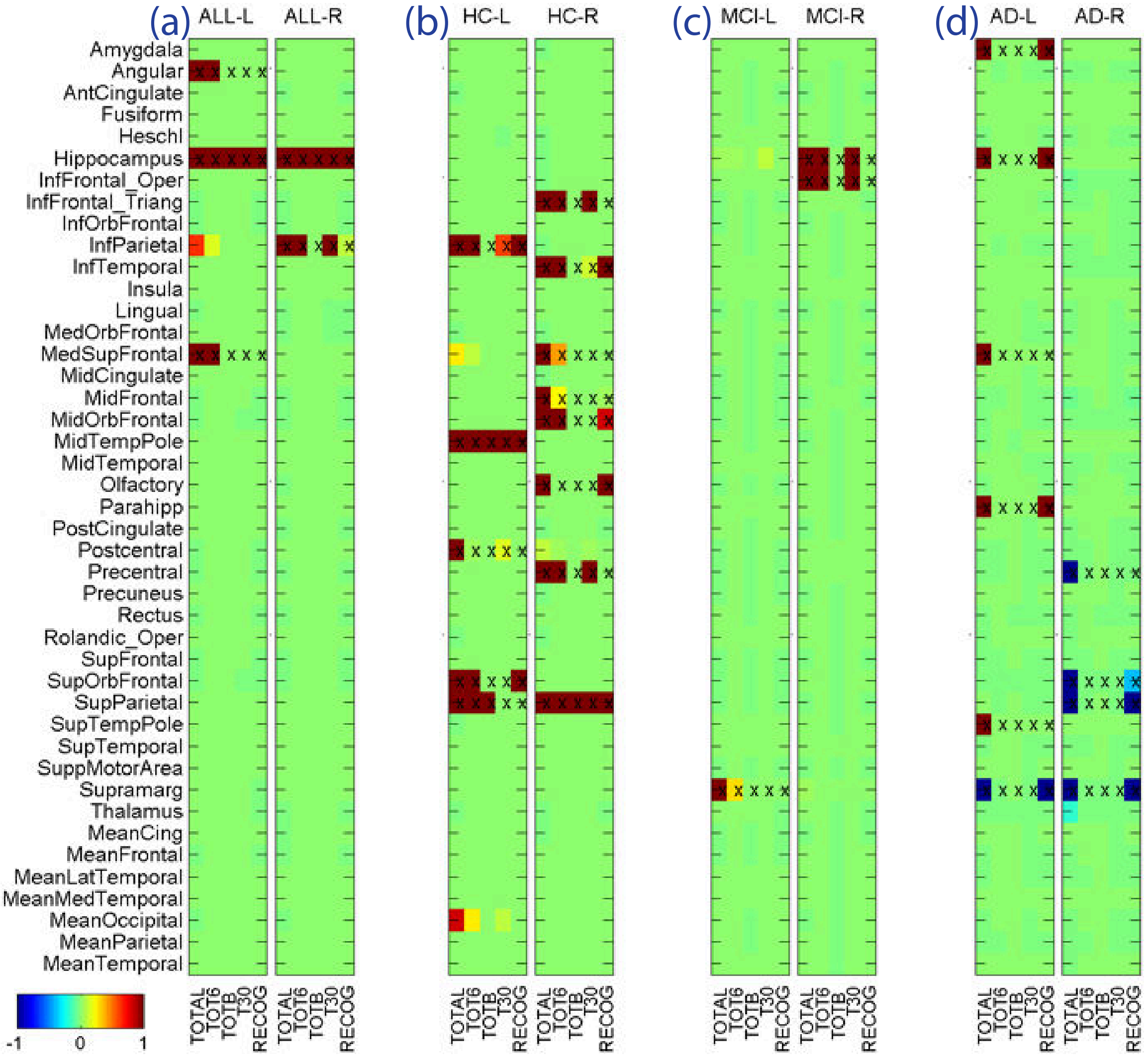


Figure 2.2. Heat map of selected features for prediction using VBM measures in (a) the total sample, (b) HC, (c) MCI, and (d) AD. In each of (a-d), regression weights (*i.e.*, coefficients) for left and right measures are visualized as two separate panels, where columns in each panel correspond to different memory scores. Since our method selects features with absolute values ≥ 1, the range of the color map is limited to [-1,1] for a more effective visualization. All selected features are marked with “x”. .

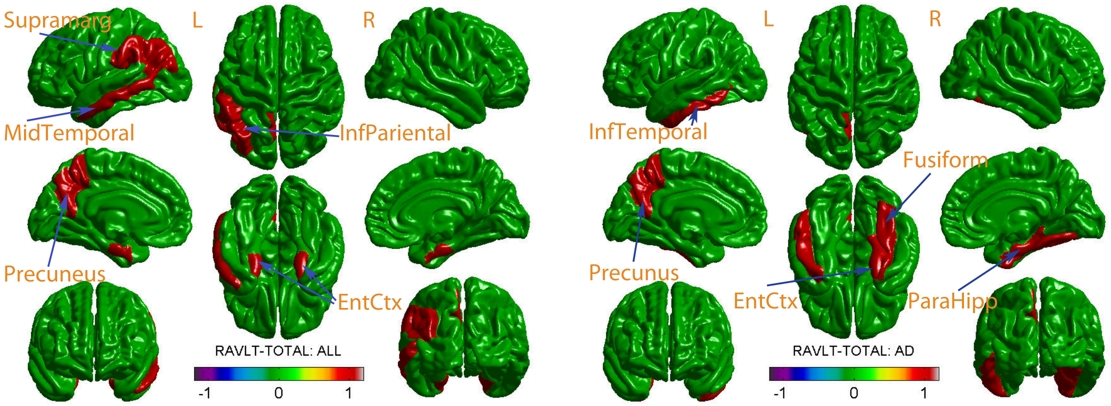


Figure 2.3. Cortical map of selected features for prediction using FreeSurfer measures in the total sample (left) and the AD sample (right). Each map only visualizes the regression weights for RAVLT-TOTAL score for individual cortical thickness measures (i.e., volume measures and mean thickness measures of larger regions are not included). Since our method selects features with absolute values ≥ 1, the range of the color map is limited to $[-1, 1]$ for a more effective visualization. .

CHAPTER 3

IDENTIFYING AD-SENSITIVE AND COGNITION-RELEVANT IMAGING BIOMARKERS VIA JOINT CLASSIFICATION AND REGRESSION

3.1 Introduction

Neuroimaging is a powerful tool for characterizing neurodegenerative process in the progression of Alzheimer's disease (AD). Pattern classification methods have been widely employed to predict disease status using neuroimaging measures [**?**, 27]. Since AD is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions, regression models have been investigated to predict clinical scores from individual magnetic resonance imaging (MRI) and/or positron emission tomography (PET) scans [8, 9]. For example, in [8], stepwise regression was performed in a pairwise fashion to relate each of MRI and FDG-PET measures of eight candidate regions to each of four Rey's Auditory Verbal Learning Test (RAVLT) memory scores.

Predicting disease status and predicting memory performance, using neuroimaging data, are both important learning tasks. Prior research typically studied these tasks *separately.* One example is to first determine disease-relevant cognitive scores and then identify imaging biomarkers associated with these scores so that interesting pathways from brain structure to cognition to symptom can potentially be discovered. However, a specific cognitive function could be related to multiple imaging measures associated with different biological pathways (some of them are not related to AD). As a result, the identified imaging biomarkers are not necessarily all disease specific. To have a better understanding of the underlying mechanism specific to AD, an interesting topic would be to only discover imaging biomarkers associated with both cognitive function and AD status.

To identify AD-sensitive and cognition-relevant imaging biomarkers, we propose a new joint classification and regression learning model to simultaneously performing two heterogeneous tasks, *i.e.*, imaging-to-disease classification and imaging-to-cognition regression. We use magnetic resonance imaging (MRI) measures as predictors and cognitive memory scores and disease status as response variables. For each individual regression or classification task, we employ a multi-task learning model in which tasks for predicting different memory performances (or those for predicting AD and control dummy variables in classification) are considered as homogeneous tasks. Different to LASSO and other related methods that mainly find the imaging features correlated to each individual memory score, our method selects the imaging features that tend to play an important role on influencing multiple homogenous tasks.

Our new method utilizes the sparse regularization to perform imaging biomarker selection and learn a sparse parameter matrix under a unified framework that integrates both heterogeneous and homogenous tasks. Specifically, by recognizing that the formation and maintenance [28] of memory are synergically accomplished by a few brain areas, such as medial temporal lobe structures, medial and lateral parietal, as well as prefrontal cortical areas, we use the $\ell_{2,1}$-norm regularization to select features that can predict most memory scores and classify AD versus control. Empirical comparison with the existing methods demonstrates that the proposed method not only yields improved performance on predicting both cognitive scores and disease status, but also discovers a small set of AD-sensitive and cognition-relevant biomarkers in accordance with prior findings.

3.2 Sparse Model for Joint Classification and Regression

When we study either regression or classification via a multi-task learning model, given a set of input variables, (*i.e.*, features, such as imaging biomarkers), we are interested in learning a set of related models (*e.g.*, associations between image biomarkers and cognitive scores) for predicting multiple homogenous tasks (such as predicting cognitive scores). Since these homogenous tasks are typically interrelated, they share a common input space. As a result, it is desirable to learn all the models jointly rather than treating each task as an independent one. Such multi-task learning methods can help discover robust patterns, especially when significant patterns in a single task become outliers for other tasks, and potentially increase the predictive power.

To identify AD-sensitive and cognition-relevant biomarkers from imaging data, we formulate a new problem to jointly learn two heterogeneous tasks: classification and regression. We propose a new sparse model for joint classification and regression to perform multivariate regression for cognitive memory scores predictions and logistic regression for disease classification tasks simultaneously.

Notation. In this chapter, we write matrices and vectors as bold uppercase and lowercase letters respectively. Given a matrix $\mathbf{M} = [m_{ij}]$, we denote its i-th row as $\mathbf{m}^i$ and j-th column as $\mathbf{m}_j$. The Frobenius norm of the matrix $\mathbf{M}$ is denoted as $\|\mathbf{M}\|_{\mathrm{F}}$, and the $\ell_{2,1}$-norm [14] of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{2,1} = \sum_i \sqrt{\sum_j m_{ij}^2} = \sum_i \|\mathbf{m}^i\|_2$.

3.2.1 Objective of Sparse Joint Classification and Regression

First, logistic regression is used for disease classification. Given the training data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, each data point $\mathbf{x}_i$ is associated with a label vector $\mathbf{y}^i = [y_{i1}, \ldots, y_{ic_1}] \in \mathbb{R}^{c_1}$. If $\mathbf{x}_i$ belongs to the k-th class, $y_{ik} = 1$, otherwise $y_{ik} = 0$. We write $\mathbf{Y} = \left[\left(\mathbf{y}^1\right)^T, \ldots, \left(\mathbf{y}^n\right)^T \right]^T \in \mathbb{R}^{n \times c_1}$. In traditional multi-class logistic regression, under a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times c_1}$, we have

$$p\left(k \mid \mathbf{x}_i, \mathbf{W}\right) = \frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{\sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i}} \quad \Longrightarrow \quad p\left(\mathbf{y}^i \mid \mathbf{x}_i, \mathbf{W}\right) = \prod_{k=1}^{c_1} \left(\frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{\sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i}} \right)^{y_{ik}},$$

where $p\left(k \mid \mathbf{x}_i, \mathbf{W}\right)$ is the probability that $\mathbf{x}_i$ belongs to the k-th class, and $p\left(\mathbf{y}^i \mid \mathbf{x}_i, \mathbf{W}\right)$ is the probability that $\mathbf{x}_i$ is associated with the given label $\mathbf{y}^i$. Therefore, the multi-class logistic loss that maximizes the Log-likelihood can be achieved by minimizing:

$$l_1\left(\mathbf{W}\right) = -\log \prod_{i=1}^{n} p\left(\mathbf{y}^i \mid \mathbf{x}_i, \mathbf{W}\right) = \sum_{i=1}^{n} \sum_{k=1}^{c_1} \left(y_{ik} \log \sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i} - y_{ik} \mathbf{w}_k^T \mathbf{x}_i \right). \tag{3.1}$$

In AD classification, we have two classes, *i.e.*, AD and health control (HC).

Second, we use multivariate least square regression to predict cognitive scores, which minimizes:

$$l_2\left(\mathbf{P}\right) = \left\| \mathbf{X}^T \mathbf{P} - \mathbf{Z} \right\|_{\mathrm{F}}^2, \tag{3.2}$$

where $\mathbf{X}$ is the data matrix, $\mathbf{Z} = \left[\left(\mathbf{z}^1\right)^T, \ldots, \left(\mathbf{z}^n\right)^T \right]^T \in \mathbb{R}^{n \times c_2}$ is the label matrix for the c_2 regression tasks, and $\mathbf{P} \in \mathbb{R}^{d \times c_2}$ is the projection matrix.

The objective for joint classification and regression to identify AD-sensitive and cognition-relevant imaging biomarkers can now be formulated as follows:

$$\min \; J\left(\mathbf{V}\right) = l_1\left(\mathbf{W}\right) + l_2\left(\mathbf{P}\right) + \gamma \left\| \mathbf{V} \right\|_{2,1}, \tag{3.3}$$

where $\mathbf{V} = [\mathbf{W} \; \mathbf{P}] \in \mathbb{R}^{d \times (c_1 + c_2)}$. Thanks to the $\ell_{2,1}$-norm regularization on $\mathbf{V}$, the biomarkers are identified across all tasks so that they are not only correlated to cognitive scores but also discriminative to disease status.

3.2.2 An Efficient Iterative Algorithm

Due to the non-smoothness of the $\ell_{2,1}$-norm term, J in Eq. (3.3) is hard to solve in general. Thus we derive an efficient iterative algorithm as follows.

Taking the derivatives of J w.r.t. $\mathbf{W}$ and $\mathbf{P}$, we set them to be zeros:

$$\frac{\partial J}{\partial \mathbf{W}} = \frac{\partial l_1(\mathbf{W})}{\partial \mathbf{W}} + 2\gamma \mathbf{D}\mathbf{W} = 0, \quad \frac{\partial J}{\partial \mathbf{P}} = 2\mathbf{X}\mathbf{X}^T\mathbf{P} - 2\mathbf{X}\mathbf{Z} + 2\gamma \mathbf{D}\mathbf{P} = 0, \tag{3.4}$$

where $\mathbf{D}$ is a diagonal matrix whose k-th diagonal element is $\frac{1}{2\left\|\mathbf{v}^k\right\|_2}$. Because $\mathbf{D}$ depends on $\mathbf{V}$, it is also an unknown variable. Following standard optimization procedures in statistical learning, we alternately optimize $\mathbf{V}$ and $\mathbf{D}$.

Algorithm 2: An efficient algorithm to solve Eq. (3.3).

```
Input: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times c_1}$, and
        $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^T \in \mathbb{R}^{n \times c_2}$.
1. Initialize $\mathbf{W} \in \mathbb{R}^{d \times c_1}$, $\mathbf{P} \in \mathbb{R}^{d \times c_2}$, and let $\mathbf{V} = [\mathbf{W}\ \mathbf{P}] \in \mathbb{R}^{d \times (c_1 + c_2)}$ ;
while not converge do
    2. Calculate the diagonal matrix $\mathbf{D}$, of which the $k$-th element is $\frac{1}{2\left\|\mathbf{v}^k\right\|_2}$;
    3. Update $\mathbf{w}$ by $\mathbf{w} - \mathbf{B}^{-1}\mathbf{a}$, where $(d \times (p-1) + u)$-th element of $\mathbf{a} \in \mathbb{R}^{dc_1 \times 1}$ is
       $\frac{\partial\left[l_1(\mathbf{W}) + \gamma\, \mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right)\right]}{\partial \mathbf{W}_{up}}$ for $1 \le u \le d, 1 \le p \le c_1$, the
       $(d \times (p-1) + u, d \times (q-1) + v)$-th element of $\mathbf{B} \in \mathbb{R}^{dc_1 \times dc_1}$ is
       $\frac{\partial\left[l_1(\mathbf{W}) + \gamma\, \mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right)\right]}{\partial \mathbf{W}_{up} \partial \mathbf{W}_{vq}}$ for $1 \le u, v \le d$ and $1 \le p, q \le c_1$. Construct the updated
       $\mathbf{W} \in \mathbb{R}^{d \times c_1}$ by the updated vector $\mathbf{w} \in \mathbb{R}^{dc_1}$, where the $(u,p)$-th element of $\mathbf{W}$
       is the $(d \times (p-1) + u)$-th element of $\mathbf{w}$;
    4. Update $\mathbf{P}$ by $\mathbf{P} = \left(\mathbf{X}\mathbf{X}^T + \gamma\mathbf{D}\right)^{-1}\mathbf{X}\mathbf{Z}$;
    5. Update $\mathbf{V}$ by $\mathbf{V} = [\mathbf{W}\ \mathbf{P}]$;
end
Output: $\mathbf{W} \in \mathbb{R}^{d \times c_1}$ and $\mathbf{P} \in \mathbb{R}^{d \times c_2}$.
```

First, we randomly initialize $\mathbf{V} \in \mathbb{R}^{d \times (c_1 + c_2)}$, upon which we calculate $\mathbf{D}$. After obtaining $\mathbf{D}$, we update the solution $\mathbf{V} = [\mathbf{W}\ \mathbf{P}]$ using Eq. (3.4). To be more precise, $\mathbf{P}$ is updated by $\mathbf{P} = \left(\mathbf{X}\mathbf{X}^T + \gamma\mathbf{D}\right)^{-1}\mathbf{X}\mathbf{Z}$. Because we cannot update $\mathbf{W}$ with a closed form solution upon Eq. (3.4), we employ Newton's method to obtain updated $\mathbf{W}$ by solving the following problem: $\min_{\mathbf{W}} l_1(\mathbf{W}) + \gamma\, \mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right)$.

Once we obtain the updated $\mathbf{V} = [\mathbf{W}\ \mathbf{P}]$, we can calculate $\mathbf{D}$. This procedure repeats until convergence. The detailed algorithm is summarized in Algorithm 3, whose convergence is proved as following.

Lemma 1 *For any vector* $\mathbf{v}$ *and* $\mathbf{v}_0$*, we have* $\|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}_0\|_2} \le \|\mathbf{v}_0\|_2 - \frac{\|\mathbf{v}_0\|_2^2}{2\|\mathbf{v}_0\|_2}$. *Proof is available in [14].*

Theorem 2 *Algorithm 3 decreases the objective value of* J *in every iteration.*

Proof. In each iteration, denote the updated $\mathbf{W}$ as $\widetilde{\mathbf{W}}$, the updated $\mathbf{P}$ as $\widetilde{\mathbf{P}}$, thus the updated $\mathbf{V}$ is $\widetilde{\mathbf{V}} = \left[\widetilde{\mathbf{W}}\ \widetilde{\mathbf{P}}\right]$. According to step 3 of Algorithm 3, we have

$$l_1\left(\widetilde{\mathbf{W}}\right) + \gamma\, \mathbf{tr}\left(\widetilde{\mathbf{W}}^T\mathbf{D}\widetilde{\mathbf{W}}\right) \le l_1(\mathbf{W}) + \gamma\, \mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right). \tag{3.5}$$

According to step 4 we know that

$$l_2\left(\widetilde{\mathbf{P}}\right) + \gamma\, \mathbf{tr}\left(\widetilde{\mathbf{P}}^T\mathbf{D}\widetilde{\mathbf{P}}\right) \le l_2(\mathbf{P}) + \gamma\, \mathbf{tr}\left(\mathbf{P}^T\mathbf{D}\mathbf{P}\right). \tag{3.6}$$

According to the definition of $\mathbf{D}$ and Lemma 1, we have the following inequality:

$$\begin{aligned} &\sum_{k=1}^{d}\left\|\tilde{\mathbf{v}}^k\right\|_2 - \sum_{k=1}^{d}\frac{\left\|\tilde{\mathbf{v}}^k\right\|_2^2}{2\left\|\mathbf{v}^k\right\|_2} \le \sum_{k=1}^{d}\left\|\mathbf{v}^k\right\|_2 - \sum_{k=1}^{d}\frac{\left\|\mathbf{v}^k\right\|_2^2}{2\left\|\mathbf{v}^k\right\|_2} \\ \Rightarrow\quad &\gamma\sum_{k=1}^{d}\left\|\tilde{\mathbf{v}}^k\right\|_2 - \gamma\, \mathbf{tr}\left(\widetilde{\mathbf{V}}^T\mathbf{D}\widetilde{\mathbf{V}}\right) \le \gamma\sum_{k=1}^{d}\left\|\mathbf{v}^k\right\|_2 - \gamma\, \mathbf{tr}\left(\mathbf{V}^T\mathbf{D}\mathbf{V}\right). \end{aligned} \tag{3.7}$$

Because $\mathbf{tr}\left(\mathbf{V}^T\mathbf{D}\mathbf{V}\right) = \mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right) + \mathbf{tr}\left(\mathbf{P}^T\mathbf{D}\mathbf{P}\right)$, by adding Eqs. (3.5–3.7) at the both sides, we arrive at

$$l_1\left(\widetilde{\mathbf{W}}\right) + l_2\left(\widetilde{\mathbf{P}}\right) + \gamma\sum_{k=1}^{d}\left\|\tilde{\mathbf{v}}^k\right\|_2 \le l_1(\mathbf{W}) + l_2(\mathbf{P}) + \gamma\sum_{k=1}^{d}\left\|\mathbf{v}^k\right\|_2 \tag{3.8}$$

Thus, Algorithm 3 decreases the value of J in Eq. (3.3) in every iteration. □ □

Because J in Eq. (3.3) is obviously lower-bounded by 0, Theorem 1 guarantees the convergence of Algorithm 3. In addition, because J is convex, Algorithm 3 converges at the global optimum of the problem.

3.3 Experimental Results

We evaluate our method by applying it to the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. The goal is to select a compact set of AD-sensitive and cognition-relevant imaging biomarkers while maintaining high predictive power.

Data preparation. We downloaded data from the ADNI database (`http://adni.loni.ucla.edu`). We used baseline MRI data, from which we extracted 56 volumetric and cortical thickness values (Fig. 5.4) using FreeSurfer (`http://surfer.nmr.mgh.harvard.edu`), as described in [16]. We included memory scores from three different cognitive assessments including Mini-Mental State Exam (MMSE), Rey's Auditory Verbal Learning Test (RAVLT), and TRAILS. Details about these assessments are available in the ADNI procedure manuals (`http://www.adni-info.org/Scientists/ProceduresManuals.aspx`).

3.3.1 Biomarker Identification

The proposed method aims to identify imaging biomarkers that are associated with both disease status and cognitive scores in a joint classification and regression framework. Here we first examine the identified biomarkers. Fig. 5.4 shows a summarization of selected features for the three experiments (one for each type of cognitive scores) where the regression/classification weights are color-mapped for each feature and each task. Fig. 3.2 visualizes the cortical maps of selected features for both classification and regression in different tasks.

Fig. 5.4 and Fig. 3.2 show that a small set of MRI measures are identified, including hippocampal volume (HippVol), entorhinal cortex thickness (EntCtx), amygdala volume (AmygVol), inferior parietal gyrus thickness (InfParietal), and middle temporal gyrus thickness (MidTemporal). These are all well-known AD-relevant biomarkers.

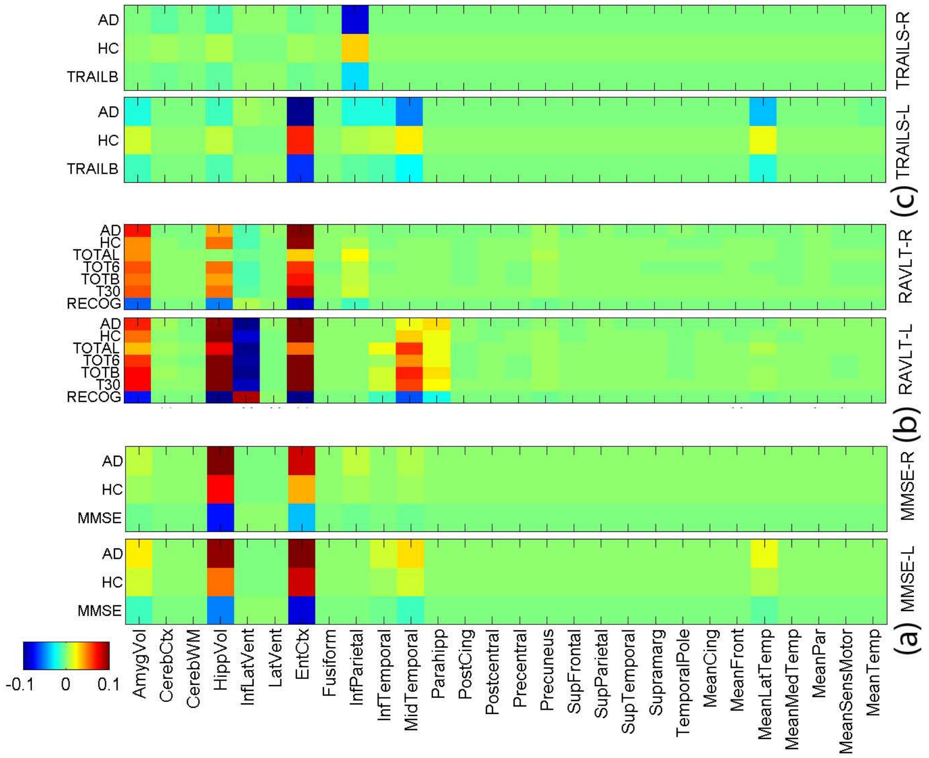


Figure 3.1. Weight maps of the joint classification and regression tasks. One binary classification task for AD and HC. Three different groups of cognitive scores for regression: (a) MMSE score, (b) RAVLT score, (c) TRAILS score. "-L" indicates the FreeSurfer biomarkers at the left side, and "-R" indicates those at the right side. .

Our method also shows that these markers are jointly associated with one or more memory scores. Although we know that MRI measures, cognitive scores and diagnosis are highly correlated, the complex relationships among them remain to be discovered for a better understanding of AD mechanism. This is one major focus of our work. As shown in Fig. 5.4, different AD-sensitive MRI measures could be related to different cognitive tasks. The proposed sparse method for joint classification and regression enables us to sort out MRI-cognition relationships while focusing on AD-sensitive markers.

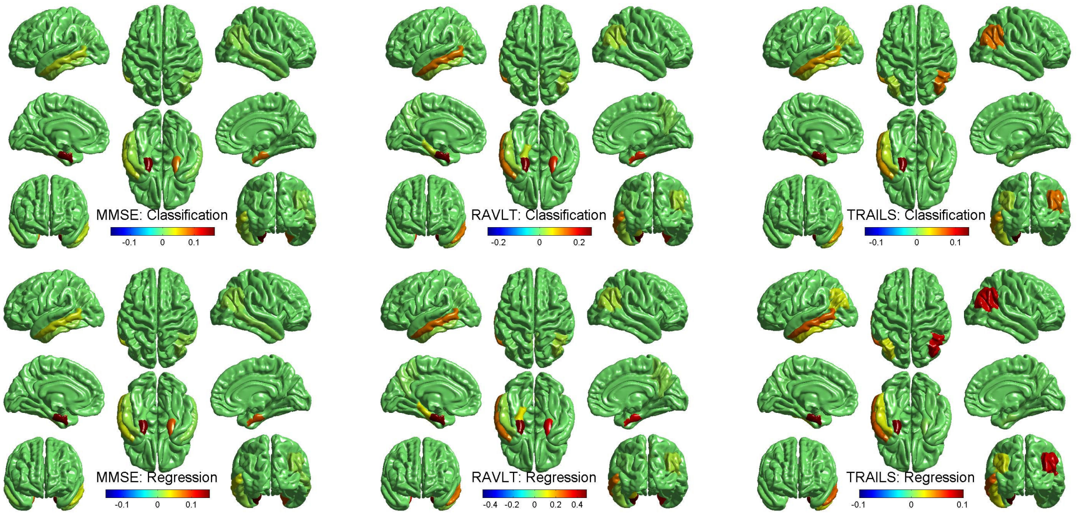


Figure 3.2. Cortical map of selected features for cognitive score prediction using FreeSurfer measures in the three joint classification and regression tasks..

3.3.2 Improved Prediction Performance

Now we evaluate the performance of joint classification and regression for AD detection and cognitive score prediction using MRI data. We performed standard 5-fold cross-validation, where the parameter γ of our method in Eq. (3.3) was fine tuned in the range of $\{10^{-5}, \ldots, 1, \ldots, 10^{5}\}$ by an internal 5-fold cross-validation in the training data of each of the 5 trials. For classification, we compared the proposed method against two baseline methods including logistic regression and support vector machine (SVM). For SVM, we implemented three different kernels including linear, polynomial and Gaussian kernels. For polynomial kernel, we searched the best results when the polynomial order varied in the range of $\{1, 2, \ldots, 10\}$; for Gaussian kernel, we fine tuned the parameter α in the same range as that for our method and fixed parameter C as 1. For regression, we compared our method against two widely used methods including multivariate regression and ridge regression. For the latter, we fine tuned its parameter in the same range as that for our method. The results are reported in Table 3.1.

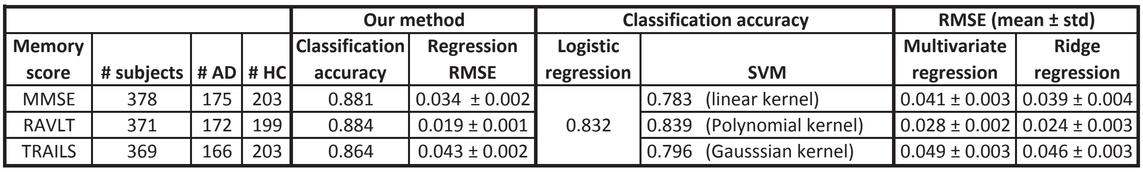
Table 3.1 shows that our method performs clearly better than both logistic regression and SVM, which are consistent with our motivations in that our method classifies participants using the information from not only MRI measures but also the reinforcement by cognitive score regression. In addition, the cognitive score regression performances of our method measured by root mean squared error (RMSE) outperform both multivariate regression and ridge regression, supporting the usefulness of joint classification and regression from another perspective. Ridge regression achieves close but slightly worse regression performance. However, it lacks the ability to identify relevant imaging markers. All these observations demonstrate the effectiveness of the proposed method in improving the performances of both AD detection and cognitive score prediction.

Mild cognitive impairment (MCI) is thought to be the prodromal stage of AD. Including MCI in this type of analyses will be an interesting future direction to help biomarker discovery for early detection of AD. We performed an initial analyis on three-class classification for AD, MCI and HC: the accuracy of our method was 0.663 and the best of other tested methods was 0.615. Apparently this is a much harder task and warrants further thorough investigation.

3.4 Conclusions

In this chapter, we have proposed a new sparse model for joint classification and regression and applied it to the ADNI cohort for identifying AD-sensitive and cognition-relevant imaging biomarkers. Our methodological contributions are threefold: 1) proposing a new learning model, joint classification and regression learning, to identify disease-sensitive and task-relevant biomarkers for analyzing multimodal data; 2) employing structural sparsity regularization to integrate heterogenous and homogenous tasks in a unified multi-task learning framework; 3) deriving a new effi-

Table 3.1.Comparison of classification and regression performance.



				Our method		Classification accuracy		RMSE (mean ± std)	
Memory score	# subjects	# AD	# HC	Classification accuracy	Regression RMSE	Logistic regression	SVM	Multivariate regression	Ridge regression
MMSE	378	175	203	0.881	0.034 ± 0.002	0.832	0.783 (linear kernel)	0.041 ± 0.003	0.039 ± 0.004
RAVLT	371	172	199	0.884	0.019 ± 0.001		0.839 (Polynomial kernel)	0.028 ± 0.002	0.024 ± 0.003
TRAILS	369	166	203	0.864	0.043 ± 0.002		0.796 (Gausssian kernel)	0.049 ± 0.003	0.046 ± 0.003

cient optimization algorithm to solve our non-smooth objective function, and coupling this with rigorous theoretical analysis on global optimum convergency. Empirical comparison with the existing methods demonstrates that our method not only yields improved performance on predicting both cognitive scores and disease status using MRI data, but also discovers a small set of AD-sensitive and cognition-relevant imaging biomarkers in accordance with prior findings.

CHAPTER 4

IDENTIFYING QUANTITATIVE TRAIT LOCI VIA GROUP-SPARSE MULTI-TASK REGRESSION AND FEATURE SELECTION

4.1 Introduction

Imaging genetics is an emergent transdisciplinary research field, where the associations between genetic variations and imaging measures as quantitative traits (QTs) or continuous phenotypes are evaluated. Compared to case-control status, the QTs have increased statistical power and are closer to the underlying biological etiology of the disease making it easier to identify underlying genes [16,29]. Genome-wide association studies (GWAS) have been increasingly performed to correlate high-throughput single nucleotide polymorphism (SNP) data to large-scale image data. While many studies employed a hypothesis-driven approach by making significant reduction in one or both data types [30], some recent studies examined these associations at the whole genome entire brain level [16]. Pairwise univariate analysis was typically used in traditional association studies to quickly provide important association information between SNPs and QTs. However, it treated the SNPs and the QTs as independent and isolated units, therefore the underlying interacting relationships between the units might be lost. Multivariate methods to examine joint effect of multi-locus genotype on a single phenotype were studied in general genetic association studies as well as several recent imaging genetic studies. This paradigm did not consider the relationship between interlinked brain phenotypes and thus still had limited power in revealing complex imaging genetic associations. In this work, taking into account the interrelated structure within and between SNPs and QTs, we propose a new frame-

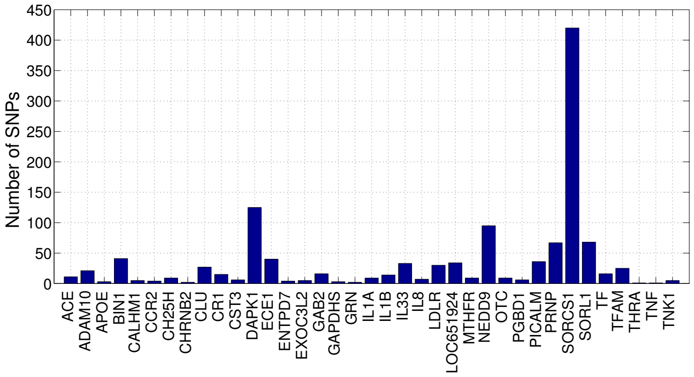


Figure 4.1. Top 37 AD risk factor genes used in this study and the numbers of their SNPs..

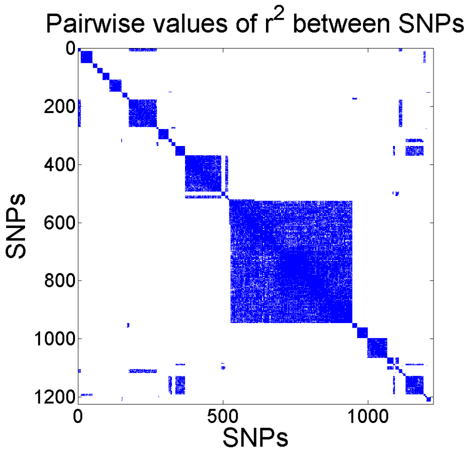


Figure 4.2. Pairwise LD correlation coefficients ($r^2 > 0.2$ in blue) among the 1224 SNPs used in this study. The SNPs clearly form groups. .

work for effectively identifying quantitative trait loci, which addresses the following challenges in imaging genetics association study.

First, traditional association studies consider all the SNPs evenly distributed and assess each SNP individually. However, certain SNPs are naturally connected via different pathways. Multiple SNPs from one gene often jointly carry out genetic functionalities. Moreover, linkage disequilibrium (LD) describes the non-random association between alleles at different loci, through which the SNPs in high LD are

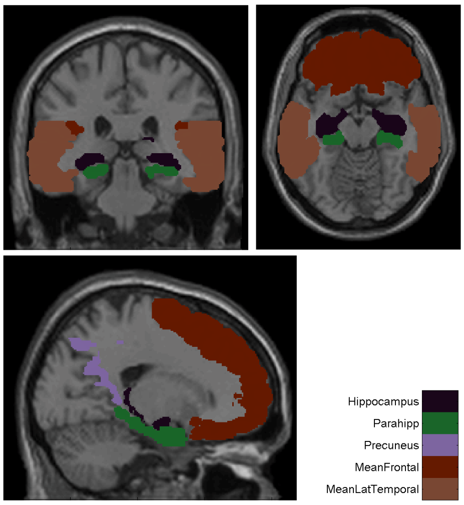


Figure 4.3.VBM ROIs used in this study are mapped onto a brain..

linked together in meiosis. Thus, instead of treating SNPs in an isolated manner, it would be beneficial to exploit the group structure among SNPs.

Second, because the functionality of the human brain typically involves more than one cerebral component, investigating each individual regional brain phenotype separately will inevitably lose the interacting relationships between them. For example, the brain's episodic memory network, including medial temporal lobe (MTL) structures, medial and lateral parietal, and prefrontal cortical areas, are normally engaged together during episodic recall [8]. In addition, accurate prediction of disease status and progression are typically implicated by multiple brain regions coupled with other biomarkers. Therefore, jointly analyzing all the imaging phenotypes via one single integral regression model is desirable to elucidate the shared mechanism that may be hidden otherwise.

By recognizing the interrelated nature of these genotypes and phenotypes, in this study, we propose a novel Sparse Multi-tAsk Regression and feaTure selection (SMART) method to identify quantitative trait loci in a mild cognitive impairment (MCI) and Alzheimer's disease (AD) study using a few important imaging QTs relevant to AD. We consider each SNP as a *feature* and each QT as a *response variable* (i.e., a *learning task*), and formulate a multi-task regression framework including multiple features (SNPs) and multiple responses (QTs). Our goal is to reveal the relationships between these genetic features and imaging phenotypes.

The proposed model consists of three major components. First, it is built upon regression analysis due to the continuous responses of the imaging phenotypes. As a result, the regression coefficients assess the relationships between SNPs and QTs. Second, in order to address the group-wise association among SNPs, inspired by group Lasso [31], we propose a new form of regularization, called as *group $\ell_{2,1}$-norm ($G_{2,1}$-norm) regularization*, in which the coefficients of the SNPs within a pre-defined group, with respect to all the QTs, are penalized as a whole via ℓ_2-norm, while ℓ_1-norm is used to sum up the group-wise penalties to enforce sparsity between groups [11]. The latter is important because in reality only a small fraction of genotypes are related to a specific phenotype. Moreover, with sparsity, outliers and irrelevant associations are inherently removed. Lastly, through enforcing $\ell_{2,1}$-norm regularization, feature selection becomes an integrated procedure across multiple learning tasks [12,13], such that the interrelationships among different imaging phenotypes are leveraged. Note that the proposed $G_{2,1}$-norm and the enforced $\ell_{2,1}$-norm couple a set of learning tasks together such that the regression analysis can be carried out jointly across all the QTs, whereas Lasso [11] and group Lasso [31] perform regression analysis separately, one task at a time.

We apply the proposed SMART method to the ADNI cohort [10] for identifying quantitative trait loci (QTLs) in MCI and AD using a set of imaging phenotypes known to be relevant to AD. Our empirical results yield not only clearly improved prediction performance in all test cases, but also a compact set of SNP predictors relevant to the imaging genotypes that are in accordance with prior studies.

4.2 Materials and Data Sources

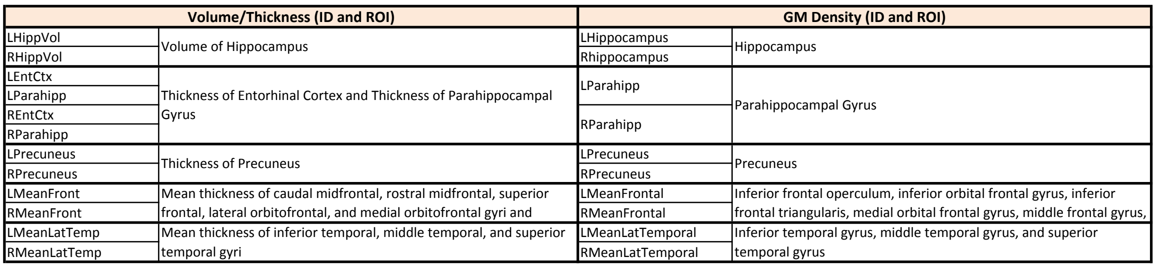
Both SNP and structural magnetic resonance imaging (MRI) data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). One goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. Following a prior study [16], 733 non-Hispanic Caucasian participants were included in this study.

4.2.1 SNP genotyping and group information

The SNP data [32], used in this study, were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA). Among all SNPs, only SNPs, belonging to the top 40 AD candidate genes listed on the AlzGene database (www.alzgene.org) as of June 10, 2010, were selected after the standard quality control (QC) and imputation steps. The QC criteria for the SNP data include (1) call rate check per subject and per SNP marker, (2) gender check, (3) sibling pair identification, (4) the Hardy-Weinberg equilibrium test, (5) marker removal by the minor allele frequency and (6) population stratification. As the second pre-processing step, the quality-controlled SNPs were imputed using the MaCH software [33] to estimate the missing genotypes. After that, the Illumina annotation information based on the Genome build 36.2 was used to select a subset of SNPs, belonging to the top 40 AD candidate genes [34].

The above procedure yielded 1224 SNPs from 37 genes. For the remaining 3 genes, no SNPs were available on the genotyping chip. The genes and the number of their SNPs are shown in Fig. 4.1. The ranking of the AlzGene database is based

Table 4.1. Quantitative traits (QTs) from "matching" ROIs: the volumetric/thickness measures (FreeSurfer) and GM density measures (VBM).



Volume/Thickness (ID and ROI)		GM Density (ID and ROI)	
LHippVol	Volume of Hippocampus	LHippocampus	Hippocampus
RHippVol		Rhippocampus	
LEntCtx	Thickness of Entorhinal Cortex and Thickness of Parahippocampal Gyrus	LParahipp	Parahippocampal Gyrus
LParahipp			
REntCtx		RParahipp	
RParahipp			
LPrecuneus	Thickness of Precuneus	LPrecuneus	Precuneus
RPrecuneus		RPrecuneus	
LMeanFront	Mean thickness of caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, and medial orbitofrontal gyri and	LMeanFrontal	Inferior frontal operculum, inferior orbital frontal gyrus, inferior frontal triangularis, medial orbital frontal gyrus, middle frontal gyrus,
RMeanFront		RMeanFrontal	
LMeanLatTemp	Mean thickness of inferior temporal, middle temporal, and superior temporal gyri	LMeanLatTemporal	Inferior temporal gyrus, middle temporal gyrus, and superior temporal gyrus
RMeanLatTemp		RMeanLatTemporal	

on SNPs instead of genes. As a result, most of the SNPs from these genes (Fig. 4.1) might be irrelevant to AD, while a small fraction of them could be risk factors for the disease and be associated with our intermediate imaging traits. Our task is to identify the SNPs in these 37 genes that predict important imaging QTs.

A straightforward observation from Fig. 4.1 shows that the SNPs are naturally divided into groups upon their belonging genes. This grouping structure of SNPs, though conveying important biological information, is seldom utilized in previous association studies that consider every SNP equally and investigate their genetic effects on imaging phenotypes separately. In this work, as one of the contributions, we aim to make use of the grouping information of SNPs in our learning model so as to achieve more lucid relationships between SNPs and neuroimaging phenotypes.

Besides grouping SNPs by genes, an alternative method could be based on LD. Through estimating non-random association of alleles at different loci (e.g., using pairwise correlation coefficients r^2, as shown in Fig. 4.2), the relationships between SNPs in terms of genetic linkage are established. For example, the group structure can be clearly observed in Fig. 4.2, where a group is defined as a block of SNPs whose pairwise $r^2 \geq 0.2$. As a result, we have 185 groups comprising 1029 SNPs, with each of the remaining 195 SNPs being isolated by itself.

In this study, we consider grouping SNPs by both genes and LD correlation coefficients r^2.

4.2.2 MRI analysis and extraction of imaging genotypes

Two widely employed automated MRI analysis techniques were used to process and extract imaging genotypes across the brain from all baseline scans of ADNI participants as previously described [16]. First, voxel-based morphometry (VBM) [35] was performed to define global gray matter (GM) density maps and extract local GM density values for target regions. Second, automated parcellation via FreeSurfer V4 [17] was conducted to define volumetric and cortical thickness values for regions of interest (ROIs) and to extract total intracranial volume (ICV). Further information is available in [16]. While a complete investigation of all VBM and FreeSurfer measures is an interesting future direction, this study is focused on a subset of these measures to test the proposed methods. Ten VBM (GM density) measures and twelve FreeSurfer measures (thickness/volume), which are known to be related to AD, are selected as QTs for identifying QTLs. These QTs are extracted from roughly matching ROIs with VBM and FreeSurfer. Table 4.1 shows the description of these QTs and Fig. 4.3 maps some of these ROIs in the brain space. All these measures were adjusted for the baseline age, gender, education, handedness, and baseline ICV using the regression weights derived from the healthy control participants.

4.3 Methods

In this section, we first systematically develop our computational models to explore the associations between SNPs and imaging phenotypes. As illustrated in Fig. 4.4, our method mainly addresses the group structure of genetic markers and joint learning across all the imaging endophenotypes, such that the learned regression model has better prediction performance and the selected SNPs are more biological meaningful. After that, we provide a new efficient algorithm to solve the proposed new multi-task regression and feature selection objective, followed by the rigorous algorithm analysis to prove its correctness and convergence.

Throughout this chapter, we write matrices as boldface uppercase letters and vectors as boldface lowercase letters. Given a matrix $\mathbf{M} = (m_{ij})$, its i-th row and j-th column are denoted as $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. The Frobenius norm and $\ell_{2,1}$-norm (also called as $\ell_{1,2}$-norm) of a matrix are defined as $||\mathbf{M}||_F = \sqrt{\sum_i ||\mathbf{m}^i||_2^2}$ and $||\mathbf{M}||_{2,1} = \sum_i \sqrt{\sum_j m_{ij}^2} = \sum_i ||\mathbf{m}^i||_2$, respectively.

4.3.1 Sparse Multi-tAsk Regression and feaTure selection

To explore the associations between SNPs and continuous imaging phenotypes, the linear (least square) regression (LR) is a standard approach. To avoid over-fitting and increase numerical stability, the ridge regression (RR) is a better option. Given the SNP data of the ADNI participants $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subseteq \Re^d$ and the selected imaging phenotypes $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\} \subseteq \Re^c$, where n is the number of participants (sample size), d is the number of SNPs (feature dimensionality) and c is the number of imaging phenotypes (tasks), the RR is designed to solve:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T \mathbf{x}_i - \mathbf{y}_i||^2 + \gamma \sum_{i=1}^{d} ||\mathbf{w}^i||^2, \tag{4.1}$$

where the entry w_{ij} of the weight matrix $\mathbf{W}$ measures the relative importance of the i-th SNP in predicting the response of the j-th imaging phenotype, and $\gamma > 0$ is a tradeoff parameter.

However, the RR model in Eq. (4.1) suffers from a number of problems when applied to evaluation of the imaging genetic associations. First, the weight matrix $\mathbf{W}$ is not sparse, therefore all the SNPs are involved in the prediction of imaging

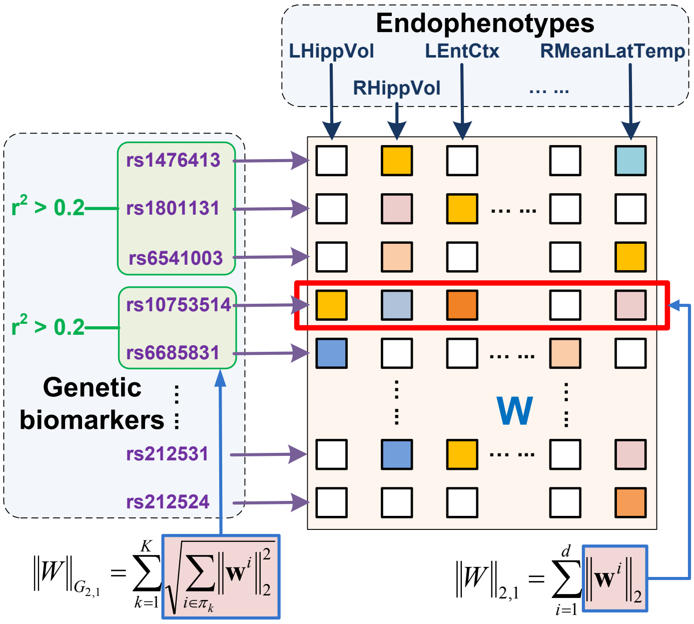


Figure 4.4. Illustration of the proposed SMART method. We incorporate the group structural information of the genetic markers through a new group $\ell_{2,1}$-norm regularization ($||\mathbf{W}||_{G_{2,1}}$), and enforce $\ell_{2,1}$-norm regularization ($||\mathbf{W}||_{2,1}$) to jointly select prominent SNPs across all endophenotypes..

phenotype responses. However, among numerous SNPs, only a small fraction of them are relevant to specific imaging QTs. Thus, it is desirable to select only relevant SNPs for more accurate prediction. Second, similar to LR, the tasks in the RR regression model are decoupled and each of them can be learned separately. As a result, the information of underlying interacting relationships between the brain regions are ignored, which, though, are essential to brain functionalities. Finally, the rows of **W** are equally treated in the RR model, which implies that the underlying structures among these SNPs are overlooked. However, it is generally believed that many SNPs are genetically linked. In order to tackle these difficulties, we propose a novel Sparse Multi-tAsk Regression and feaTure selection (SMART) method to exploit the interrelated structures within and between the genotypes and phenotypes.

4.3.1.1 Group-sparsity for genetic association

The objective of RR model in Eq. (4.1) uses Frobenious norm for regularization, which penalizes all the coefficients in a flat manner thereby all the SNPs are evenly treated. However, SNPs on the same chromosome with close distance tend to be inherited together and correlated with each other. For example, as shown in Fig. 4.4, the pairwise LD correlation coefficients r^2 between "rs1476413", "rs1801131" and "rs6541003" are greater than 0.2, thus they are more homogeneous and should be considered together when we predict the responses of the imaging QTs. Motivated by sparse learning, such as Lasso [11] and group Lasso [31], we propose a new form of regularization as follows:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T \mathbf{x}_i - \mathbf{y}_i||_2^2 + \gamma \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2}, \tag{4.2}$$

where the SNPs, *i.e.*, features, are partitioned into K groups $\Pi = \{\pi_k\}_{k=1}^K$, such that $\{\mathbf{w}^i\}_{i=1}^{m_k} \in \pi_k$ are genetically linked, and m_k is the number of SNPs in π_k. Two types of genetic links are used here to group SNPs: (1) SNPs are naturally divided into groups based on their belonging or nearest genes. (2) SNPs are grouped by thresholding the pairwise LD correlation coefficients r^2, e.g., in this work, the neighboring SNPs whose $r^2 \geq 0.2$ form a group.

Without loss of generality, $\{\pi_k\}_{k=1}^K$ are ordered and concatenated. Denote $\mathbf{W} = \begin{bmatrix} \mathbf{W}^1 \\ \cdots \\ \mathbf{W}^K \end{bmatrix}$, where $\mathbf{W}^k \in \mathbb{R}^{m_k \times c} (1 \leq k \leq K)$, we can write Eq. (4.2) as following:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T \mathbf{x}_i - \mathbf{y}_i||_2^2 + \gamma \sum_{k=1}^{K} ||\mathbf{W}^k||_F, \tag{4.3}$$

which can be written in matrix form as following:

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma||\mathbf{W}||_{G_{2,1}}, \tag{4.4}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$, and $||\cdot||_{G_{2,1}}$ is our proposed *group $\ell_{2,1}$-norm ($G_{2,1}$-norm)* of a matrix with respect to a partition Π and defined as:

$$||\mathbf{W}||_{G_{2,1}} = \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} = \sum_{k=1}^{K} ||\mathbf{W}^k||_F \ . \tag{4.5}$$

Note that: the $G_{2,1}$-norm defined above is different from the regularization term in group Lasso. Given a partition of the features, the group Lasso enforces group-wise sparsity for each learning task separately, whereas the $G_{2,1}$-norm defined in Eq. (4.5) penalizes the regression coefficients of a group of features across all the learning tasks jointly. As a result, the biological group-level structural information among SNPs are incorporated into our multi-task learning model.

Moreover, because the ℓ_1-norm across all the group-wise penalties are used in $G_{2,1}$-norm, similar to Lasso and group Lasso, sparsity is enforced among biological groups. This is important in identifying relevant genotypes for specific phenotypes, because only a small fraction of SNPs are related to certain imaging phenotypes. From the perspective of sparsity learning, the Lasso and group Lasso have flat sparsity, the $\ell_{2,1}$-norm has structured sparsity, and the $G_{2,1}$-norm has structured sparsity across feature groups.

4.3.1.2 Individual structured sparsity for joint feature selection

Although the objective in Eq. (4.4) takes into account the group structure of the SNP data through the proposed $G_{2,1}$-norm, the feature selection across tasks are still not completely addressed, because $G_{2,1}$-norm penalizes the coefficients flatly within each group of SNPs. To be more specific, within a given group, say π_k, Frobenious norm $||\mathbf{W}^k||_F$ is used, which is the same as ridge regression that uses Frobenious norm over the whole projection matrix $\mathbf{W}$. In an important group, certain features could be irrelevant; on the other hand, in a less important group, some features could be significant to tasks. Thus, we enforce additional structured sparsity to our learning model for jointly selecting features across multiple tasks via a $\ell_{2,1}$-norm regularization [12, 13, 36, 37]:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T \mathbf{x}_i - \mathbf{y}_i||_2^2 + \gamma_1 \sum_{k=1}^{K} ||\mathbf{W}^k||_F + \gamma_2 \sum_{i=1}^{d} ||\mathbf{w}^i||_2, \tag{4.6}$$

which can be concisely rewritten in matrix form as:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T \mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 ||\mathbf{W}||_{G_{2,1}} + \gamma_2 ||\mathbf{W}||_{2,1} \ . \tag{4.7}$$

In Eq. (4.7), the first term measures the regression loss. The second term couples all the regression coefficients of a group of features over all the c tasks together, which incorporates the grouping information on features (SNPs) due to the genetic linkage. Finally, the third term penalizes all c regression coefficient of each individual feature as whole to select features across multiple learning tasks.

We call Eq. (4.7) as Sparse Multi-tAsk Regression and feaTure selection (SMART) method with illustration in Fig. 4.4.

4.3.2 A new efficient optimization algorithm

Because the number of genetic markers can be very large, we need an efficient algorithm to solve Eq. (4.7). Existing algorithms usually reformulate such sparsity problem as a second order cone programming (SOCP) or semidefinite programming (SDP) problem, which can be solved by interior point method or the bundle method. However, solving SOCP or SDP is computationally very expensive, which limits their use in practice. Here, we propose an efficient algorithm to solve our objective function in Eq. (4.7).

Taking the derivative with respect to $\mathbf{W}$, and setting the derivative to zero, we have[1]

$$\mathbf{X}\mathbf{X}^T\mathbf{W} - \mathbf{X}\mathbf{Y}^T + \gamma_1\mathbf{D}\mathbf{W} + \gamma_2\tilde{\mathbf{D}}\mathbf{W} = \mathbf{0}, \tag{4.8}$$

where $\mathbf{D}$ is a block diagonal matrix with the k-th diagonal block as $\frac{1}{2\|\mathbf{W}^k\|_F}\mathbf{I}_k$, $\mathbf{I}_k$ is an identity matrix with size of m_k, $\tilde{\mathbf{D}}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{w}^i\|_2}$. Thus we have

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D} + \gamma_2\tilde{\mathbf{D}})^{-1}\mathbf{X}\mathbf{Y}^T, \tag{4.9}$$

where $\mathbf{W}$ can be efficiently obtained by solving the linear equation $(\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D} + \gamma_2\tilde{\mathbf{D}})\mathbf{W} = \mathbf{X}\mathbf{Y}^T$, and the matrix inversion that is computationally expensive is not involved.

Note that $\mathbf{D}$ and $\tilde{\mathbf{D}}$ in Eq. (4.9) depend on $\mathbf{W}$ and thus are also unknown variables. We propose an iterative algorithm to solve this problem, which is described in Algorithm 3.

4.3.3 Analysis of the algorithm

Now we prove that Algorithm 3 converges to the global optimum.

Lemma 2 *For any matrices $\mathbf{M}$ and $\mathbf{M}_0$ with the same size, we have* $||\mathbf{M}||_F - \frac{||\mathbf{M}||_F^2}{2||\mathbf{M}_0||_F} \leq ||\mathbf{M}_0||_F - \frac{||\mathbf{M}_0||_F^2}{2||\mathbf{M}_0||_F}$.

[1] When $\|\mathbf{W}^k\|_F = 0$, the k-th diagonal block of $\mathbf{D}$ can be regularized as $\frac{1}{2\sqrt{\|\mathbf{W}^k\|_F^2+\varsigma}}\mathbf{I}_k$. Similarly, when $\mathbf{w}^i = 0$, the i-th diagonal element of $\tilde{\mathbf{D}}$ can be regularized as $\frac{1}{2\sqrt{\|\mathbf{w}^i\|_2^2+\varsigma}}$. Then the derived algorithm can be proved to minimize $\sum_{i=1}^n ||\mathbf{W}^T\mathbf{x}_i - \mathbf{y}_i||_2^2 + \gamma_1 \sum_{k=1}^K \sqrt{\|\mathbf{W}^k\|_F^2 + \varsigma} + \gamma_2 \sum_{i=1}^d \sqrt{\|\mathbf{w}^i\|_2^2 + \varsigma}$. It is easy to see that this problem is reduced to problem (4.6) when $\varsigma \to 0$.

```
Algorithm 3: Algorithm to solve Eq. (4.7).
Input: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n] \in \mathbb{R}^{c\times n}$
Initialize $\mathbf{W}_1 \in \mathbb{R}^{d\times c}$, $t = 1$ ;
while not converge do
    1. Calculate the block diagonal matrix $\mathbf{D}_t$, where the $k$-th diagonal is
       $\frac{1}{2\|\mathbf{W}_t^k\|_F}\mathbf{I}_k$; Calculate the diagonal matrix $\tilde{\mathbf{D}}_t$, where the $i$-th diagonal element
       is $\frac{1}{2\|\mathbf{w}_t^i\|_2}$ ;
    2. $\mathbf{W}_{t+1} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D}_t + \gamma_2\tilde{\mathbf{D}}_t)^{-1}\mathbf{X}\mathbf{Y}^T$ ;
    3. $t = t + 1$ ;
Output: $\mathbf{W}_t \in \mathbb{R}^{d\times c}$.
```

Proof: Obviously, $-(||\mathbf{M}||_F - ||\mathbf{M}_0||_F)^2 \leq \mathbf{M}$, so we have

$$\begin{aligned}
&-(||\mathbf{M}||_F - ||\mathbf{M}_0||_F)^2 \leq \mathbf{M} \\
&\Rightarrow 2||\mathbf{M}||_F||\mathbf{M}_0||_F - ||\mathbf{M}||_F^2 \leq ||\mathbf{M}_0||_F^2 \\
&\Rightarrow ||\mathbf{M}||_F - \tfrac{||\mathbf{M}||_F^2}{2||\mathbf{M}_0||_F} \leq ||\mathbf{M}_0||_F - \tfrac{||\mathbf{M}_0||_F^2}{2||\mathbf{M}_0||_F}
\end{aligned}$$

which completes the proof. □

Theorem 3 *Algorithm 3 decreases the objective value in each iteration.*

Proof: In each iteration t, according to step 2 we have

$$\begin{aligned}
&||\mathbf{W}_{t+1}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 Tr\mathbf{W}_{t+1}^T\mathbf{D}_t\mathbf{W}_{t+1} + \gamma_2 Tr\mathbf{W}_{t+1}^T\tilde{\mathbf{D}}_t\mathbf{W}_{t+1} \\
&\leq ||\mathbf{W}_t^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 Tr\mathbf{W}_t^T\mathbf{D}_t\mathbf{W}_t + \gamma_2 Tr\mathbf{W}_t^T\tilde{\mathbf{D}}_t\mathbf{W}_t \\
&\Rightarrow ||\mathbf{W}_{t+1}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1\sum_{k=1}^{K}\frac{||\mathbf{W}_{t+1}^k||_F^2}{2||\mathbf{W}_t^k||_F} + \gamma_2\sum_{i=1}^{d}\frac{||\mathbf{w}_{t+1}^i||_2^2}{2||\mathbf{w}_t^i||_2} \\
&\leq ||\mathbf{W}_t^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1\sum_{k=1}^{K}\frac{||\mathbf{W}_t^k||_F^2}{2||\mathbf{W}_t^k||_F} + \gamma_2\sum_{i=1}^{d}\frac{||\mathbf{w}_t^i||_2^2}{2||\mathbf{w}_t^i||_2}. \qquad (4.10)
\end{aligned}$$

Applying Lemma 3 twice to Eq. (4.10), we have the following

$$\begin{aligned}
&||\mathbf{W}_{t+1}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 \sum_{k=1}^{K} ||\mathbf{W}_{t+1}^k||_F + \gamma_2 \sum_{i=1}^{d} ||\mathbf{w}_{t+1}^i||_2 \\
&\leq ||\mathbf{W}_t^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 \sum_{k=1}^{K} ||\mathbf{W}_t^k||_F + \gamma_2 \sum_{i=1}^{d} ||\mathbf{w}_t^i||_2. \qquad (4.11)
\end{aligned}$$

Thus, Algorithm 3 decreases the objective value in each iteration. □

Algorithm 3 stops when the following criterion is satisfied:

$$||\mathbf{W}_{t+1} - \mathbf{W}_t||_F \,/\, \max\left(||\mathbf{W}_t||_F, 1\right) \leq \mathrm{Tol}, \qquad (4.12)$$

where $\mathrm{Tol} = 10^{-4}$ is empirically selected in our experiments.

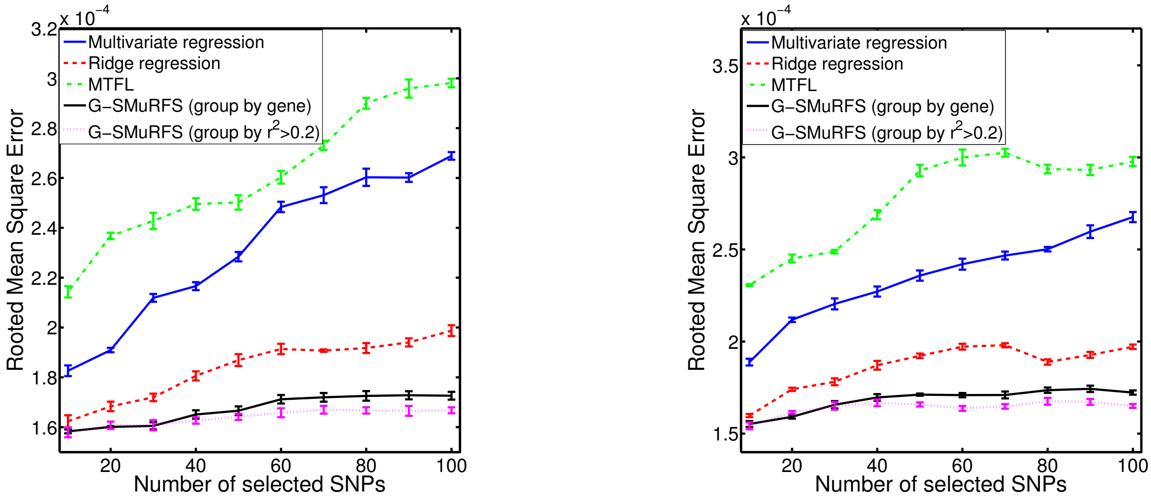
Upon convergence, $\mathbf{W}_t$, $\mathbf{D}_t$ and $\tilde{\mathbf{D}}_t$ will satisfy Eq. (4.9). As the problem of solving Eq. (4.7) is a convex problem, satisfying the Eq. (4.9) indicates that $\mathbf{W}_t$ is a global optimum solution to Eq. (4.7). Therefore, Algorithm 3 converges to the global optimum of Eq. (4.7). Since we have a closed form solution in each iteration, our algorithm converges very fast, which makes our method is suitable for not only candidate SNP but also genome-wide association studies.

4.4 Experimental Results and Discussions

In this section, we evaluate the proposed SMART method by applying it to the data from the ADNI cohort, where a wide range of SNPs are examined and selected to predict the response of the MRI imaging phenotypes. The goal is to select a compact set of SNPs while maintaining high predictive power.

4.4.1 Improved imaging phenotype prediction

We first evaluate the proposed method in predicting the continuous responses of candidate neuroimaging phenotypes. Given two sets of imaging phenotypes, FreeSurfer and VBM, we conduct experiments on each of them separately.



(a) FreeSurfer imaging genotypes.

(b) VBM imaging genotypes.

Figure 4.5. Performance comparison: The mean and standard deviation (SD) of the root mean square errors (RMSEs) obtained from 5 cross-validation trials in each experiment are plotted, where each error bar indicates ± 1 SD..

We compare our method against multivariate linear regression, ridge regression, and multi-task feature learning (MTFL) [13] method. The former two are the most widely used methods in statistical learning and medical image analysis. The latter one is a method most related to the proposed method in that it also selects features (SNPs) across tasks, however it only uses $\ell_{2,1}$-norm regularization whereby group information is not taken into account. Therefore, MTFL method can be seen as a special case of the proposed method by setting $\gamma_1 = 0$ in Eq. (4.7).

We group SNPs using two methods: (1) SNPs annotated with the same gene are grouped together; (2) SNPs within the same LD block are grouped together, where $r^2 \geq 0.2$ is used in this work. For each test case, we conduct standard 5-fold cross-validation and report the average results. For each of the 5 trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$ for ridge regression, MTFL method, and our method. For each trial, from the learned coefficient matrix we sum the absolute

Table 4.2. The results (p-values) of t-tests for performance comparison between our methods and three competing methods.

	FreeSurfer biomarkers		VBM biomarkers	
	Group by gene	Group by $r^2 > 0.2$	Group by gene	Group by $r^2 > 0.2$
MLR	7.08×10^{-5}	3.13×10^{-5}	9.28×10^{-6}	3.96×10^{-6}
RR	1.31×10^{-2}	2.21×10^{-3}	2.12×10^{-2}	1.85×10^{-3}
MTFL	6.57×10^{-7}	2.41×10^{-7}	5.82×10^{-7}	2.63×10^{-7}

values of the coefficients of a single SNP over all the tasks as the SNP weight, from which we pick up the top $\{10, 20, \ldots, 100\}$ SNPs to predict the regression responses for the test data. The performance of each trial is assessed by root mean square error (RMSE), a widely used measurement for regression analysis. For each experiment, the mean and standard deviation (SD) of the RMSEs obtained from the 5 trials are reported in Fig. 4.5, where each error bar indicates ± 1 SD. Detailed RMSE results of each fold in cross-validation are available in the supplementary document at http://ranger.uta.edu/%7eheng/imaging-genetics/.

The proposed SMART methods consistently outperform three competing methods in both FreeSurfer and VBM cases (Fig. 4.5), while the cross-validation trials in each experiment perform very similarly to one another (see the error bars in Fig. 4.5). For a formal comparison, t-test is performed and the resulting p-values are reported in Table 4.2, from which we can see that our methods are significantly better than three completing methods. Moreover, the predictive performances of our methods are considerably stable, whereas those of the other methods are sensitive to experimental conditions. These results clearly demonstrate the advantage of the proposed SMART method in predicting phenotypic responses.

A more careful observation shows that the regression performance of our method when using $r^2 > 0.2$ to group SNPs is better than that of our method when grouping SNPs by genes. While gene is the most natural way to group SNPs, different segments within the same gene may have different functions (e.g., bases for different isoforms) and mixing them together may perturb the prediction. Grouping by LD blocks using r^2 yields more homogeneous groups and has a potential to boost the prediction power.

Fig. 4.6 shows heat maps of prediction errors on each QT. While all these QTs are AD-relevant, Fig. 4.6 indicates that they are affected in different degrees by genetic factors. QTs that are better predicted by SNPs include GM density measures of the parahippocampal gyrus and frontal region in VBM analyses and thickness measures of the frontal region, lateral temporal region and precuneus in FreeSurfer analyses. The VBM and FreeSurfer measures of a certain region yield similar results in some cases (e.g., frontal region), but may provide different information in other cases (e.g., parahippocampal gyrus). Thus, performing both VBM and FreeSurfer analyses can help identify useful imaging phenotypes and guide further investigation to better elucidate the underlying disease mechanism, from gene, to brain structure and function, and to symptoms.

4.4.2 Genetic marker selection

Shown in Fig. 4.7 are the regression coefficients for top 10 selected SNPs. First, these SNPs are either AlzGene candidates or proximal to the candidates; however, little is known about their underlying mechanisms in relation to AD. The results shown in Fig. 4.7 can help identify relevant QTs for each SNP and has a potential to gain biological insights from gene to brain to symptoms. Second, as expected, the *APOE* SNP rs429358 shows the strongest association with all QTs in each experiment; and the hippocampal measures exhibit the strongest association with the *APOE* SNP.

Clearly, the proposed approach is able to identify the most important AD genetic risk factor via imaging QTs as well as the best-known neurodegenerative marker. Third, besides confirming the prior findings, our method also yielded new discoveries such as the associations between *APOE* and other eminent AD markers including entorhinal cortex and parahippocampal gyrus. These associations were not identified in our prior massive univariate analyses on the same data (Shen et al., 2010), indicating that the proposed multi-locus method has increased power to discover interesting imaging QTs. In sum, the above evidence demonstrates not only the effectiveness of the proposed method but also the strength of using imaging QTs in genetic association study.

Quite a few SNPs from the *SORCS1* gene are selected as the top 10 hits in each experiment, however the large size of the gene (Fig. 4.1) may play a role. Fig. 4.8 shows an LD plot with location maps for a group of 46 *SORCS1* SNPs, where two top hits (red spikes) are highlighted for each of the FreeSurfer and VBM experiments. Although *SORCS1* has been associated with diabetes and AD, the top ranked *SORCS1* SNPs in Fig. 4.7 have not been reported in prior association studies. Thus, this gene together with its SNPs warrants further investigation in independent cohorts. Due to the nature of our method, an epistasis analysis on these top hits would be appropriate for investigation in future studies.

4.5 Conclusions

In this chapter, we have proposed a novel Sparse Multi-tAsk Regression and feaTure selection (SMART) method to perform both regression analysis for predicting continuous responses of brain imaging measures and selecting relevant SNPs in an MCI/AD study. Different from traditional regression methods that ignore the interrelated structures within genotyping and imaging data, our method studies the

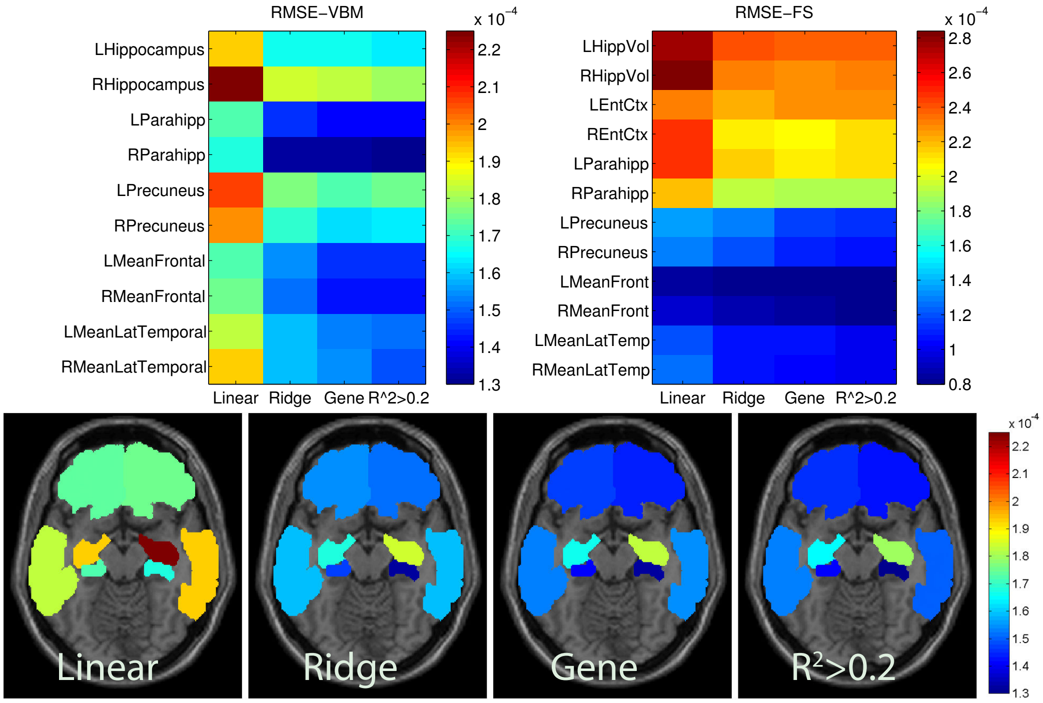


Figure 4.6. Top panels show the heat maps of root mean square errors (RMSEs) for predicting VBM (top left) and FreeSurfer (top right) measures using linear regression, ridge regression, our G-SMuRFS method with SNPs grouped by gene, and G-SMuRFS with SNPs grouped by $r^2 > 0.2$, where top 10 SNPs were used in our G-SMuRFS methods. In the bottom panel, RMSEs for predicting VBM measures using four methods are mapped onto the brain volume. .

associations between SNPs and imaging phenotypes within a single regression framework and shared common subspace. Through enforcing a new form of regularization using $G_{2,1}$-norm that takes into account both group-level structural information inside SNP data and sparsity among SNP groups, our learning model is able to exploit additional information to achieve both enhanced predictive performance and improved feature (SNP) selection capability. Besides, $\ell_{2,1}$-norm regularization is used in our model to jointly select SNPs relevant to important imaging phenotypes. An efficient algorithm to solve the proposed objective is presented with rigorous proof of its correctness and convergence. Our experiments using the SNP and MRI data from the

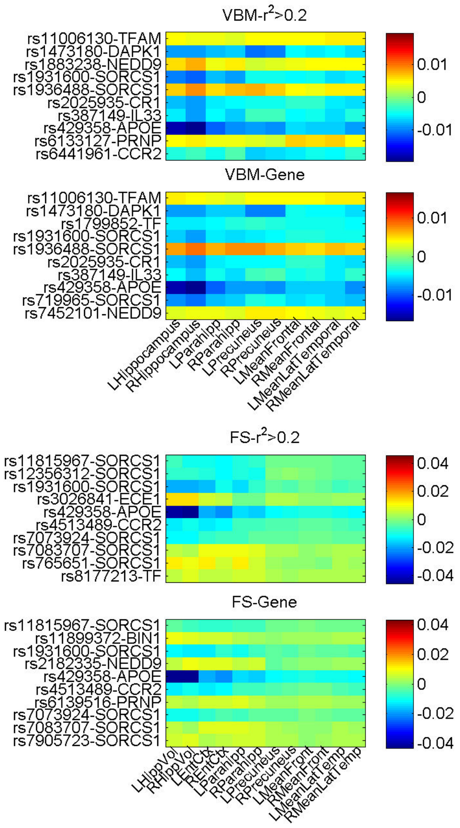


Figure 4.7. Regression coefficients are visualized for top 10 selected SNPs in each of the four experiments (from top to bottom): (1) Group by $r^2 > 0.2$, regression on VBM measures, (2) group by gene, regression on VBM measures, (3) group by $r^2 > 0.2$, regression on FreeSurfer measures, and (4) group by gene, regression on FreeSurfer measures. .

ADNI cohort yielded the following promising results: (1) the prediction performance of SMART method was consistently better than conventional multi-variate linear regression and ridge regression, (2) a compact set of SNP predictors were identified in each test case, warranting further investigation in independent cohorts for confirmation, and (3) these selected SNPs could predict the responses of multiple imaging phenotypes at the same time and had a potential to serve as useful genetic risk factors for AD. These promising results were consistent with our theoretical foundation and

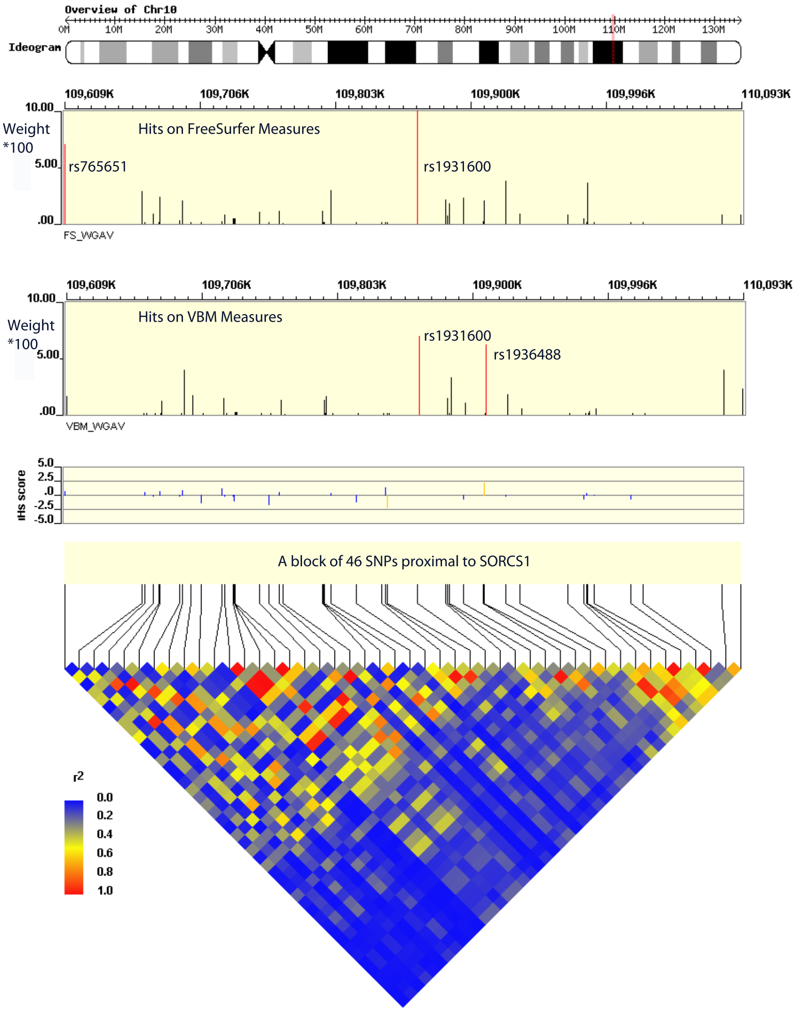


Figure 4.8. Pair-wise Linkage Disequilibrium (LD) in a group of 46 SNPs proximal to SORCS1. Numerical values r^2 of the LD maps are determined by Haploview and visualized with WGAViewer. The top panel is the ideogram of the chromosome and the vertical red line represents the relative location of the locus of interest. In the second panel, regression coefficients*100 is plotted for each SNP for the FreeSurfer data, where two top hits rs765651 and rs1931600 are labeled with red lines. In the third panel, regression coefficients*100 is plotted for each SNP for the VBM data, where two top hits rs1931600 and rs1936488 are labeled with red lines. The fourth panel shows the recent selection score [1]. The bottom figure demonstrates the LD pattern among 46 SNPs. .

in accordance with some prior studies, which demonstrated the effectiveness of the proposed method.

One important future direction of this work could be to explore the possibility of simultaneously employing multiple SNP grouping schemes or more generally adopting a pre-defined network/pathway strategy and see whether these approaches can further improve the prediction performance. Other potential future directions include (1) application of SMART method to additional imaging phenotypes (*e.g.*, PET, fMRI data), and (2) building a principled sparse learning framework to reveal complex relationships among multiple data sources available in the ADNI database, including genetic, cerebrospinal fluid (CSF), plasma, imaging, and cognitive data sets to study AD at a system biology level.

CHAPTER 5

IDENTIFYING DISEASE SENSITIVE AND QUANTITATIVE TRAIT RELEVANT BIOMARKERS FROM MULTI-DIMENSIONAL HETEROGENEOUS IMAGING GENETICS DATA VIA SPARSE MULTI-MODAL MULTI-TASK LEARNING

5.1 Introduction

Recent advances in acquiring multi-modal brain imaging and genome-wide array data provide exciting new opportunities to study the influence of genetic variation on brain structure and function. Research in this emerging field, known as *imaging genetics*, holds great promise for a system biology of the brain to better understand complex neurobiological systems, from genetic determinants to cellular processes to the complex interplay of brain structure, function, behavior and cognition. Analysis of these multi-modal data sets will facilitate early diagnosis, deepen mechanistic understanding and improved treatment of brain disorders.

Machine learning methods have been widely employed to predict Alzheimer's disease (AD) status using imaging genetics measures [5–7, 27]. Since AD is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions, regression models have also been investigated to predict clinical scores from structural, such as magnetic resonance imaging (MRI), and/or molecular, such as fluorodeoxyglucose positron emission tomography (FDG-PET), neuroimaging data [8, 9]. For example, [8] performed stepwise regression in a pairwise fashion to relate each of MRI and FDG-PET measures of eight candidate regions to each of four Rey's Auditory Verbal Learning Test (RAVLT) memory scores. This univariate ap-

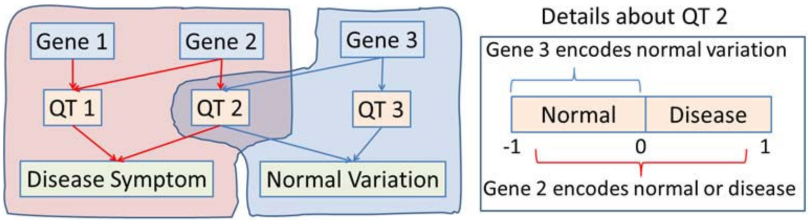


Figure 5.1. A simplified schematic example of two pathways from gene to QTs to phenotypic endpoints: the red one is disease relevant while the blue one yields only normal variation. Traditional two-stage imaging genetic strategy identifies QT 1 and QT 2 first and then Genes 1, 2, 3. Our new method will identify only disease relevant genes (*i.e.*, Gene 1 and Gene 2); and Gene 3 won't be identified because it cannot be used to classify disease status..

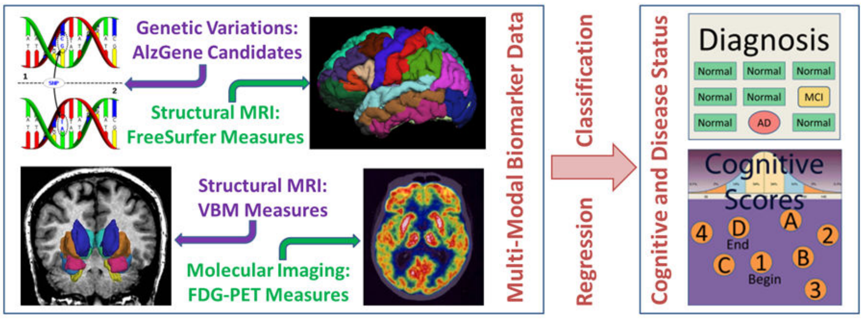


Figure 5.2. The proposed sparse multi-modal multi-task feature selection method will identify biomarkers from multi-modal heterogeneous data resources. The identified biomarkers could predict not only disease status, but also cognitive functions to help researchers better understand the underlying mechanism from gene to brain structure and function, and to cognition and disease..

proach, however, did not consider either interrelated structures within imaging data or those within cognitive data. Using relevance vector regression, [9] jointly analyzed the voxel-based morphometry (VBM) features extracted from the entire brain to predict each selected clinical score, while the investigations of different clinical scores are independent from each other.

One goal of imaging genetics is to identify genetic risk factors and/or imaging biomarkers via intermediate quantitative traits (QTs, *e.g.* cognitive memory scores used in this chapter) on the chain from gene to brain to symptom. Thus, both disease classification and QT prediction are important machine learning tasks. Prior imaging genetics research typically employs a two-step procedure for identifying risk factors and biomarkers: one first determines disease-relevant QTs, and then detects the biomarkers associated with these QTs. Since a QT could be related to many genetic or imaging markers on different pathways that are not all disease specific (*e.g.*, QT 2 and Gene 3 in Fig. 5.1), an ideal scenario would be to discover only those markers associated with both QT and disease status for a better understanding of the underlying biological pathway specific to the disease.

On the other hand, identifying genetic and phenotypic biomarkers from large-scale multi-dimensional heterogeneous data is an important biomedical and biological research topic. Unlike simple feature selection working on a single data source, multi-modal learning describes the setting of learning from data where observations are represented by multiple types of feature sets. Many multi-modal methods have been developed for classification and clustering purposes, such as co-training [38–41] and multi-view clustering [42, 43]. However, they typically assume that the multi-modal feature sets are conditionally independent, which does not hold in many real-world applications such as imaging genetics. Considering different representations give rise to different kernel functions, several Multiple Kernel Learning (MKL) approaches [44–53] have been recently studied and employed to integrate heterogeneous data and select multi-type features. However, such models train a single weight for all features from the same modality, *i.e.,* all features from the same data source are weighted equally, when they are combined with the features from other sources. This limitation often yields inadequate performance.

To address the above challenges, we propose a new sparse multi-modal multi-task learning algorithm that integrates heterogeneous genetic and phenotypic data effectively and efficiently to identify disease-sensitive and cognition-relevant biomarkers from multiple data sources. Different to LASSO [11], group LASSO [31] and other related methods that mainly find the biomarkers correlated to each individual QT (memory score), we consider predicting each memory score as a regression task and select biomarkers that tend to play an important role in influencing multiple tasks. A joint classification and regression multi-task learning model is utilized to select the biomarkers correlated to memory scores and disease categories simultaneously.

Sparsity regularizations have recently been widely investigated and applied to multi-task learning models [12, 13, 54–57]. Sparse representations are typically achieved by imposing non-smooth norms as regularizers in the optimization problems. From the view of sparsity organization, we have two types: 1) The flat sparsity is often achieved by ℓ_0-norm or ℓ_1-norm regularizer or trace norm in matrix/tensor completion. Optimization techniques include LARS [58], linear gradient search [59], proximal methods [60]. 2) The structured sparsity is usually obtained through different sparse regularizers such as $\ell_{2,1}$-norm [55–57], $\ell_{2,0}$-norm [61], $\ell_{\infty,1}$-norm [62], (also denoted as $\ell_{1,2}$-norm, $\ell_{1,\infty}$-norm in different papers) and group ℓ_1-norm [31] which can be solved by methods in [54, 63]. We propose a new combined structured sparse regularization to integrate features from different modalities and to learn a weight for each feature leading to a more flexible scheme for feature selection in data integration, which is illustrated in Fig. 5.3. In our combined structured sparse regularization, the group ℓ_1-norm regularization (blue circles in Fig. 5.3) learns the feature global importance, *i.e.* the modal-wise feature importance of every data modality on each class (task), and the $\ell_{2,1}$-norm regularization (red circles in Fig. 5.3) explores the feature local importance, *i.e.* the importance of each feature for multiple classes/tasks. The proposed method is applied to identify AD-sensitive biomarkers associated to the cognitive scores by integrating heterogeneous genetic and phenotypic data (as shown in Fig. 5.2). Our empirical results yield clearly improved performance on predicting both cognitive scores and disease status.

5.2 Identifying Disease Sensitive and QT Relevant Biomarkers from Heterogeneous Imaging Genetics Data

Pairwise univariate correlation analysis can quickly provide important association information between genetic/phenotypic data and QTs. However, it treats the features and the QTs as independent and isolated units, therefore the underlying interacting relationships between the units might be lost. We propose a new sparse multi-modal multi-task learning model to reveal genetic and phenotypic biomarkers, which are disease sensitive and QT-relevant, by simultaneously and systematically taking into account an ensemble of SNPs (Single-nucleotide polymorphism) and phenotypic signatures and jointly performing two heterogeneous tasks, *i.e.* biomarker-to-QT regression and biomarker-to-disease classification. The QTs studied in this chapter are the cognitive scores.

In multi-task learning, given a set of input variables (*i.e.*, features such as SNPs and MRI/PET measures), we are interested in learning a set of related models (*e.g.* relations between genetic/imaging markers and cognitive scores) to predict multiple outcomes (*i.e.*, tasks such as predicting cognitive scores and disease status). Because these tasks are relevant, they share a common input space. As a result, it is desirable to learn all the models jointly rather than treating each task as independent and fitting each model separately, such as Lasso [11] and group Lasso [31]. Such multi-task learning can discover robust patterns (because significant patterns in a single task could be outliers for other tasks) and potentially increase the predictive power.

In this chapter, we write matrices as uppercase letters and vectors as boldface lowercase letters. Given a matrix $W = [w_{ij}]$, its ith row and jth column are denoted

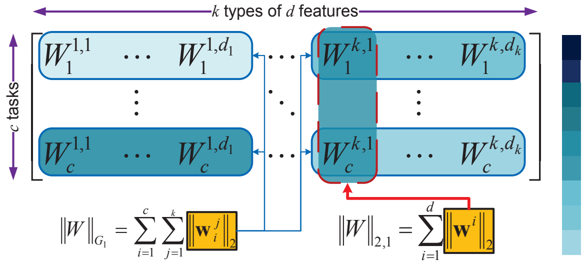

k types of d features
c tasks
$W_1^{1,1}$ … W_1^{1,d_1}
$W_c^{1,1}$ … W_c^{1,d_1}
$W_1^{k,1}$ … W_1^{k,d_k}
$W_c^{k,1}$ … W_c^{k,d_k}
$\|W\|_{G_1} = \sum_{i=1}^{c}\sum_{j=1}^{k}\|\mathbf{w}_i^j\|_2$
$\|W\|_{2,1} = \sum_{i=1}^{d}\|\mathbf{w}^i\|_2$

Figure 5.3. Illustration of the feature weight matrix W^T. The elements in matrix with deep blue color have large values. The group ℓ_1-norm (G_1-norm) emphasizes the learning of the ***group-wise*** weights for a type of features (*e.g.*, all the SNPs features, or all the MRI imaging features, or all the FDG-PET imaging features) corresponding to each task (*e.g.*, the prediction for a disease status or a memory score) and the $\ell_{2,1}$-norm accentuates the ***individual*** weight learning cross multiple tasks..

as $\mathbf{w}^i$ and $\mathbf{w}_j$, respectively. The $\ell_{2,1}$-norm of the matrix W is defined as $||W||_{2,1} = \sum_{i=1} ||\mathbf{w}^i||_2$ (also denoted as $\ell_{1,2}$-norm by other researchers).

5.2.1 Heterogeneous Data Integration via Combined Structured Sparse Regularizations

First, we will systematically propose our new multi-modal learning method to integrate and select the genetic and phenotypic biomarkers from large-scale heterogeneous data. In the supervised learning setting, we are given n training samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, where $\mathbf{x}_i = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^k)^T \in \Re^d$ is the input vector including all features from a total of k different modalities and each modality j has d_j features ($d = \sum_{j=1}^k d_j$). $\mathbf{y}_i \in \Re^c$ is the class label vector of data point $\mathbf{x}_i$ (only one element in $\mathbf{y}_i$ is 1, and others are zeros), where c is the number of classes (tasks). Let

$X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$ and $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_c] \in \Re^{c \times n}$. Different to MKL, we directly learn a $d \times c$ parameter matrix as:

$$W = \begin{bmatrix} \mathbf{w}_1^1 & ... & \mathbf{w}_c^1 \\ ... & ... & ... \\ \mathbf{w}_1^k & ... & \mathbf{w}_c^k \end{bmatrix} \in \Re^{d \times c}, \tag{5.1}$$

where $\mathbf{w}_p^q \in \Re^{d_q}$ indicates the weights of all features in the q-th modality with respect to the p-th task (class). Typically we can use a convex loss function $\mathcal{L}(X, W)$ to measure the loss incurred by W on the training samples. Compared to MKL approaches that learn one weight for one kernel matrix representing one modality, our method will learn the weight for each feature to capture the local feature importance. Because the features come from heterogeneous data sources, we impose the regularizer $\mathcal{R}(W)$ to capture the interrelationships of modalities and features as:

$$\min_W \ \mathcal{L}(X, W) + \gamma \mathcal{R}(W), \tag{5.2}$$

where γ is a trade-off parameter. In heterogeneous data fusion, from multi-view perspective of view, the features of a specific view (modality) can be more or less discriminative for different tasks (classes). Thus, we propose a new group ℓ_1-norm (G_1-norm) as a regularization term in Eq. (5.2), which is defined over W as following:

$$\|W\|_{G_1} = \sum_{i=1}^{c} \sum_{j=1}^{k} ||\mathbf{w}_i^j||_2, \tag{5.3}$$

which is illustrated by the blue circles in Fig. 5.3. Then the Eq. (5.2) becomes:

$$\min_W \ \mathcal{L}(X, W) + \gamma_1 \|W\|_{G_1}. \tag{5.4}$$

Because the group ℓ_1-norm uses ℓ_2-norm within each modality and ℓ_1-norm between modalities, it enforces the sparsity between different modalities, *i.e.* if one modality of features are not discriminative for certain tasks, the objective in Eq. (5.4) will assign zeros (in ideal case, usually they are very small values) to them for corresponding tasks; otherwise, their weights are large. This new group ℓ_1-norm regularizer captures the global relationships between data modalities.

However, in certain cases, even if most features in one modality are not discriminative for the classification or regression tasks, a small number of features in the same modality can still be highly discriminative. From the multi-task learning point of view, such important features should be shared by all/most tasks. Thus, we add an additional $\ell_{2,1}$-norm regularizer into Eq. (5.4) as:

$$\min_{W} \ \mathcal{L}(X, W) + \gamma_1 \left\| W \right\|_{G_1} + \gamma_2 \left\| W \right\|_{2,1} . \tag{5.5}$$

The $\ell_{2,1}$-norm was popularly used in multi-task feature selection [55, 63]. Because the $\ell_{2,1}$-norm regularizer impose the sparsity between all features and non-sparsity between tasks, the features that are discriminative for all tasks will get large weights.

Our regularization items consider the heterogeneous features from both group-wise and individual viewpoints. Fig. 5.3 visualizes the matrix W^T as a demonstration. In Fig. 5.3, the elements with deep blue color have large values. The group ℓ_1-norm emphasizes the group-wise weights learning corresponding to each task and the $\ell_{2,1}$-norm accentuates the individual weight learning cross multiple tasks. Through the combined regularizations, for each task (class), many features (not all of them) in the discriminative modalities and a small number of features (may not be none) in the non-discriminative modalities will learn large weights as the important and discriminative features.

The multi-dimensional data integration has been increasingly important to many biological and biomedical studies. So far, the MKL methods are most widely used. Due to the learning model deficiency, the MKL methods cannot explore both modality-wise importance and individual importance of features simultaneously. Our new structured sparse multi-modal learning method integrates the multi-dimensional data in a more efficient and effective way. The loss function $\mathcal{L}(X, W)$ in Eq. (5.8) can be replace by either least square loss function or logistic regression loss function to perform regression/classification tasks.

5.2.2 Joint Disease Classification and QT Regression

Because we are interested in identifying the disease sensitive and QT relevant biomarkers, we consider performing both logistic regression for classifying disease status and multivariate regression for predicting cognitive memory scores simultaneously [2]. A similar model was used in [64] for heterogeneous multi-task learning. Regular multi-task learning only considers homogeneous tasks such as regression or classification individually. Joint classification and regression can be regarded as a learning paradigm for handling heterogeneous tasks.

First, logistic regression is used for disease classification, which minimizes the following loss function:

$$\mathcal{L}_1(W) = \sum_{i=1}^{n} \sum_{k=1}^{c_1} \left(y_{ik} \log \sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i} - y_{ik} \mathbf{w}_k^T x_i \right). \tag{5.6}$$

Here, we perform three binary classification tasks for the following three diagnostic groups respectively ($c_1 = 3$): AD, Mild Cognitive Impairment (MCI), and health control (HC).

Second, we use the traditional multivariate least squares regression model to predict memory scores. Under the regression matrix $P \in \Re^{d \times c_2}$, the least squares loss is defined by

$$\mathcal{L}_2(P) = \left\| X^T P - Z \right\|_F^2 , \tag{5.7}$$

where X is the data points matrix, P is the coefficient matrix of regression with c_2 tasks, the label matrix $Z = \left[(\mathbf{z}^1)^T, (\mathbf{z}^2)^T, \cdots, (\mathbf{z}^n)^T\right]^T \in \Re^{n \times c_2}$.

We perform the joint classification and regression tasks, the disease sensitive and QT relevant biomarker identification task can be formulated as the following objective:

$$\begin{aligned} \min_V \; & \sum_{i=1}^{n} \sum_{k=1}^{c_1} \left(y_{ik} \log \sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i} - y_{ik} \mathbf{w}_k^T \mathbf{x}_i \right) \\ & + \left\| X^T P - Z \right\|_F^2 + \gamma_1 \left\| V \right\|_{G_1} + \gamma_2 \left\| V \right\|_{2,1}, \end{aligned} \tag{5.8}$$

where $V = [W \;\; P] \in \Re^{d \times (c_1 + c_2)}$. As a result, the identified biomarkers will be correlated to memory scores and also be discriminative to disease categories.

Because the objective in Eq. (5.8) is a non-smooth problem and cannot be easily solved in general, we derive a new efficient algorithm to solve this problem in the next subsection.

5.2.3 Optimization Algorithm

We take the derivatives of Eq. (5.8) with respect to W and P respectively, and set them to zeros, we have

$$\frac{\partial \mathcal{L}_1(W)}{\partial W} + 2\gamma_1 \sum_{i=1}^{c_1} D_i \mathbf{w}_i + 2\gamma_2 D W = 0, \tag{5.9}$$

$$2XX^T P - 2XZ + 2\gamma_1 \sum_{i=c_1+1}^{c_2} D_i \mathbf{p}_i + 2\gamma_2 D P = 0, \tag{5.10}$$

where $D_i (1 \leq i \leq c_1 + c_2)$ is a block diagonal matrix with the k-th diagonal block as $\frac{1}{2\left\|\mathbf{v}_i^k\right\|_2} I_k$ (I_k is a d_k by d_k identity matrix), D is a diagonal matrix with the k-th diagonal element as $\frac{1}{2\left\|\mathbf{v}^k\right\|_2}$. Because $D_i (1 \leq i \leq c_1 + c_2)$ and D depend on $V = [\; W \quad P \;]$, they are also unknown variables to be optimized. In this chapter, we provide an iterative algorithm to solve Eq. (5.8). First, we guess a random solution

$V \in \Re^{d\times(c_1+c_2)}$, then we calculate the matrices $D_i(1 \le i \le c_1 + c_2)$ and D according to the current solution V. After obtaining the $D_i(1 \le i \le c_1 + c_2)$ and D, we can update the solution $V = [\ W \quad P\ ]$ based on Eq. (5.9). Specifically, the i-th column of P is updated by $\mathbf{p}_i = (XX^T + \gamma_1 D_i + \gamma_2 D)^{-1} X\mathbf{z}_i$. We cannot update W with a closed form solution based on Eq. (5.9), but we can obtained the updated W by the Newton's method. According to Eq. (5.9), we need to solve the following problem:

$$\min_{W}\ \mathcal{L}_1(W) + \gamma_1 \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i + \gamma_2 Tr(W^T D W). \tag{5.11}$$

Similar to the traditional method in the logistic regression [65, 66], we can use the Newton's method to obtain the solution W.

For the first term, the traditional logistic regression derivatives can be applied to get the first and second order derivatives [66].

For the second term, the first and second order derivatives are

$$\begin{aligned} \frac{\partial \sum\limits_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i}{\partial W_{up}} &= 2D_p(u,u)W_{up}\,, \\ \frac{\partial \sum\limits_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i}{\partial W_{up} \partial W_{vq}} &= 2D_p(u,u)\delta_{uv}\delta_{pq}, \end{aligned} \tag{5.12}$$

where $D_p(u,u)$ is the u-th diagonal element of D_p.

For the third term, the first and second order derivatives are

$$\begin{aligned} \frac{\partial Tr(W^T D W)}{\partial W_{up}} &= 2D(u,u)W_{up}\,, \\ \frac{\partial Tr(W^T D W)}{\partial W_{up} \partial W_{vq}} &= 2D(u,u)\delta_{uv}\delta_{pq}. \end{aligned} \tag{5.13}$$

After obtaining the updated solution $V = [\ W \quad P\ ]$, we can calculate the new matrices $D_i(1 \le i \le c_1 + c_2)$ and D. This procedure is repeated until the algorithm converges. The detailed algorithm is listed in Algorithm 3. We will prove that the above algorithm will converge to the global optimum.

5.2.4 Algorithm Analysis

To prove the convergence of the proposed algorithm, we need a lemma as follows.

Lemma 3 *For any vectors* $\mathbf{v}$ *and* $\mathbf{v}_0$*, we have the following inequality:* $\|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}_0\|_2} \leq \|\mathbf{v}_0\|_2 - \frac{\|\mathbf{v}_0\|_2^2}{2\|\mathbf{v}_0\|_2}$.

Proof: Obviously, $-(\|\mathbf{v}\|_2 - \|\mathbf{v}_0\|_2)^2 \leq 0$, so we have

$$\begin{aligned} & -(\|\mathbf{v}\|_2 - \|\mathbf{v}_0\|_2)^2 \leq 0 \Rightarrow 2\|\mathbf{v}\|_2 \|\mathbf{v}_0\|_2 - \|\mathbf{v}\|_2^2 \leq \|\mathbf{v}_0\|_2^2 \\ & \Rightarrow \|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}_0\|_2} \leq \|\mathbf{v}_0\|_2 - \frac{\|\mathbf{v}_0\|_2^2}{2\|\mathbf{v}_0\|_2}, \end{aligned} \tag{5.14}$$

which completes the proof. □

Then we prove the convergence of the algorithm, which is described in the following theorem.

Theorem 4 *The algorithm decreases the objective value of problem (5.8) in each iteration.*

Proof: In each iteration, suppose the updated W is $\tilde{W}$, and the updated P is $\tilde{P}$, then the updated V is $\tilde{V} = [\ \tilde{W} \quad \tilde{P}\ ]$. From Step 3 in the Algorithm 3, we know that:

$$\begin{aligned} & \mathcal{L}_1(\tilde{W}) + \gamma_1 \sum_{i=1}^{c_1} \tilde{\mathbf{w}}_i^T D_i \tilde{\mathbf{w}}_i + \gamma_2 Tr(\tilde{W}^T D \tilde{W}) \\ & \leq \mathcal{L}_1(W) + \gamma_1 \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i + \gamma_2 Tr(W^T D W). \end{aligned} \tag{5.15}$$

According to Step 4, we have:

$$\begin{aligned} & \left\| X^T \tilde{P} - Y \right\|_F^2 + \gamma_1 \sum_{i=1}^{c_2} \tilde{\mathbf{p}}_i^T D_i \tilde{\mathbf{p}}_i + \gamma_2 Tr(\tilde{P}^T D \tilde{P}) \\ & \leq \left\| X^T P - Y \right\|_F^2 + \gamma_1 \sum_{i=1}^{c_2} \mathbf{p}_i^T D_i \mathbf{p}_i + \gamma_2 Tr(P^T D P). \end{aligned} \tag{5.16}$$

Based on the definitions of $D_i(1 \leq i \leq c_1 + c_2)$ and D, and Lemma 1, we have two following inequalities:

$$\begin{aligned}
& \sum_{k=1}^{K} \left\|\tilde{\mathbf{v}}_i^k\right\|_2 - \sum_{k=1}^{K} \frac{\left\|\tilde{\mathbf{v}}_i^k\right\|_2^2}{2\left\|\mathbf{v}_i^k\right\|_2} \leq \sum_{k=1}^{K} \left\|\mathbf{v}_i^k\right\|_2 - \sum_{k=1}^{K} \frac{\left\|\mathbf{v}_i^k\right\|_2^2}{2\left\|\mathbf{v}_i^k\right\|_2} \\
\Rightarrow\ & \sum_{k=1}^{K} \left\|\tilde{\mathbf{v}}_i^k\right\|_2 - \tilde{\mathbf{v}}_i^T D_i \tilde{v}_i \leq \sum_{k=1}^{K} \left\|\mathbf{v}_i^k\right\|_2 - \mathbf{v}_i^T D_i \mathbf{v}_i \\
\Rightarrow\ & \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{K} \left\|\tilde{\mathbf{v}}_i^k\right\|_2 - \gamma_1 \sum_{i=1}^{c_1+c_2} \tilde{\mathbf{v}}_i^T D_i \tilde{\mathbf{v}}_i \\
\leq\ & \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{K} \left\|\mathbf{v}_i^k\right\|_2 - \gamma_1 \sum_{i=1}^{c_1+c_2} \mathbf{v}_i^T D_i \mathbf{v}_i, \qquad (5.17)
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{k=1}^{d} \left\|\tilde{\mathbf{v}}^k\right\|_2 - \sum_{k=1}^{d} \frac{\left\|\tilde{\mathbf{v}}^k\right\|_2^2}{2\left\|\mathbf{v}^k\right\|_2} \leq \sum_{k=1}^{d} \left\|\mathbf{v}^k\right\|_2 - \sum_{k=1}^{d} \frac{\left\|\mathbf{v}^k\right\|_2^2}{2\left\|\mathbf{v}^k\right\|_2} \\
\Rightarrow\ & \gamma_2 \sum_{k=1}^{d} \left\|\tilde{\mathbf{v}}^k\right\|_2 - \gamma_2 Tr(\tilde{V}^T D \tilde{V}) \\
\leq\ & \gamma_2 \sum_{k=1}^{d} \left\|\mathbf{v}^k\right\|_2 - \gamma_2 Tr(V^T D V). \qquad (5.18)
\end{aligned}$$

Note that the following two equalities:

$$\begin{aligned}
\sum_{i=1}^{c_1+c_2} \mathbf{v}_i^T D_i \mathbf{v}_i &= \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i + \sum_{i=1}^{c_2} \mathbf{p}_i^T D_i \mathbf{p}_i, \\
Tr(V^T D V) &= Tr(W^T D W) + Tr(P^T D P),
\end{aligned} \qquad (5.19)$$

then by adding Eqs. (5.15–5.18) in the both sides, we arrive at

$$\begin{aligned}
& \mathcal{L}_1(\tilde{W}) + \mathcal{L}_2(\tilde{P}) + \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{K} \left\|\tilde{\mathbf{v}}_i^k\right\|_2 + \gamma_2 \sum_{k=1}^{d} \left\|\tilde{\mathbf{v}}^k\right\|_2 \\
\leq\ & \mathcal{L}_1(W) + \mathcal{L}_2(P) + \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{K} \left\|\mathbf{v}_i^k\right\|_2 + \gamma_2 \sum_{k=1}^{d} \left\|\mathbf{v}^k\right\|_2.
\end{aligned}$$

Therefore, the algorithm decreases the objective value of problem (5.8) in each iteration. □

In the convergence, W, P, $D_i(1 \leq i \leq c_1+c_2)$ and D satisfy the Eq. (5.9). As the Eq. (5.8) is a convex problem, satisfying the Eq. (5.9) indicates that $V = [\ W \quad P\ ]$ is a global optimum solution to the Eq. (5.8). Therefore, the Algorithm 3 will converge to the global optimum of the Eq. (5.8). Because our algorithm has the closed form solution in each iteration, the convergency is very fast.

5.3 Empirical Studies and Discussions

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.ucla.edu`). One goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see `www.adni-info.org`. Following a prior imaging genetics study [16], 733 non-Hispanic Caucasian participants were included in this study. We empirically evaluate the proposed method by applying it to the ADNI cohort, where a wide range of multi-modal biomarkers are examined and selected to predict memory performance measured by five RAVLT scores and classify participants into health control (HC), MCI and AD.

5.3.1 Experimental Design

Overall Setting. Our primary goal is to identify relevant genetic and imaging biomarkers that can classify disease status and predict memory scores (Fig. 5.2). We describe our genotyping, imaging and memory data in Section 5.3.1; present the identified biomarkers in Section 5.3.2; discuss the disease classification in Section 5.3.3; and demonstrate the memory score prediction in Section 5.3.4.

Genotyping Data. The single nucleotide polymorphism (SNP) data [32] were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA). Among all SNPs, only SNPs, belonging to the top 40 AD candidate genes listed on the AlzGene database (www.alzgene.org) as of June 10, 2010, were selected after the standard quality control (QC) and imputation steps. The QC criteria for the SNP data include (1) call rate check per subject and per SNP marker, (2) gender check, (3) sibling pair identification, (4) the Hardy-Weinberg equilibrium test, (5) marker removal by the minor allele frequency and (6) population stratification. The quality-controlled SNP were then imputed using the MaCH software to estimate the missing genotypes. After that, the Illumina annotation information based on the Genome build 36.2 was used to select a subset of SNPs, belonging or proximal to the top 40 AD candidate genes. This procedure yielded 1224 SNPs, which were annotated with 37 genes [?]. For the remaining 3 genes, no SNPs were available on the genotyping chip.

Imaging Biomarkers. In this study, we use the baseline structural MRI and molecular FDG-PET scans, from which we extract imaging biomarkers. Two widely employed automated MRI analysis techniques were used to process and extract imaging genotypes across the brain from all baseline scans of ADNI participants as previously described [16]. First, voxel-based morphometry (VBM) [35] was performed to define global gray matter (GM) density maps and extract local GM density values for 86 target regions (Fig. 5.4(a)). Second, automated parcellation via freeSurfer V4 [17] was conducted to define 56 volumetric and cortical thickness values (Fig. 5.4(b)) and to extract total intracranial volume (ICV). Further information about these measures is available in [16]. All these measures were adjusted for the baseline age, gender, education, handedness, and baseline ICV using the regression weights derived from the healthy control participants. For PET images, following [67], mean glucose

Table 5.1.RAVLT cognitive measures as responses in multi-task learning.

Task ID	Description of RAVLT scores
TOTAL	Total score of the first 5 learning trials
TOT6	Trial 6 total number of words recalled
TOTB	List B total number of words recalled
T30	30 minute delay total number of words recalled
RECOG	30 minute delay recognition score

metabolism (CMglu) measures of 26 regions of interest (ROIs) in the Montreal Neurological Institute (MNI) brain atlas space were employed in this study (Fig. 5.4(c)).

Memory Data. The cognitive measures we use to test the proposed method are the baseline RAVLT memory scores from all ADNI participants. The standard RAVLT format starts with a list of 15 unrelated words (List A) repeated over five different trials and participants are asked to repeat. Then the examiner presents a second list of 15 words (List B), and the participant is asked to remember as many words as possible from List A. Trial 6, termed as 5 minute recall, requests the participant again to recall as many words as possible from List A, without reading it again. Trial 7, termed as 30 minute recall, is administrated in the same way as Trial 6, but after a 30 minute delay. Finally, a recognition test with 30 words read aloud, requesting the participant to indicate whether or not each word is on List A. The RAVLT has proven useful in evaluating verbal learning and memory. Table 5.1 summarizes five RAVLT scores used in our experiments.

Participant Selection. In this study, we included only participants with no missing data for all above four types (views) of features and cognitive scores, which resulted in a set of 345 subjects (88 HC, 174 MCI and 88 AD). The feature sets extracted from baseline multimodal data of these subjects are summarized in Table 6.1.

Table 5.2.Multi-modal feature sets as predictors in multi-view learning.

View ID (Feature Set ID)	Modality	# Features
VBM	MRI	86
FreeSurfer	MRI	56
FDG-PET	FDG-PET	26
SNP	Genetics	1244

5.3.2 Biomarker Identifications

The proposed heterogeneous multi-task learning scheme aims to identify genetic and phenotypic biomarkers that are associated with both cognition (*e.g.*, RAVLT in this study) and disease status in a joint regression and classification framework. Here we first examine the identified biomarkers. Shown in Fig. 5.4 is a summarization of selected features for all four data types, where the regression/classification weights are color-mapped for each feature and each task.

In Fig. 5.4(a), many VBM measures are selected to be associated with disease status, which is in accordance with known global brain atrophy pattern in AD. The VBM measures associated with RAVLT scores seem to be a subset of those disease sensitive markers, showing a specific memory circuitry contributing to the disease, as well as suggesting that the disease is implicated by not only this memory function but also other complicated factors. Evidently, the proposed method could have a potential to offer deep mechanistic understandings. Shown in Fig. 5.5 is a comparison between RAVLT-relevant markers and AD-relevant markers and their associated weights mapped onto a standard brain space.

Fig. 5.4(b) shows the identified markers from the FreeSurfer data. In this case, a small set of markers are discovered. These markers, such as hippocampal volume, amygdala volume, and entorhinal cortex thickness, are all well-known AD-relevant markers, showing the effectiveness of the proposed method. These markers are also shown to be associated with both AD and RAVLT. The FDG-PET findings (Fig. 5.4(c)) are also interesting and promising. The AD-relevant biomarkers include angular, hippocampus, middle temporal, and post cingulate regions, which agrees with prior findings (*e.g.*, [67]). Again, a subset of these markers are also relevant to RAVTL scores.

As to the genetics, only top findings are shown in Fig. 5.4(d). The APOE E4 SNP (rs429358), the best known AD risk factor, shows the strongest link to both disease status and RAVLT scores. A few other important AD genes, including recently discovered and replicated PICALM and BIN1, are also included in the results. For those newly identified SNPs, further investigation in independent cohorts should be warranted.

5.3.3 Improved Disease Classification

We classify the selected participants of ADNI cohort using the proposed methods by integrating the four different types of data. We report the classification performances of our method. We compare our methods against several most recent multiple kernel learning (MKL) methods that are able to make use of multiple types of data including SVM ℓ_∞ MKL method [51], SVM ℓ_1 MKL [49], SVM ℓ_2 MKL method [47], least square (LSSVM) ℓ_∞ MKL method [48], LSSVM ℓ_1 MKL method [46] and LSSVM ℓ_2 MKL method [45]. We also compare a related method, Heterogeneous Multi-task Learning (HML) method [64], which simultaneously conducts classification and regression like our method. However, because this method is designed for homogenous input data and is not able to deal with multiple types of data at the same time, we concatenate the four types of features as its input. In addition, we report the classification performances by our method and SVM on each individual types of data as baselines. SVM on a simple concatenation of all four types of features are also reported. In our experiments, we conduct three-class classification, which is more desirable and more challenging than binary classifications using each pair of three categories.

We conduct standard 5-fold cross-validation and report the average results. For each of the 5 trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters. The parameters of our methods (γ_1 and γ_2 in Eq. (5.8)) are optimized in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$. For SVM method and MKL methods, one Gaussian kernel is constructed for each type of features ($i.e.$, $\mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||_2^2\right)$), where the parameters γ are fine tuned in the same range used as our method. We implement the MKL methods using the codes published by [45]. Following [45], in LSSVM ℓ_∞ and ℓ_2 methods, the regularization parameter λ is estimated jointly as the kernel coefficient of an identity matrix; in LSSVM ℓ_1 method, λ is set to 1; in all other SVM approaches, the C parameter of the box constraint is set to 1. We use LIBSVM[1] software package to implement SVM. We implement HML method following the details in its original work, and set the parameters to be optimal. The classification performances measured by classification accuracy of all compared methods in AD detection are reported in Table 5.3.

A first glance at the results shows that our methods consistently outperform all other compared methods, which demonstrates the effectiveness of our methods in early AD detection. In addition, the methods using multiple data sources are generally better than their counterparts using one single type of data. This confirms the usefulness of data integration in AD diagnosis. Moreover, our methods always outperform the MKL methods in these experiments, although both take advantage of multiple data sources. This observation is consistent with our theoretical analysis.

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

Table 5.3. Classification performance comparison between the proposed method and related methods for distinguishing HC, MCI and AD.

Methods	Accuracy (mean+std)
SVM (SNP)	0.561 ± 0.026
SVM (FreeSurfer)	0.573 ± 0.012
SVM (VBM)	0.541 ± 0.032
SVM (PET)	0.535 ± 0.026
SVM (all)	0.575 ± 0.019
HML (all)	0.638 ± 0.019
SVM ℓ_∞ MKL method	0.624 ± 0.031
SVM ℓ_1 MKL method	0.593 ± 0.042
SVM ℓ_2 MKL method	0.561 ± 0.037
LSSVM ℓ_∞ MKL method	0.614 ± 0.031
LSSVM ℓ_1 MKL method	0.585 ± 0.018
LSSVM ℓ_2 MKL method	0.577 ± 0.033
Our method (SNP)	0.673 ± 0.021
Our method (FreeSurfer)	0.689 ± 0.029
Our method (VBM)	0.669 ± 0.031
Our method (PET)	0.621 ± 0.028
Our method	**0.726 ± 0.032**

That is, our methods not only assign proper weight to each type of data, but also consider the relevance of the features inside each individual type of data. In contrast, the MKL methods address the former while not taking into account the latter.

5.3.4 Improved Memory Performance Prediction

Now we evaluate the memory performance prediction capability of the proposed method. Because the cognitive scores are continuous, we evaluate the proposed method via regression and compare it to two baseline methods, *i.e.*, multivariate linear regression (MRV) and ridge regression. Because both MRV and ridge regression are for single-type input data, we conduct regression on each of the four types of features and a simple concatenation of them. Similarly, we also predict memory per-

Table 5.4. Comparison of memory prediction performance measured by average RMSEs (smaller is better).

Test case	TOTAL	TOT6	TOTB	T30	RECOG
MRV (SNP)	6.153	2.476	2.168	2.201	3.483
MRV (FreeSurfer)	5.928	2.235	2.039	2.088	3.339
MRV (VBM)	6.093	2.289	2.142	2.137	3.394
MRV (PET)	6.246	2.514	2.237	2.215	3.615
MRV (all)	5.909	2.232	1.992	2.032	3.306
Ridge (SNP)	6.076	2.416	2.147	2.117	3.368
Ridge (FreeSurfer)	5.757	2.203	2.004	2.017	3.237
Ridge (VBM)	5.976	2.147	2.038	2.129	3.249
Ridge (PET)	6.153	2.443	2.186	2.107	3.515
Ridge (all)	5.704	2.143	1.989	1.994	3.193
Our method (SNP)	5.991	2.201	2.008	2.001	3.107
Our method (FreeSurfer)	5.601	2.106	1.947	1.886	3.015
Our method (VBM)	5.715	2.011	1.899	1.974	3.041
Our method (PET)	6.013	2.241	2.017	2.017	3.331
Our method (all)	5.506	1.984	1.886	1.841	2.989

formance by our method on the same test conditions. When multiple-type input data is used, as demonstrated in Section 5.3.2, our method automatically and adaptively select the prominent biomarkers for regression. For each test case, we conduct standard 5-fold cross-validation and report the average results. For each of the 5 trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$ for both ridge regression and our method. For our method, in each trial, from the learned coefficient matrix we sum the absolute values of the coefficients of a single feature over all the tasks as the overall weight, from which we pick up the features with non-zero weights (*i.e.*, $w > 10^{-3}$) to predict regression responses for test data. The performance assessed by root mean square error (RMSE), a widely used measurement for statistical regression analysis, are reported in Table 5.4.

From Table 5.4 we can see that the proposed method always has better memory prediction performance. Among the test cases, the FreeSurfer imaging measures and VBM imaging measure have similar predictive power, which are better than those of PET imaging measures and SNP features. In general, combining the four types of features are better than only using one type of data. Because our method adaptively weight each type of data and each feature inside a type of data, it has the least regression error when using all available input data. These results, again, demonstrated the usefulness of our method and data integration in early AD diagnosis.

5.4 Conclusions

In this chapter, we proposed a novel sparse multi-modal multi-task learning method to identify the disease sensitive biomarkers via integrating heterogeneous imaging genetics data. We utilized the joint classification and regression learning model to identify the disease sensitive and QT relevant biomarkers. We introduced a novel combined structured sparsity regularization to integrate heterogeneous imaging genetics data, and derived a new efficient optimization algorithm to solve our non-smooth objective function and followed with the rigorous theoretical analysis on the global convergency. The empirical results showed our method improved both memory scores prediction and disease classification accuracy.

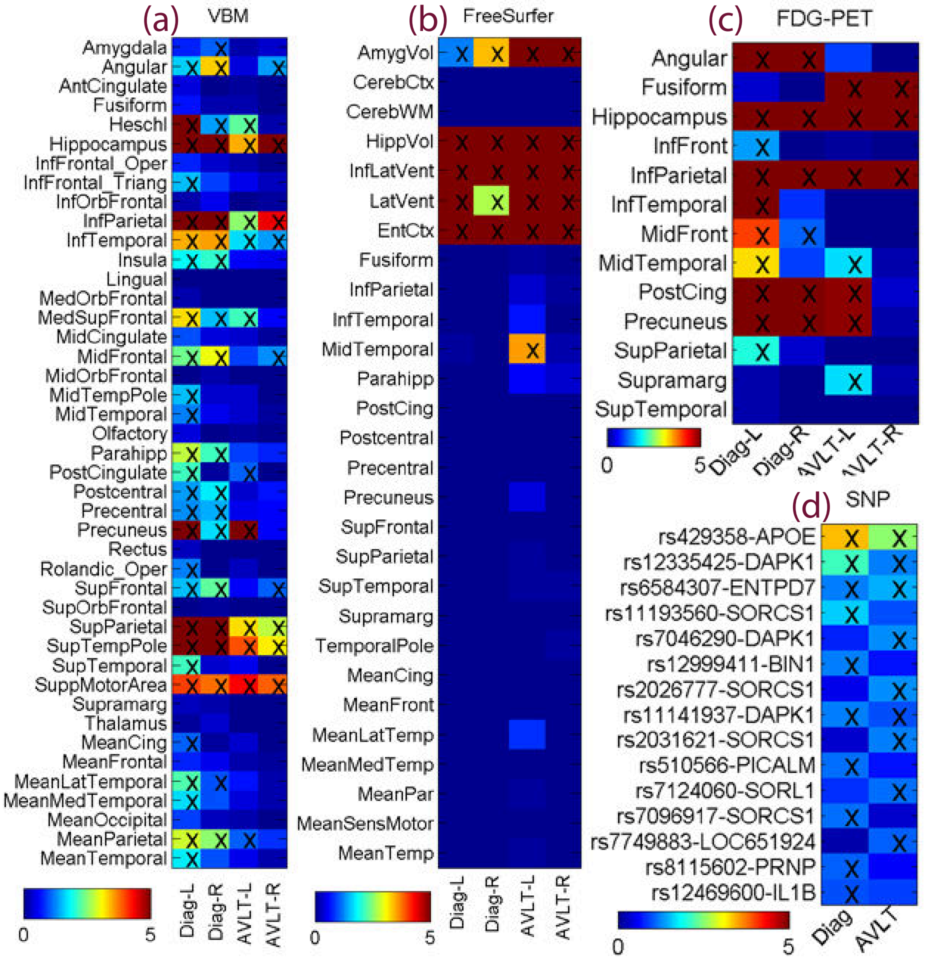
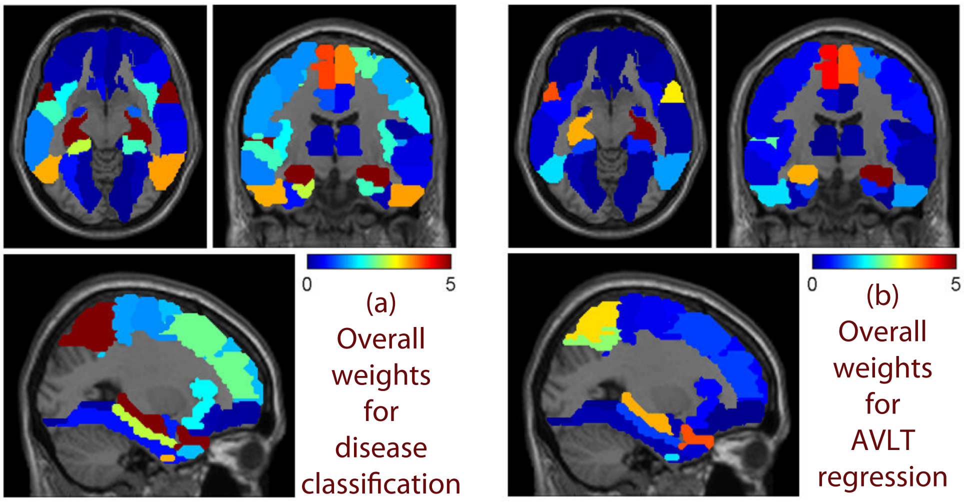


Figure 5.4. Weight maps for multi-modal data: (a) VBM measures from MRI, (b) FreeSurfer measures from MRI, (c) glucose metabolism from FDG-PET, and (d) top SNP findings. Weights for disease classification were labeled as Diag-L (left side), Diag-R (right side), or Diag; and weights for RAVLT regression were labeled as AVLT-L, AVLT-R or AVLT. In (a-c), weights were normalized by dividing the corresponding threshold used for feature selection, and thus all selected features had normalized weights ≥ 1 and were marked with “x”. In (d), only top SNPs were shown, weights were normalized by dividing the weight of the 10th top SNP, and the top 10 SNPs for either classification or regression task had normalized weights ≥ 1 and were marked with “x”..



(a) Overall weights for disease classification

(b) Overall weights for AVLT regression

Figure 5.5. VBM weights of joint regression of AVLT scores and classification of disease status were mapped onto brain..

CHAPTER 6

FROM PHENOTYPE TO GENOTYPE: AN ASSOCIATION STUDY OF LONGITUDINAL PHENOTYPIC MARKERS TO ALZHEIMER'S DISEASE RELEVANT SNPS

6.1 Introduction

Neuroimaging genetics is an emerging research field, where brain imaging is used as quantitative phenotypes to investigate the role of genetic variation in brain structure and function. It holds great promise for a systems biology of the brain to better understand complex neurobiological systems, from genetic determinants to cellular processes to the complex interplay of brain structure, function, behavior and cognition. Disorders of the nervous system are associated with complex neurobiological changes, which may lead to profound alterations at all levels of organization.

Genome-wide association studies (GWAS) have been increasingly performed to correlate high-throughput SNP data to large-scale imaging data. To facilitate such association analysis, many studies employed a hypothesis-driven approach [30] by making significant reduction in one or both data types. For example, some whole brain studies focused on a small number of genetic variables, *e.g.*, [68–71], and some whole genome studies examined a limited number of imaging variables, *e.g.*, [29, 72, 73]. Many SNPs have been identified as risk factors for Alzheimer's Disease (AD), see those in the AlzGene database (www.alzgene.org).

So far most studies focus on selecting and associating SNPs to AD status or imaging phenotypes. Very few studies have been done to directly examine how the SNP values change when phenotypic measures are varied, *i.e.*, via regression of SNP values on phenotypic measures. This alternative approach may have a potential to help us discover important imaging genetic associations from a different perspective. In this study, we perform such an initial analysis for finding phenotypic imaging markers which are related to SNPs from or proximal to AlzGene candidates.

Neuroimaging measures have been widely studied to predict disease status and/or cognitive performance [5, 6]. However, whether these measures coupled with their longitudinal profiles have sufficient power to infer relevant genotype groups is still an under-explored yet important topic in AD research. A simple strategy typically used in longitudinal studies (*e.g.*, [74]) is to analyze a single summarized value such as average change, rate of change, or slope. This approach may be inadequate to distinguish the complete dynamics of cognitive trajectories and thus become unable to identify the underlying genetic structure.

With these observations, in this work, we propose a new task-correlated longitudinal sparse regression framework to effectively identify the longitudinal phenotypic markers related to candidate AD SNPs. Based on the emerging structured sparse learning techniques, which has been effectively applied in imaging genetics studies, the new combined structured sparse regularizations are introduced to tackle the longitudinal phenotypic patterns and biological genotypic correlations. The proposed new computational biology model consists of three major components. First, due to the serial measures of the imaging phenotypes over time, we propose a novel longitudinal regression analysis method. As a result, the regression coefficients assess the relationships between longitudinal phenotypes and their genetic makeups. Second, certain SNPs are naturally correlated via different ways, *e.g.* multiple SNPs from one single gene often jointly carry out similar genetic functionalities, SNPs in high linkage disequilibrium (LD) are linked together in meiosis. To incorporate such SNP correlations in our association studies, we propose to use the trace/nuclear norm regularization [75] to approximately minimize the rank of regression coefficient matrix, such that the coefficients of phenotypes associated to correlated SNPs are linearly dependent. Lastly, through enforcing the $\ell_{2,1}$-norm regularization, the imaging feature selection across most SNPs are coupled [12, 13], so that the identified imaging phenotypes are longitudinally stable and have common influence on all the SNPs.

We apply the proposed method to the ADNI cohort [15] for identifying longitudinal phenotypes using a set of SNPs based on the AlzGene database. Our empirical results yield not only clearly improved prediction performance in all test cases, but also a compact set of associations between phenotypes and genotypes that are in accordance with prior research findings.

6.2 Materials and Data Sources

Both SNP and structural magnetic resonance imaging (MRI) data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). One goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, we refer interested readers to www.adni-info.org.

6.2.1 SNP genotypes

The SNP data used in this study [32] were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA). Among all SNPs, only SNPs, belonging to the top 40 AD candidate genes listed on the AlzGene database (www.alzgene.org) as of June 10, 2010, were selected after the standard quality control (QC) and imputation steps. The QC criteria for the SNP data include (1) call rate check per subject and per SNP marker, (2) gender check, (3) sibling pair identification, (4) the Hardy-Weinberg equilibrium test, (5) marker removal by the minor allele frequency and (6) population stratification. As the second pre-processing step, the quality-controlled SNPs were imputed using the MaCH software [33] to estimate the missing genotypes. After that, the Illumina annotation information based on the Genome build 36.2 was used to select a subset of SNPs, belonging to the top 40 AD candidate genes [34]. The above procedure yielded 1224 SNPs from 37 genes. For the remaining 3 genes, no SNPs were available on the genotyping chip.

6.2.2 MRI analysis and extraction of imaging phenotypes

Two widely employed automated MRI analysis techniques were used to process and extract imaging genotypes across the brain from all the MRI scans of ADNI participants as previously described [16]. First, voxel-based morphometry (VBM) [35] was performed to define modulated gray matter (GM) maps and extract local GM values for target regions. Second, automated parcellation via FreeSurfer V4 [17,76] was conducted to define volumetric and cortical thickness values for regions of interest (ROIs) and to extract total intracranial volume (ICV). Further information is available in [16]. The time points examined in this study for imaging markers included baseline (BL), Month 6 (M6), Month 12 (M12) and Month 24 (M24). All the participants with no missing BL/M6/M12/M24 MRI measurements were included in this study. Fig. 6.2 shows the names of these ROIs in the brain space. All these measures were adjusted for baseline ICV using the regression weights derived from the healthy control (HC) participants.

6.3 Task-Correlated Longitudinal Sparse Regression

For the association study of longitudinal imaging phenotypes to the genotypes, the input imaging features are a set of matrices $\mathcal{X} = \{X_1, X_2, \ldots, X_T\} \in \mathbb{R}^{d\times n\times T}$ corresponding to the measurements at T consecutive time points, where X_t is the imaging measurements for a certain type of imaging markers, such as VBM or FreeSurfer markers used in this study, at time $t\,(1 \leq t \leq T)$. Obviously, $\mathcal{X}$ is a tensor data with d imaging features, n subject samples and T time points. The output genetic variations described by c SNPs for the n subject samples forms a matrix $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T \in \mathbb{R}^{n\times c}$, where the $\mathbf{y}_i \in \mathbb{R}^c$ is the SNPs values of the i-th subject sample. Our goal is to learn from $\{\mathcal{X}, Y\}$ a model that can reveal the associations between the longitudinal imaging phenotypes $\mathcal{X}$ and the genotypes Y.

A straightforward method for relating imaging phenotypes and SNPs is to perform regression at each time point separately, which, though, does not take into account the valuable information conveyed by the longitudinal patterns of the phenotypic inputs. To overcome this limitation, different from previous studies that learned the regression coefficient matrix for each time point individually, we aim to learn a unified longitudinal regression model to find the genetic features which are associated to the longitudinal imaging patterns over all the measurement time points. To this end, we expect to learn a coefficient tensor (a stack of coefficient matrices) $\mathcal{B} = \{B_1, \cdots, B_T\} \in \mathbb{R}^{d\times c\times T}$ to reveal the temporal changes of the coefficient matrices. In this chapter, we propose to use the low-rank structured sparse regularizations to explore the temporal patterns and the interrelatedness between SNPs in a new task-correlated longitudinal sparse regression model.

6.3.1 Task-correlated longitudinal sparse regression using low-rank structured sparse regularizations

The simplest model to associate the the phenotypic markers to the genotypes is the multivariate regression model, which solves the following optimization problem:

$$\min_{\mathcal{B}} \; J_0 = \mathcal{L}\left(\mathcal{B}\right) + \gamma ||\mathcal{B}||_2^2 = \mathcal{L}\left(\mathcal{B}\right) + \gamma \sum_{t=1}^{T} \sum_{k=1}^{d} ||\mathbf{b}_t^k||_2^2. \tag{6.1}$$

where $\mathbf{b}_t^k$ denotes the k-th row of coefficient matrix B_t at time t, and $\mathcal{L}\left(\mathcal{B}\right)$ is the proposed longitudinal loss and defined as:

$$\mathcal{L}\left(\mathcal{B}\right) = ||\mathcal{B} \otimes_1 \mathcal{X}^T - Y||_F^2 = \sum_{t=1}^{T} ||X_t^T B_t - Y||_F^2. \tag{6.2}$$

Because the objective J_0 in Eq. (6.1) can be decoupled for each individual time point and does not consider the longitudinal correlations between the imaging features and the SNPs, it is not suitable for longitudinal data analysis and feature selection. Because the selected imaging markers with temporal changes are desired to connect all the SNPs, the T groups of regression tasks at different time points should not be decoupled and have to be performed simultaneously. Thus we introduce the structured sparse regularization [12–14] into the longitudinal data regression and feature selection model as following:

$$\min_{\mathcal{B}} J_1 = \mathcal{L}\left(\mathcal{B}\right) + \gamma \sum_{k=1}^{d} \sqrt{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2}, \tag{6.3}$$

Apparently, J_1 in Eq. (6.3) can no longer be decoupled over time dimension. Upon solution, the imaging features with common influences to all the SNPs across all the time points will be identified out due to the second term in Eq. (6.3), which essentially is a tensor extension of the widely used $\ell_{2,1}$-norm for matrices.

To further take into account that many SNPs are interrelated together and their effects on brain structure or disease progression could overlap, we expect to further develop J_1 in Eq. (6.3) to leverage the useful information conveyed by the SNPs correlations. Mathematically speaking, due to the interrelatedness among the SNPs, the learning vector $(\mathbf{b}_t)_j$ should have certain correlations, where $(\mathbf{b}_t)_j$ denotes the j-th column of B_t. Namely, the coefficient matrices $B_t\,(1 \leq t \leq T)$ should be of low-rank. Given a general n-mode tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$, we denote $\text{unfold}_k(\mathcal{T}) = T_{(k)} \in \mathbb{R}^{I_k \times (I_1 \ldots I_{k-1} I_{k+1} \ldots I_n)}$ as the unfolding operation along its k-th mode. Then we can achieve our goal by minimizing the rank of $B_{(1)} = [B_1, B_2, \ldots, B_T] \in \mathbb{R}^{d \times (c \times T)}$ induced from $\mathcal{B}$, which leads to the following optimization problem:

$$\min_{\mathcal{B}} J_2 = \mathcal{L}(\mathcal{B}) + \gamma_1 \sum_{k=1}^{d} \sqrt{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2} + \gamma_2 \|B\|_* , \tag{6.4}$$

where $\|\|_*$ denotes the trace-norm of a matrix, and without ambiguity we drop the subscript of the matrix $B_{(1)}$ for notation brevity. Given a matrix $M \in \mathbb{R}^{n \times m}$ and its singular values $\sigma_i\,(1 \leq i \leq \min(n, m))$, the trace-norm of M is defined as $\|M\|_* = \sum_{i=1}^{\min(n,m)} \sigma_i = \text{Tr}\left(MM^T\right)^{\frac{1}{2}}$. It has been shown that [75] the trace-norm is the best convex approximation of the rank-norm. Therefore, the third term of J_2 in Eq. (6.4) indeed minimizes the rank of the unfolded learning model $\mathcal{B}$, such that the correlations among the SNPs are captured. Due to its both capabilities for imaging marker selection and task correlation integration, we call J_2 defined in Eq. (6.4) as the proposed *task-correlated longitudinal sparse regression model.*

6.3.2 A New Optimization Algorithm and Its Global Convergence

Because our new objective J_2 is non-smooth, the problem in Eq. (6.4) is difficult to solve in general. Some existing methods, such as LARS [58], linear gradient search [59], proximal [60] methods, can solve it, but not efficiently. Thus, in this subsection we derive a new efficient algorithm to solve J_2 with rigorous proof of its global convergence.

Taking the derivative of J_2 *w.r.t* B_t and set it to zeros, we have:

$$2X_tX_t^TB_t - 2X_tY + 2\gamma_1 DB_t + 2\gamma_2\bar{D}B_t = 0, \tag{6.5}$$

where D is a diagonal matrix with $D(k,k) = \frac{1}{2\sqrt{\sum_{t=1}^T ||\mathbf{b}_t^k||_2^2}}$ and $\bar{D} = \left(BB^T\right)^{-1/2}/2$. Thus, we can derive:

$$B_t = (X_tX_t^T + \gamma_1 D + \gamma_2\bar{D})^{-1}X_tY. \tag{6.6}$$

When the time t changes from 1 to T, we can compute $B_t\,(1 \leq t \leq T)$ by Eq. (6.6). Because D and $\bar{D}$ depend on B and can be seen as latent variables, we propose an iterative algorithm to obtain the global optimum solutions of $\mathcal{B}$ in Algorithm 4.

Algorithm 4: A new algorithm to minimize J_2 in Eq. (6.4).

```
Data: $\mathcal{X} \in \mathbb{R}^{d\times n\times T}$, $Y \in \mathbb{R}^{n\times c}$.
1. Initialize $\mathcal{B}^{(0)} \in \mathbb{R}^{d\times c\times T}$ using the regression results at each individual time point.
repeat
    2. Calculate the diagonal matrix $D$, where the $k$-th diagonal element is computed as $\frac{1}{2\sqrt{\sum_{t=1}^T ||\mathbf{b}_t^k||_2^2}}$.
    3. Calculate $\bar{D} = \frac{1}{2}\left(BB^T\right)^{-\frac{1}{2}}$.
    4. Update $B_t$ by $B_t = (X_tX_t^T + \gamma_1 D + \gamma_2\bar{D})^{-1}X_tY$.
until Converges
Result: $\mathcal{B} = \{B_1, B_2, \ldots, B_T\} \in \mathbb{R}^{d\times c\times T}$.
```

We summarize the convergence of Algorithm 4 as following.

Theorem 5 *Algorithm 4 monotonically decreases J_2 in Eq.* (6.4) *in each iteration, and converges to the globally optimal solution.*

Proof: In Algorithm 4, in each iteration we denote the updated B_t as $\tilde{B}_t$ and the updated $\mathcal{L}$ as $\tilde{\mathcal{L}}$. From step 4 we know that:

$$\begin{aligned} &\tilde{\mathcal{L}} + \gamma_1 \sum_{t=1}^{T} \mathrm{Tr}(\tilde{B}_t^T D \tilde{B}_t) + \gamma_2 \sum_{t=1}^{T} \mathrm{Tr}(\tilde{B}_t^T \bar{D} \tilde{B}_t) \leq \\ &\mathcal{L} + \gamma_1 \sum_{t=1}^{T} \mathrm{Tr}(B_t^T D B_t) + \gamma_2 \sum_{t=1}^{T} \mathrm{Tr}(B_t^T \bar{D} B_t). \end{aligned} \tag{6.7}$$

In each iteration, denote the updated B as $\tilde{B}$ and the updated $\mathbf{b}_t^k$ as $\tilde{\mathbf{b}}_t^k$, according to the definitions of D and $\bar{D}$, we can write:

$$\begin{aligned} &\tilde{\mathcal{L}} + \frac{\gamma_1}{2} \sum_{k=1}^{d} \frac{|| \sum_{t=1}^{T} \tilde{\mathbf{b}}_t^k ||_2^2}{\sqrt{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2}} + \frac{\gamma_2}{2} \mathrm{Tr}\left(\tilde{B}\tilde{B}^T \left(BB^T\right)^{-\frac{1}{2}} \right) \leq \\ &\mathcal{L} + \frac{\gamma_1}{2} \sum_{k=1}^{d} \frac{|| \sum_{t=1}^{T} \mathbf{b}_t^k ||_2^2}{\sqrt{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2}} + \frac{\gamma_2}{2} \mathrm{Tr}\left(BB^T \left(BB^T\right)^{-\frac{1}{2}} \right). \end{aligned} \tag{6.8}$$

Following [14], it can be verified that:

$$\begin{aligned} &\sqrt{\sum_{t=1}^{T} ||\tilde{\mathbf{b}}_t^k||_2^2} - \frac{\sum_{t=1}^{T} ||\tilde{\mathbf{b}}_t^k||_2^2}{2\sqrt{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2}} \leq \\ &\sqrt{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2} - \frac{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2}{2\sqrt{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2}}. \end{aligned} \tag{6.9}$$

$$\begin{aligned} &\mathrm{Tr}\left(\tilde{B}\tilde{B}^T\right)^{\frac{1}{2}} - \mathrm{Tr}\tilde{B}\tilde{B}^T \left(BB^T\right)^{-\frac{1}{2}} \leq \\ &\mathrm{Tr}\left(BB^T\right)^{\frac{1}{2}} - \mathrm{Tr}BB^T \left(BB^T\right)^{-\frac{1}{2}}. \end{aligned} \tag{6.10}$$

Adding the both sides of Eqs. (6.8–6.10) together, we obtain:

$$\begin{aligned} &\tilde{\mathcal{L}} + \gamma_1 \sum_{k=1}^{d} \sqrt{\sum_{t=1}^{T} ||\tilde{\mathbf{b}}_t^k||_2^2} + \gamma_2 \mathrm{Tr}\left(\tilde{B}\tilde{B}^T\right)^{\frac{1}{2}} \leq \\ &\mathcal{L} + \gamma_1 \sum_{k=1}^{d} \sqrt{\sum_{t=1}^{T} ||\mathbf{b}_t^k||_2^2} + \gamma_2 \mathrm{Tr}\left(BB^T\right)^{\frac{1}{2}}. \end{aligned} \tag{6.11}$$

Thus, our algorithm decreases the objective value of Eq. (6.4) in each iteration. When the objective value keeps unchange, Eq. (6.5) is satisfied, *i.e.*, the KKT condition of the objective is satisfied. Our algorithm reaches one of the optimal solution. Because our objective in Eq. (6.4) is a convex problem, our Algorithm 4 will converge to one of the globally optimal solution. □

Computational analysis. In the iteration loop of Algorithm 4, steps 2 is computationally trivial. Step 3 solves a singular value decomposition (SVD) problem, and step 4 solves a system of linear equations, both of which, thereby the whole algorithm, are well studied in literature and can be solved very efficiently by existing numerical packages.

6.4 Experimental Results and Discussions

In this section, we evaluate the proposed method by applying it to the ADNI cohort, where a wide range of imaging markers measured over a period of two years are examined and associated to SNPs that are relevant to AD. The goal is to discover a compact set of phenotypic imaging markers that are closely related to AD sensitive genotypes encoded by SNPs.

6.4.1 Improved prediction of SNPs from longitudinal phenotypic imaging markers

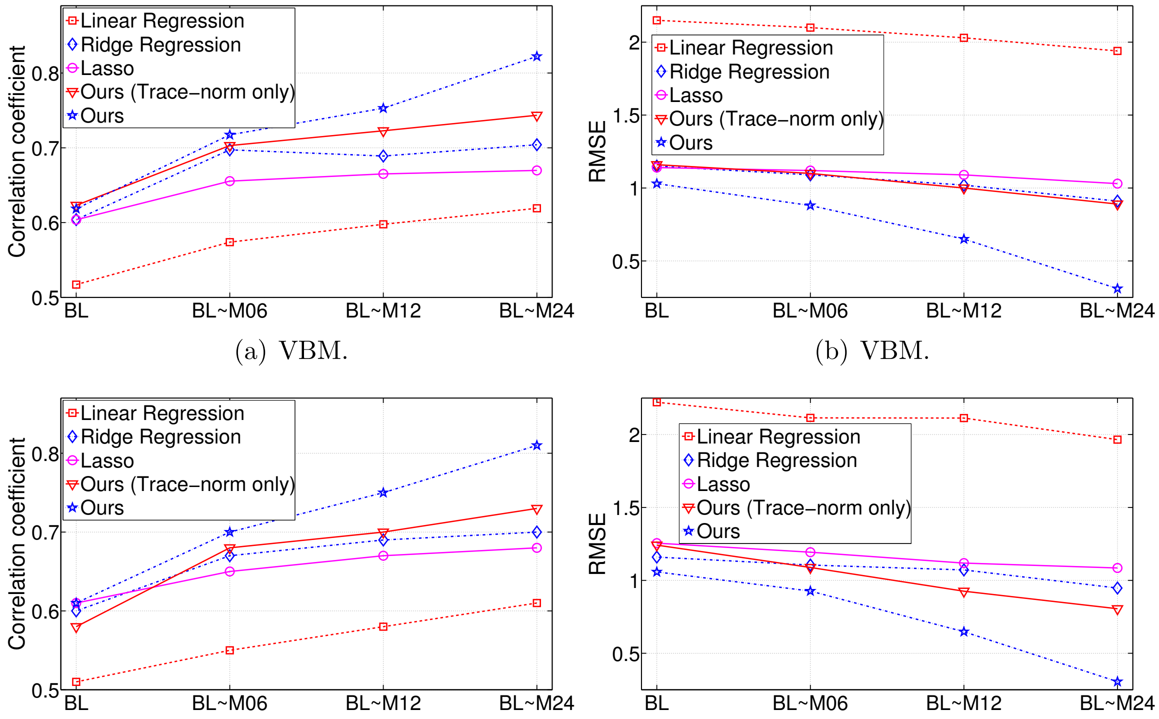
We first evaluate the proposed method by applying it to the ADNI cohort to predict the SNPs of the participants from each of their two types of imaging phenotypes, *i.e.*, VBM markers and FreeSurfer markers, tracked over four different time points, including baseline (BL) and 6/12/24-month (M06/M12/M24). Because some subjects of the ADNI cohort do not have complete imaging marker measurements over all the four time points, in our experiments we use the subject samples that have both

Table 6.1. Numbers of participants in the experiments using two different types of imaging markers.

Imaging phenotypes	# total	# AD	# MCI	# HC
VBM	424	86	194	144
FreeSurfer	474	100	216	158

SNPs data and complete imaging measurements. As a result, two subsets of ADNI subjects are included in our experiments, one for each type of imaging phenotypes, as detailed in Table 6.1.

We compare the proposed method against its three close counterparts including multivariate linear regression (LR) method, ridge regression (RR) method, and least absolute shrinkage and selection operator (Lasso) [11] method. LR method is the most broadly used association model in both statistical learning and imaging genetics. RR method is the regularized version of LR model to avoid over-fitting. Lasso method replaces the squared ℓ_2-norm regularization in RR method by the ℓ_1-norm regularization, from which sparse results can be achieved [11]. Different to these compared methods, our new association model imposes structured sparsity via the tensor $\ell_{2,1}$-norm regularization for phenotypic marker selection and the trace-norm regularization for capturing the interrelationships among different SNPs. We implement two versions of the proposed method as follows. First, we implement our method by only imposing the trace-norm regularization, denoted as "Ours (Trace-norm only)", which only makes use of the SNPs' correlations, but does not select longitudinal imaging markers. Second, we implement the full version of the proposed method, denoted as "Ours", which solves the problem in Eq. (6.4). For measuring the regression performance of the five compared association models, we use a 5-fold cross-validation strategy by computing the Pearson's correlation coefficient (CORR) and the root



(a) VBM.

(b) VBM.

(c) FreeSurfer.

(d) FreeSurfer.

Figure 6.1. Regression performance with respect to the use of different number of longitudinal time points by three different methods..

mean square error (RMSE) between the predicted and the actual SNPs values, which are reported in Fig. 6.1.

As can be seen from Fig. 6.1, if we only use the baseline data, the proposed method is reduced into a conventional multi-task regression model, which appears as a matrix but not a tensor and achieves only the slightly better performance than the RR and Lasso methods. On the other hand, by using the longitudinal data, the performance of the proposed method is significantly improved, *e.g.*, for predicting SNPs using the longitudinal data over all the four time points, the proposed (BL∼M24) method achieves the CORR of 0.793 and 0.812 and the RMSE of 0.314 and 0.301, respectively, which are much better than the case of using only the baseline data.

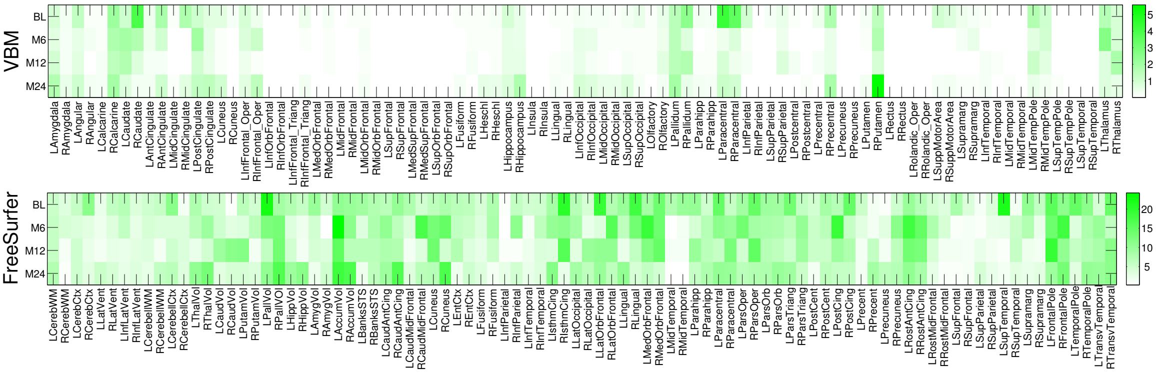


Figure 6.2. Weight maps of the association between imaging markers and the SNPs learned by the proposed method..

In addition, Fig. 6.1 also shows that the usage of longitudinal data can improve the performances of all the LR, RR and Lasso methods, although the improvements are much less than the proposed method.

These results demonstrate the effectiveness of using longitudinal data for improved regression from imaging phenotypes to genotypes, especially by the proposed method, which has the capability to make use of the input data through longitudinal feature selection and the integration of the interrelatedness among the SNPs.

6.4.2 Identification of longitudinal imaging markers

One primary goal of this study is to identify a subset of imaging phenotypes that are highly correlated to certain SNPs to capture important imaging genomic associations in AD research. Thus, we examine the phenotypic imaging markers identified by the proposed methods, which are relevant to the genotypes encoded by SNPs.

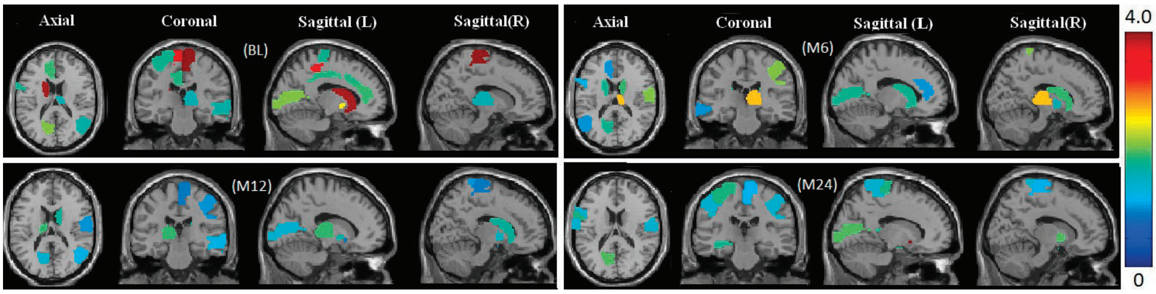


Figure 6.3. Visualization of top 10 VBM features selected by the proposed method at four different time points. The colors of the selected brain regions show the regression coefficients of the corresponding VBM markers..

6.4.2.1 Identified imaging markers with high AD risks.

Shown in Fig. 6.2 are the overall regression coefficients for all the VBM and FreeSurfer measures with respect to the 1224 SNPs used in this study. Because these SNPs are AlzGene candidates or proximal to the candidates, the results in Fig. 6.2 can help identify SNP-relevant imaging phenotypes and have a potential to gain biological insights from gene to brain to symptoms. Besides, the top 10 selected VBM imaging features, as well as their association coefficients, are visualized in Fig. 6.3 by mapping them onto the human brain.

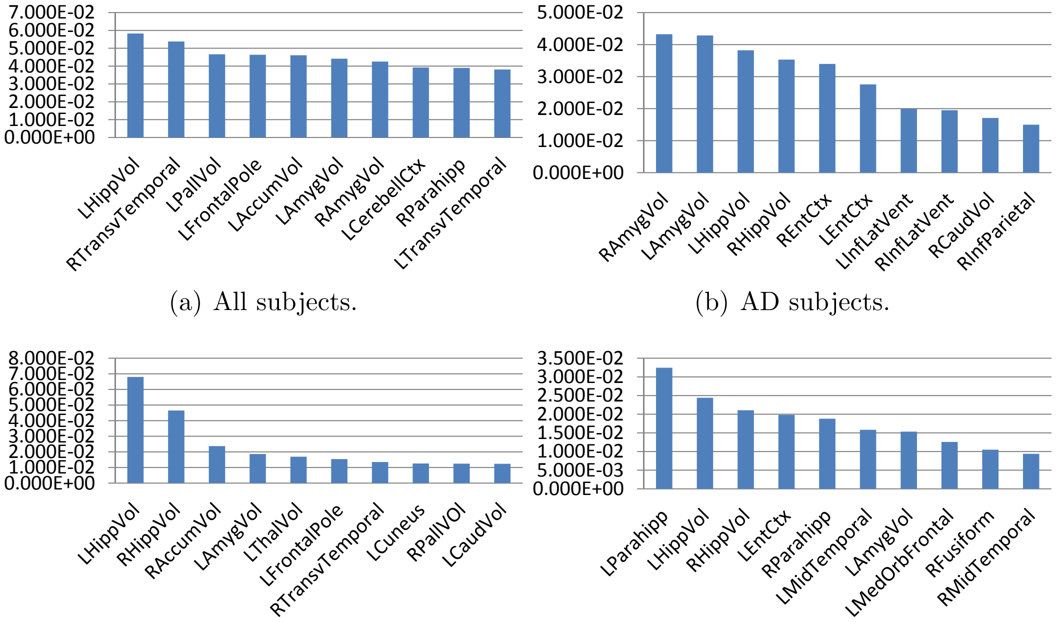
A first glance at the association weigh maps shows that the selected imaging markers have clear patterns that span across all the four studied time points, which demonstrates that these phenotypic markers are longitudinally stable thus can serve as screening target over the course of AD progression. We also observe that hippocampal measures (LHippocampus, RHippocampus, LHippVol and RHippVol) are identified, which is in accordance with the fact that in the pathological pathway of AD, medial temporal lobe including hippocampus is firstly affected, followed by progressive neocortical damage. The thickness measures of isthmus cingulate (LIsthmCing and RIsthmCing), frontal pole (LFrontalPole and RFrontalPole) and posterior cingulate gyrus (LPostCingulate and RPostCingulate) are also selected, which, again, is accordance with the fact that the GM atrophy of these regions is high in AD. In summary, the identified longitudinally stable markers strongly agree with the existing findings, which warrants the correctness of the discovered phenotype-genotype associations, and reveals the complex relationships among MRI measures, genetic variations, and diagnosis status. This is of clear importance for theoretical research and clinical practices for a better understanding of AD mechanism.

6.4.2.2 Case studies: markers identified for rs423958-APOE and rs11136000-CLU.

We provide two case studies to show the top 10 FreeSurfer markers associated with two major AD risk SNPs: rs423958-APOE and rs11136000-CLU. We explore the associations between the FreeSurfer markers and the two SNPs in four different subject groups induced from the ADNI data, *i.e.*, the groups of All, AD, MCI and HC participants respectively. The number of the subjects in each group is available in Table 6.1. We select the imaging markers by their average regression coefficients over all the four time points. The top 10 FreeSurfer markers relevant to rs423958-APOE and their regression coefficients are shown in Fig. 6.4, and those relevant to rs11136000-CLU are shown in Fig. 6.5. From Fig. 6.4 we can see that most of the top 10 FreeSurfer markers for rs423958-APOE in the four different testing groups are well known AD sensitive phenotypes, such as hippocampal volume in All, AD, MCI and HC groups, amygdala volume in All, AD, MCI and HC groups, accumbens volume in All and MCI groups, entorhinal cortex thickness in AD and HC groups. Similar patterns are also observed for rs11136000-CLU, as shown in Fig. 6.5. Although data is not shown due to space limit, our VBM analyses have also yielded similar results. The complete imaging marker identification results by our method for both VBM and FreeSurfer markers on the top 10 identified SNPs are available at the author's website at http://ranger.uta.edu/%7eheng/imgsnp/. These results have again demonstrated the promise of the proposed method in term of its capability to identify imaging markers relevant to AD sensitive SNPs.

6.5 Conclusions

Elucidating the associations between longitudinal phenotypic imaging markers and AD sensitive SNPs is of important value for both scientific research and clinical practice. In this chapter, we presented a new *task-correlated longitudinal sparse regression* method to identify longitudinal imaging markers to AD relevant SNPs. In our newly proposed regression model, we imposed a tensor $\ell_{2,1}$-norm regularization extended from the standard matrix $\ell_{2,1}$-norm to capture the temporal patterns in the longitudinal data over all the tasks of interest, and meanwhile imposed the trace-norm regularization onto the unfolded coefficient tensor such that the interrelatedness among the SNPs during the progression of AD conversion is addressed. Due to the additional time dimension of the input data and the non-smoothness of the tensor $\ell_{2,1}$-norm and trace-norm, solving the formulated objective of our new method was very challenging. Therefore we presented an efficient iterative algorithm and rigorously proved its convergence to the global optimum. We applied the proposed method to the ADNI cohort and evaluated it in both SNPs prediction and longitudinal imaging marker identification. The clearly improved regression performance in the prediction and highly suggestive imaging markers selected by our new method have validated its effectiveness.



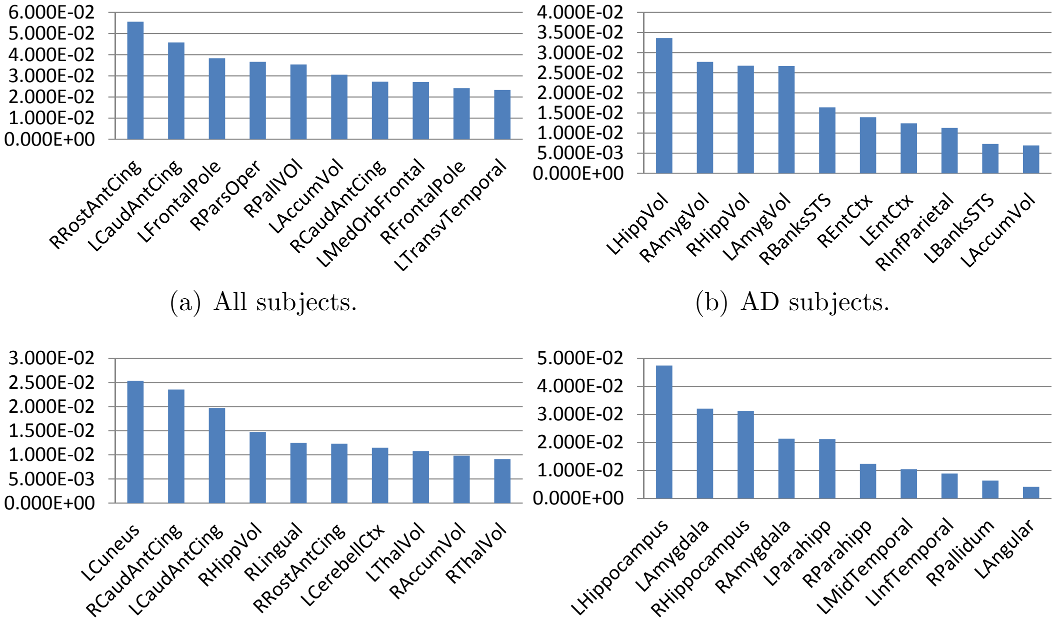
(a) All subjects.

(b) AD subjects.

(c) MCI subjects.

(d) HC subjects.

Figure 6.4. Top 10 FreeSurfer markers identified for rs423958-APOE..



(a) All subjects.

(b) AD subjects.

(c) MCI subjects.

(d) HC subjects.

Figure 6.5. Top 10 FreeSurfer markers identified for rs11136000-CLU..

CHAPTER 7

CONCLUSIONS AND FUTURE WORKS

My research mainly focus on imaging genetics, where our goal is to elucidate the complicated interrelationships among the genotypes, phenotypes, cognitive measurements and diagnosis status on the ADNI platform to help early AD detection. We have built up a unified framework using sparse learning theories, by which we can identify disease relevant biomarkers upon different inputs of the available data. To summarize, from task perspective, we have designed models that can deal with homogeneous tasks for cognitive measure regression and heterogeneous tasks to additionally deal with AD status classification; from data perspective, we have devised models that can deal with both single-modal and multi-modal input data; we also developed model to deal with longitudinal data, which is a new, yet important, direction for imaging genetics.

Recent advances in acquiring multi-modal brain imaging and genome-wide array data have provided exciting new opportunities to study the influence of genetic variation on brain structure and function. Analysis of these multi-modal data sets will facilitate early diagnosis, deepen mechanistic understanding and improved treatment of brain disorders. In the future, I will harness the opportunities of designing principled structured sparse learning and multi-task learning approaches to reveal sophisticated relationships among multi-modal imaging genetic data sets and addressing critical challenges of dimensionality, scalability, diversity, complexity, and heterogeneity in order to realize the full potential of the data. My new multi-modal learning methods will also be applied to solve the emerging multi-dimensional biological data integration, such as the The Cancer Genome Atlas (TCGA) data analysis.

REFERENCES

[1] B. F. Voight, S. Kudaravalli, *et al.*, “A map of recent positive selection in the human genome,” *PLoS Biol*, vol. 4, no. 3, p. e72, 2006.

[2] H. Wang, F. Nie, H. Huang, S. Risacher, A. Saykin, and L. Shen, “Identifying ad-sensitive and cognition-relevant imaging biomarkers via joint classification and regression,” *Proceedinds of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2011)*, pp. 115–123, 2011.

[3] H. Wang, F. Nie, H. Huang, C. Risacher, S.and Ding, A. Saykin, L. Shen, and ADNI, “Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance,” in *Proceedings of the 13th IEEE International Conference on Computer Vision (ICCV 2011)*, 2011, pp. 557–562.

[4] H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. Risacher, A. Saykin, L. Shen, and ADNI, “Identifying Quantitative Trait Loci via Group-Sparse Multi-Task Regression and Feature Selection: An Imaging Genetics Study of the ADNI Cohort,” *Bioinformatics*, 2011 (in press).

[5] L. Shen, Y. Qi, and *et al.*, “Sparse bayesian learning for identifying imaging biomarkers in AD prediction,” *Med Image Comput Comput Assist Interv*, vol. 13, no. Pt 3, pp. 611–8, 2010.

[6] N. Batmanghelich, B. Taskar, and C. Davatzikos, “A general and unifying framework for feature construction, in image-based pattern classification,” *Inf Process Med Imaging*, vol. 21, pp. 423–34, 2009.

[7] C. Hinrichs, V. Singh, *et al.*, “Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset,” *Neuroimage*, vol. 48, no. 1, pp. 138–49, 2009.

[8] K. Walhovd, A. Fjell, and *et al.*, “Multi-modal imaging predicts memory performance in normal aging and cognitive decline,” *Neurobiol Aging*, vol. 31, no. 7, pp. 1107–1121, 2010.

[9] C. M. Stonnington, C. Chu, and *et al.*, “Predicting clinical scores from magnetic resonance scans in alzheimer’s disease,” *Neuroimage*, vol. 51, no. 4, pp. 1405–13, 2010.

[10] M. W. Weiner, P. S. Aisen, and *et al.*, “The alzheimer’s disease neuroimaging initiative: progress report and future plans,” *Alzheimers Dement*, vol. 6, no. 3, pp. 202–11 e7, 2010.

[11] R. Tibshirani, “Regression shrinkage and selection via the LASSO,” *J. Royal. Statist. Soc B.*, vol. 58, pp. 267–288, 1996.

[12] G. Obozinski, B. Taskar, and M. Jordan, “Multi-task feature selection,” *Technical report, Department of Statistics, University of California, Berkeley*, 2006.

[13] A. Argyriou, T. Evgeniou, and M. Pontil, “Multi-task feature learning,” *NIPS*, pp. 41–48, 2007.

[14] F. Nie, H. Huang, X. Cai, and C. Ding, “Efficient and Robust Feature Selection via Joint l2,1-Norms Minimization,” in *NIPS*, 2010.

[15] S. G. Mueller, M. W. Weiner, and *et al.*, “Ways toward an early diagnosis in alzheimer’s disease: The alzheimer’s disease neuroimaging initiative (adni),” *Alzheimers Dement*, vol. 1, no. 1, pp. 55–66, 2005.

[16] L. Shen, S. Kim, and *et al.*, “Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort,” *NeuroImage*, vol. 53, no. 3, pp. 1051 – 1063, 2010.

[17] B. Fischl, D. H. Salat, and *et al.*, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–55, 2002.

[18] P. S. Aisen, R. C. Petersen, and *et al.*, "Clinical core of the alzheimer's disease neuroimaging initiative: progress and plans," *Alzheimers Dement*, vol. 6, no. 3, pp. 239–46, 2010.

[19] M. Ball, V. Hachinski, A. Fox, A. Kirshen, M. Fisman, W. Blume, V. Kral, H. Fox, and H. Merskey, "A new definition of Alzheimer's disease: a hippocampal dementia," *Lancet*, vol. 325, no. 8419, pp. 14–16, 1985.

[20] A. Convit, M. de Leon, C. Tarshish, S. De Santi, A. Kluger, H. Rusinek, and A. George, "Hippocampal volume losses in minimally impaired elderly." *Lancet*, vol. 345, no. 8944, p. 266, 1995.

[21] J. Barnes, R. Scahill, and *et al.*, "Increased hippocampal atrophy rates in AD over 6 months using serial MR imaging," *Neurobiol Aging*, vol. 29, no. 8, pp. 1199–1203, 2008.

[22] S. De Santi, M. de Leon, and *et al.*, "Hippocampal formation glucose metabolism and volume losses in MCI and AD," *Neurobiol of aging*, vol. 22, no. 4, pp. 529–539, 2001.

[23] R. Petersen, C. Jack Jr, and *et al.*, "Memory and MRI-based hippocampal volumes in aging and AD," *Neurology*, vol. 54, no. 3, p. 581, 2000.

[24] C. Van Petten, "Relationship between hippocampal volume and memory ability in healthy individuals across the lifespan: review and meta-analysis," *Neuropsychologia*, vol. 42, no. 10, pp. 1394–1413, 2004.

[25] R. Buckner and D. Carroll, "Self-projection and the brain," *Trends Cogn. Sci.*, vol. 11, no. 2, pp. 49–57, 2007.

[26] D. Hassabis and E. Maguire, "Deconstructing episodic memory with construction," *Trends Cogn. Sci.*, vol. 11, no. 7, pp. 299–306, 2007.

[27] Y. Fan, N. Batmanghelich, C. M. Clark, and C. Davatzikos, "Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline," *Neuroimage*, vol. 39, no. 4, pp. 1731–43, 2008.

[28] M. Moscovitch, L. Nadel, and *et al.*, "The cognitive neuroscience of remote episodic, semantic and spatial memory," *Curr Opin Neurobiol*, vol. 16, no. 2, pp. 179–190, 2006.

[29] S. G. Potkin, J. A. Turner, G. Guffanti, A. Lakatos, F. Torri, D. B. Keator, and F. Macciardi, "Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations," *Cogn Neuropsychiatry*, vol. 14, no. 4, pp. 391–418, 2009.

[30] D. Glahn, P. Thompson, and J. Blangero, "Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function," *Human brain mapping*, vol. 28, no. 6, pp. 488–501, 2007.

[31] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of The Royal Statistical Society Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[32] A. J. Saykin, L. Shen, T. M. Foroud, S. G. Potkin, S. Swaminathan, S. Kim, S. L. Risacher, K. Nho, M. J. Huentelman, D. W. Craig, P. M. Thompson, J. L. Stein, J. H. Moore, L. A. Farrer, R. C. Green, L. Bertram, J. Jack, C. R., M. W. Weiner, and ADNI, "Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans," *Alzheimers Dement*, vol. 6, no. 3, pp. 265–73, 2010.

[33] Y. Li, C. J. Willer, *et al.*, "Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genet Epidemiol*, vol. 34, no. 8, pp. 816–34, 2010.

[34] L. Bertram, M. B. McQueen, *et al.*, "Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database," *Nat Genet*, vol. 39, no. 1, pp. 17–23, 2007.

[35] J. Ashburner and K. Friston, "Voxel-based morphometry–the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[36] K. Puniyani, S. Kim, and E. Xing, "Multi-population GWA mapping via multi-task regularized regression," *Bioinformatics*, vol. 26, no. 12, p. i208, 2010.

[37] S. Lee, J. Zhu, and E. Xing, "Adaptive Multi-Task Lasso: with Application to eQTL Detection," *Advances in neural information processing systems*, 2010.

[38] S. Abney, "Bootstrapping," *Annual Meeting of the Association for Computational Linguistics*, 2002.

[39] U. Brefeld and T. Scheffer, "Co-em support vector learning," *International Conference on Machine Learning*, 2004.

[40] R. Ghani, "Combining labeled and unlabeled data for multi-class text categorization," *International Conference on Machine Learning*, 2002.

[41] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine Learning*, 2000.

[42] S. Bickel and T. Scheffer, "Multi-view clustering," *IEEE International Conference on Data Mining*, 2004.

[43] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[44] C. Hinrichs, V. Singh, G. Xu, and S. Johnson, “Mkl for robust multi-modality ad classification,” in *Proceedings of the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part II*, 2009.

[45] S. Yu, T. Falck, A. Daemen, L. Tranchevent, J. Suykens, B. De Moor, and Y. Moreau, “L 2-norm multiple kernel learning and its application to biomedical data fusion,” *BMC bioinformatics*, vol. 11, no. 1, p. 309, 2010.

[46] J. Suykens, T. Van Gestel, and J. De Brabanter, *Least squares support vector machines*. World Scientific Pub Co Inc, 2002.

[47] M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg, “Non-sparse multiple kernel learning,” in *NIPS*, 2008.

[48] J. Ye, S. Ji, and J. Chen, “Multi-class discriminant kernel learning via convex programming,” *JMLR*, vol. 9, pp. 719–758, 2008.

[49] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, “Learning the kernel matrix with semidefinite programming,” *JMLR*, vol. 5, pp. 27–72, 2004.

[50] F. Bach, G. Lanckriet, and M. Jordan, “Multiple Kernel Learning, Conic Duality, and the SMO Algorithm,” *ICML*, 2004.

[51] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, “Large scale multiple kernel learning,” *JMLR*, vol. 7, pp. 1531–1565, 2006.

[52] A. Zien and C. Ong, “Multiclass multiple kernel learning,” *ICML*, pp. 1191–1198, 2007.

[53] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, “More efficiency in multiple kernel learning,” *ICML*, pp. 775–782, 2007.

[54] C. Micchelli, J. Morales, and M. Pontil, “A Family of Penalty Functions for Structured Sparsity,” *NIPS*, 2010.

[55] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statistics and Computing*, vol. 20, pp. 231–252, 2010.

[56] L. Sun, J. Liu, J. Chen, and J. Ye, "Efficient recovery of jointly sparse vectors," in *Advances in Neural Information Processing Systems 22*, 2009, pp. 1812–1820.

[57] S. Kim and E. Xing, "Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity," *ICML*, 2010.

[58] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[59] J. Liu, J. Chen, and J. Ye, "Large-scale sparse logistic regression," in *SIGKDD09*, 2009, pp. 547–556.

[60] A. Beck and M. Teboulle., "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[61] D. Luo, C. Ding, and H. Huang, "Towards Structural Sparsity: An Explicit l_2/l_0 Approach," *ICDM*, pp. 344–353, 2010.

[62] A. Quattoni, X. Carreras, M. Collins, and T. Darrell, "An Efficient Projection for $l_{1,\infty}$ Regularization," *ICML*, 2009.

[63] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[64] X. Yang, S. Kim, and E. P. Xing, "Heterogeneous multitask learning with joint sparsity constraints," *NIPS*, 2009.

[65] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: fast algorithms and generalization bounds," *IEEE Trans Pattern Anal Mach Intell.*, vol. 27, no. 6, pp. 957–68, 2005.

[66] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient l1 regularized logistic regression," in *The 21st National Conference on Artificial Intelligence (AAAI)*, 2006.

[67] S. Landau, D. Harvey, C. Madison, R. Koeppe, E. Reiman, N. Foster, M. Weiner, and W. Jagust, "Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI," *Neurobiol Aging Epub*, 2009.

[68] A. Hariri, E. Drabant, and D. Weinberger, "Imaging genetics: perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing," *Biological Psychiatry*, vol. 59, no. 10, pp. 888–897, 2006.

[69] C. Brun, N. Leporé, and *et al.*, "Mapping the regional influence of genetics on brain structure variability–a tensor-based morphometry study," *Neuroimage*, vol. 48, no. 1, pp. 37–49, 2009.

[70] N. Filippini, A. Rao, and *et al.*, "Anatomically-distinct genetic associations of APOE $\varepsilon 4$ allele load with regional cortical atrophy in Alzheimer's disease," *NeuroImage*, vol. 44, pp. 724–728, 2009.

[71] T. Nichols and B. Inkster, "Comparison of Whole Brain Multiloci Association Methods," *NeuroImage*, vol. 47, p. S161, 2009.

[72] S. Seshadri, A. DeStefano, and *et al.*, "Genetic correlates of brain aging on MRI and cognitive test measures: a genome-wide association and linkage analysis in the Framingham Study," *BMC Medical Genetics*, vol. 8, no. Suppl 1, p. S15, 2007.

[73] S. Baranzini, J. Wang, and *et al.*, "Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis," *Human molecular genetics*, 2008.

[74] S. L. Risacher, L. Shen, J. D. West, S. Kim, B. C. McDonald, L. A. Beckett, D. J. Harvey, J. Jack, C. R., M. W. Weiner, A. J. Saykin, and ADNI, "Longitu-

dinal MRI atrophy biomarkers: relationship to conversion in the ADNI cohort," *Neurobiol Aging*, vol. 31, no. 8, pp. 1401–18, 2010.

[75] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[76] B. Fischl, M. Sereno, and A. Dale, "Cortical surface-based analysis. ii: Inflation, flattening, and a surface-based coordinate system." *Neuroimage*, vol. 9, no. 2, pp. 195–207, 1999.

BIOGRAPHICAL STATEMENT

Hua Wang is currently a Ph.D. candidate in the Computer Science and Engineering Department, University of Texas at Arlington, under the supervision of Dr. Heng Huang. He received a B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China in 1999 and a M.S. degree in Signal Processing from Nanyang Technological University, Singapore in 2003. Before he went to UTA in 2007, he worked in Motorola as an embedded software engineer for four years. Hua's current research areas are machine learning and data mining, as well as their applications in bioinformatics, medical image analysis, and computer vision.