MINING BRAIN IMAGING AND GENETICS DATA VIA STRUCTURED SPARSE LEARNING

Jingwen Yan

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

October 2015

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Huanmei Wu, PhD, Chair

Doctoral Committee

Li Shen, PhD

April 29, 2015

Shiaofen Fang, PhD

Xiaowen Liu, PhD

DEDICATION

Dedicated to my mother and husband

for all the love and support along the way

ACKNOWLEDGEMENTS

First and foremost I would like to express my deepest gratitude to my advisor, Dr. Li Shen, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. It was him who brought me into the computational neuroscience field and guided me through all the barriers that came across my research. He has provided me with great research insights and exceptional enthusiasm as well as encouragement throughout my PhD study. This work would never be materialized without him.

I also give my thanks to all the committee members for supporting this thesis work and all the constructive suggestions and feedback, that have been very helpful to keep this work on track and finally make the goal accomplished timely.

I would like to express my appreciation to all my colleagues and professors from Center for Neuroimaging: Dr. Andrew J. Saykin, Dr. Kim Sungeun, Dr. Shannon L. Risacher, Dr. Kwangsik Nho and many others, who have provided me very valuable domain expertise from neurological, biological and genetic perspectives, as well as many invaluable medical data sources. I have learned considerably from their multi-perspective insights into problems. Also I am very thankful to my collaborators: Prof. Heng Huang from University of Texas at Arlington and Hui Zhang from Indiana University Pervasive Technology Institute, for many valuable discussions on algorithms and big data frameworks.

This work has been supported by an NSF grant (IIS-1117335) for "III:Small: Collaborative Research: A Large-scale Data Mining Framework for Genome-wide Mapping of Multi-modal Phenotypic Biomarkers and Outcome Prediction", and also

in part by an NIH grant (R01 LM011360) for "Bioinformatics Strategies for Multidimensional Brain Imaging Genetics".

Finally I would like to express my special thanks and appreciation to my family. My mom, Yuhong Jing, is always supporting me and encouraging me with her best capabilities and wishes. I would like to conclude by extending my appreciation to my husband, Jing Liu, who was always there cheering me up and stood by me through the good times and bad. And things are always getting better with him along the side.

Jingwen Yan

MINING BRAIN IMAGING AND GENETICS DATA VIA STRUCTURED SPARSE LEARNING

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by gradual loss of brain functions, usually preceded by memory impairments. It has been widely affecting aging Americans over 65 old and listed as 6th leading cause of death. More importantly, unlike other diseases, loss of brain function in AD progression usually leads to the significant decline in self-care abilities. And this will undoubtedly exert a lot of pressure on family members, friends, communities and the whole society due to the time-consuming daily care and high health care expenditures. In the past decade, while deaths attributed to the number one cause, heart disease, has decreased 16 percent, deaths attributed to AD has increased 68 percent. And all of these situations will continue to deteriorate as the population ages during the next several decades.

To prevent such health care crisis, substantial efforts have been made to help cure, slow or stop the progression of the disease. The massive data generated through these efforts, like multimodal neuroimaging scans as well as next generation sequences, provides unprecedented opportunities for researchers to look into the deep side of the disease, with more confidence and precision. While plenty of efforts have been made to pull in those existing machine learning and statistical models, the correlated structure and high dimensionality of imaging and genetics data are generally ignored or avoided through targeted analysis. Therefore their performances on imaging genetics study are quite limited and still have plenty to be improved.

The primary contribution of this work lies in the development of novel prior knowledge-guided regression and association models, and their applications in various neurobiological problems, such as identification of cognitive performance related imaging biomarkers and imaging genetics associations. In summary, this work has achieved the following research goals: (1) Explore the multimodal imaging biomarkers toward various cognitive functions using group-guided learning algorithms, (2) Development and application of novel network structure guided sparse regression model, (3) Development and application of novel network structure guided sparse multivariate association model, and (4) Promotion of the computation efficiency through parallelization strategies.

Huanmei Wu, PhD, Chair

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1

1.1 Data mining in Alzheimer research 1

1.2 Contribution 2

1.2.1 Biomarker discovery 3

1.2.2 Genetic mechanism exploration 4

1.2.3 Data intensive computing 5

1.3 Organization 6

CHAPTER 2 PREVIOUS WORK 8

2.1 Biomarker discovery 8

2.2 Genetic mechanism exploration 14

CHAPTER 3 LOCALIZED CORTICAL BIOMARKERS FOR PREDICTING COGNITIVE OUTCOMES USING GROUP $\ell_{2,1}$-NORM 19

3.1 Background 19

3.2 Materials and Methods 21

3.2.1 Neuroimaging and Cognition Data 21

3.2.2 G-SMuRFS 23

3.3 Experimental Results and Discussion 24

3.3.1 Experimental Setting 24

3.3.2 Results and Discussion 25

3.4 Conclusion 31

CHAPTER 4 JOINT IDENTIFICATION OF IMAGING AND PROTEOMICS BIOMARKERS USING NG-L21 35

4.1 Background 35
4.2 Methods 36
4.3 Results 40
4.3.1 Data and Experimental Setting 40
4.3.2 Experimental Results 41
4.4 Conclusion 43
CHAPTER 5 STRUCTURE-AWARE SCCA IN IMAGING GENETICS ASSOCIATION ANALYSIS 45
5.1 Background 45
5.2 Structure-aware SCCA (S2CCA) 46
5.3 Experimental Results 49
5.3.1 Results on Simulation Data 49
5.3.2 Results on Real Neuroimaging Genetics Data 51
5.4 Conclusion 53
CHAPTER 6 TRANSCRIPTOME-GUIDED AMYLOID IMAGING GENETICS VIA KG-SCCA 55
6.1 Materials and Data Sources 56
6.1.1 Imaging and Genotyping Data 57
6.1.2 Amyloid Pathway-based Gene Co-expression Network in the Brain 58
6.2 Methods 61
6.3 Experimental Results and Discussions 66
6.3.1 Results on Simulation Data 67
6.3.2 Results on Real Imaging Genetic Data 69

6.4 Conclusion 71

CHAPTER 7 DATA INTENSIVE COMPUTING IN BRAIN IMAGING GENETICS STUDY 73

7.1 SCCA for Imaging Genetics 74

7.2 Accelerating SCCA at XSEDE 75

7.2.1 Available Acceleration Strategies 77

7.2.2 Accelerating SCCA with MKL and Offload Model 78

7.3 SCCA for Large Brain Imaging Genetics Data 80

7.3.1 Simulated Imaging Genetics Data 81

7.3.2 Scaling to Large Datasets with Data Parallel Strategy 83

7.4 Conclusion 87

CHAPTER 8 CONCLUSIONS 89

8.1 Summary 89

8.2 Future directions of research 91

REFERENCES 93

CURRICULUM VITAE

LIST OF TABLES

3.1 List of surface thickness measures 22

3.2 Participant characteristics 23

3.3 Cross validation performance comparison 32

4.1 Participant characteristics 41

4.2 RAVLT scores. 41

4.3 Average correlation coefficient between predicted and actual scores . . . 42

5.1 Five-fold cross-validation performance on synthetic data 50

5.2 Participant characteristics. 52

5.3 Five-fold cross validation results on real data 53

6.1 Participant characteristics. 56

6.2 Five-fold cross-validation performance on synthetic data 66

6.3 Five-fold cross validation results on real data 69

7.1 Translation of benchmark number to R-25 benchmark description for all R-25 plots. 79

7.2 A set of 4 × 3 experiments running SCCA analytics over combinations of 4 genotype data sets and 3 phenotype data sets 86

7.3 The aforementioned 4 × 3 experiments running improved SCCA analytics with data parallel approach 86

LIST OF FIGURES

1.1 Overview of the dissertation work 6

1.2 Organization of this thesis 7

3.1 Scatter plots of actual and predicted (by G-SMuRFS) cognitive scores . . 26

3.2 Histogram of regression weights of all cortical measures 27

3.3 Number of "high impact" cortical markers 28

3.4 Example G-SMuRFS regression weights color-coded on the cortical surface 29

3.5 Example SurfStat t-statistic color-coded and mapped onto the cortical surface. 30

4.1 NG-L21 schematic 37

4.2 Heat maps of regression weights 44

5.1 Five-fold trained weights of $\mathbf{u}$ and $\mathbf{v}$ 51

5.2 Comparison between S2CCA and PMD on identified canonical vectors . 54

6.1 Amyloid pathway-based gene co-expression networks 59

6.2 Co-expression Network 60

6.3 Five-fold trained weights of $\mathbf{u}$ and $\mathbf{v}$ on synthetic data 67

6.4 Five-fold trained weights of $\mathbf{u}$ and $\mathbf{v}$ on real data 70

6.5 Mapping canonical loading generated by KG-SCCA onto the brain. . . . 71

7.1 Adopting offload model on Stampede cluster at XSEDE 76

7.2 Basic vectorized and matrixed operations can be obtained significant speed-up with MLK and offload on MIC. 79

7.3 Correlation structure of the simulated genotype data 82

7.4 Example image data 83

7.5 Comparison of SCCA speed-up for different datasets combinations 84

7.7 Ground truth and identified associations 88

Chapter 1

INTRODUCTION

1.1 DATA MINING IN ALZHEIMER RESEARCH

As one of the most common brain dementia, Alzheimer's disease (AD) is a neurodegenerative disorder characterized by gradual loss of brain functions, usually preceded by memory impairments. Based on the latest report [60], AD has been widely affecting aging Americans over 65 years old and has been officially listed as the 6th leading cause of death. Due to the significant decline of self-care abilities during disease, it is not only the patients who suffer, but also the family members, friends, communities and the whole society due to time-consuming daily care and high health care expenditures needed. In the past decade, while deaths attributed to the number one cause, heart disease, has decreased 16 percent, deaths attributed to Alzheimer's disease has increased 68 percent. And all of these situations will continue to deteriorate as the population ages during the next several decades. In pursuit of prevention of such health care crisis, substantial efforts have been made to help cure, slow or stop the progression of the disease. And the massive data generated through these efforts, like multimodal brain imaging scans as well as genome sequences, provides us an unprecedented opportunity to look into the deep side of the disease, with more confidence and precision.

Recently, machine learning and statistical methods have been vastly pulled into AD research area to help advance the pattern mining out of the data ocean and to further facilitate the understanding of the underlying disease mechanism. The majority of the applications fall into three categories: (1) Disease status classification,

(2) Predictive analysis of disease progression, and (3) Associative analysis to explore the genetic mechanism of various quantitative traits (QTs). While substantial efforts have been made to directly apply the existing models, there are certain limitations worth our attention. Since all those models are not originally designed for disease study, direct application may be problematic due to the special correlation structure of data sets as well as the skyrocketing dimensionality. Detailed review of previous work can be found in Chapter 2. In this thesis, we particularly focus on the development of novel computational models that best fit the needs of disease study and their applications in prediction analysis as well as association analysis.

1.2 CONTRIBUTION

In this thesis, we harness the opportunities of developing knowledge guided sparse learning frameworks to take the best advantage of the massive sophisticated data sets to realize their full potential and address the challenges of dimensionality, scalability, diversity, heterogeneity and ultimate pattern interpretability. With the growing multimodal brain imaging, genetics, proteomics, and clinical outcome data in AD research area [75], exploring effective imaging biomarkers for various cognitive function changes and further examining their underlying genetic variations through imaging genetics study would be a promising test bed for exploration, application and validation of the proposed methods. Continuous phenotypic measures from imaging data, fluid biomarkers, and cognitive status have great potential to serve as useful intermediate traits on the chain of causality from genes to symptoms. Upon these datasets, this work is dedicated to developing efficient and effective strategies for integrative analysis of high dimensional heterogeneous imaging, multi-omics and clinical

data. More specifically, the proposed methods are designed, developed, accelerated and finally applied to solve the problems in the following two areas: (1) Biomarker discovery, and (2) Genetic mechanism exploration (Fig. 1.1).

1.2.1 BIOMARKER DISCOVERY

The first part of this thesis focuses on developing novel mathematical models to examine the predictive power of multi-resolution and multimodal measures, and to explore the role of prior knowledge in improving the predictive performances.

Multi-resolution measures: When performing a specific task such as 'moving fingers', only a few brain regions get activated. If so, averaged measures of brain regions, widely used in existing predictive studies, will possibly smooth out those activation signals. Similarly as undesirable is the huge computation cost of high resolution voxel-level measures. Therefore, we introduced the intermediate level measures, where each brain region was further partitioned into small patches using clustering algorithms. This strategy was applied to cortical thickness measures, which were further used to predict the memory performances through a new regression model. With the assumption that only a few brain regions participate in the memory specific tasks, spatial information of patches was incorporated to limit the number of selected brain regions. This work reveals more localized brain markers and their improved performance in prediction analysis.

Multimodal measures: Brain imaging measures are now of great diversity from multiple perspectives: functional and structural changes in the brain, cerebrospinal fluid, blood plasma, etc. While prior work mostly focused on examining them individually for potential AD markers, it is of more interest to see whether these multimodal

datasets could complement each other and offer improved prediction power. We examined several multimodal measures from the existing database: cortical thickness, volume, gray matter density, and CSF proteomic measures. The combination of the cortical thickness, volume and CSF proteomic data was found to have the best prediction performance against all the other combinations, indicating possible complementary information between imaging and proteomic measures and suggesting the necessity of further efforts in multimodal studies.

Prior knowledge in learning models: Prior spatial information, normally modeled as group structure, has already been well studied. In this work, we exploited other prior correlation structures within brain that take the form of networks, by proposing a new sparse learning model, network guided $\ell_{2,1}$ norm (NG-L21). Under this framework, two prior correlation structures were examined regarding their power to improve the prediction performance of multimodal imaging phenotypes toward memory performances: (1) correlation structure based on imaging measures, and (2) correlation structure based on gene co-expression profiles. These work confirmed the importance of prior knowledge in leading the model to yield better prediction performance and more meaningful biomarker patterns.

1.2.2 GENETIC MECHANISM EXPLORATION

While genotype data, such as single nucleotide polymorphisms (SNPs), store the most deep-inside signals, multimodal brain scans, blood plasma, and CSF measures capture phenotypic outcomes that are respectively and partially explained by relevant genetic variations. But how the underlying genetic variations (such as SNPs) affect the external phenotypes remains a mystery.

Genome wide association study (GWAS) of quantitative-traits (QTs) is widely applied by exhaustively examining the correlation between all possible pairs between genotypic and phenotypic measures, without embracing those complicated multiple-SNP-multiple-QT relationships. To address this issue, we proposed two models: Structure-aware Sparse Canonical Correlation Analysis (S2CCA) and Knowledge-Guided Sparse Canonical Correlation Analysis (KG-SCCA), together with a new efficient iterative algorithm. Both of them were validated against other existing models based on synthetic datasets. KG-SCCA was further applied to relate the apolipoprotein E (*APOE*, a major AD risk gene) SNPs with amyloid deposition in brain regions, and yielded promising results. This is the first work reporting the localized amyloid deposition related to *APOE* SNPs.

1.2.3 DATA INTENSIVE COMPUTING

High dimensionality of the raw imaging and genomics data has greatly held back the application of machine learning methods. Designing highly scalable algorithms to take these thousands of millions data is not easy and requires long-term effort to accomplish. Fortunately big data science opens another door for vast scientists struggling with explosive data. Our initial efforts to promote the usage of these new models into large datasets application by taking advantage of the parallelization strategies have been made by coupling Math Kernel Library (MKL) with Xeon Phi co-processor on Stampede supercomputer in Texas Advanced Computing Center. For serial tasks, this framework consistently provides at least two-fold speedups without any code change, which sheds light on the potential brain-wide and genome-wide applications.

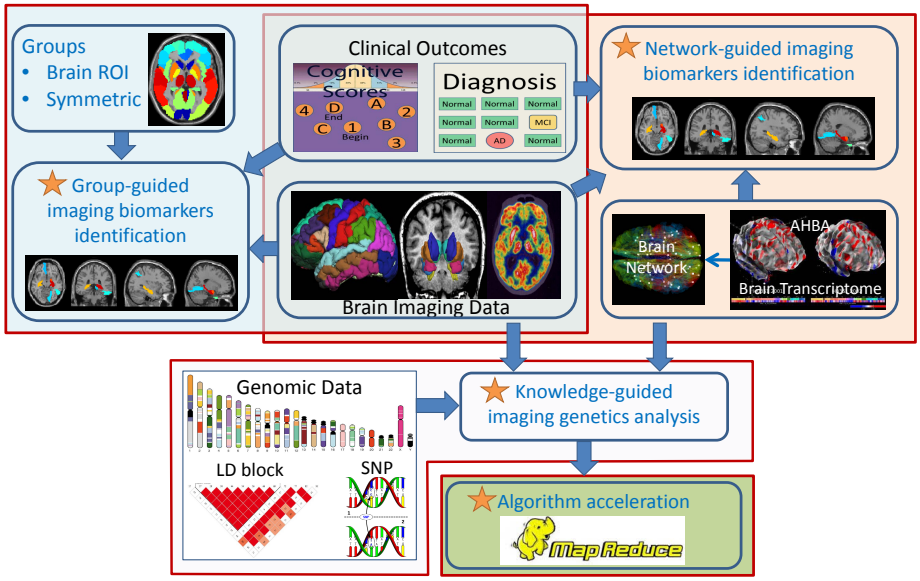


Figure 1.1: Overview of the dissertation work

1.3 ORGANIZATION

The chapters of the thesis are organized as shown in Fig. 1.2. Chapter 1 gives an brief overview of machine learning in Alzheimer's disease study and a summary of contributions. In Chapter 2, we will review those existing data mining techniques, their applications and limitations in Alzheimer research. Chapter 3 and Chapter 4 are for disease biomarker discovery, in which we will evaluate G-SMuRFS through a multi-resolution study and further propose a network guided multivariate predictive model to explore multimodal imaging biomarkers. In Chapter 5 and Chapter 6, we will discuss the association studies and propose two novel structure-guided sparse canonical correlation analysis models, based on which we explore the complex multi-phenotype-multi-genotype relationships. Chapter 7 describes our initial efforts in data intensive computing. In Chapter 8, we will summarize our work and discuss some potential future directions.

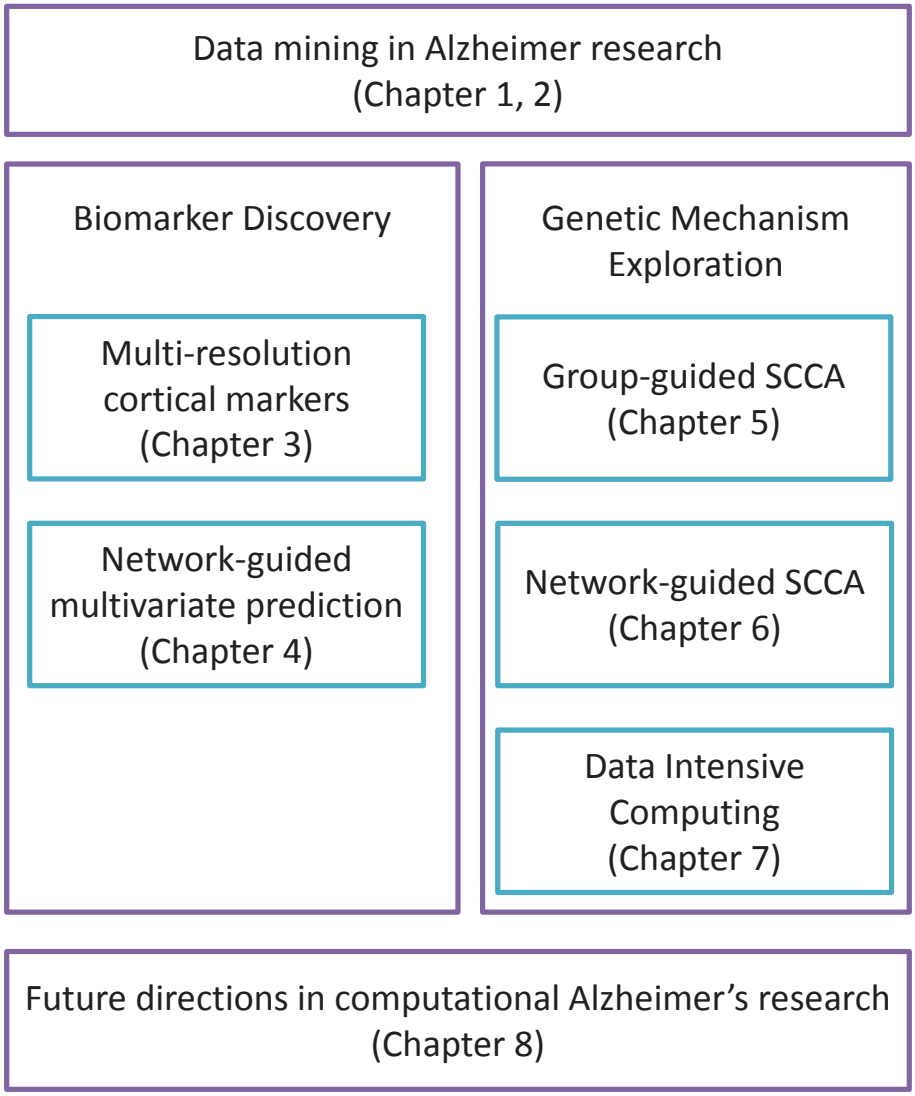


Figure 1.2: Organization of this thesis

Chapter 2

PREVIOUS WORK

In this chapter we will review existing machine learning models and further discuss their applications as well as limitations in disease study. Throughout this section, 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 the $\ell_{2,1}$-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. Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subseteq \Re^d$ be imaging measures and $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\} \subseteq \Re^c$ be cognitive outcomes (regression analysis) or genetic variation measurements (association analysis), where n is the number of samples, d is the number of $\mathbf{X}$ features (feature dimensionality) and c is the number of $\mathbf{Y}$ features (tasks or variations). Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$.

2.1 BIOMARKER DISCOVERY

With the advances in neuroimaging and genetic sequencing technology, imaging and genetics data have explosively grown in the last decades, including structural magnetic resonance imaging (MRI), functional MRI, positron emission tomography (PET), etc. And these developments are of great significance for enabling multi-perspective view of brain structures and activities. First accompanied are the predictive analyses that mostly utilize the multiple modal imaging or proteomic measures as predictors to classify the disease status, mostly between healthy controls and AD patients [8, 17, 26, 52, 85].

However, it is argued later that this simple grouping strategy may not reveal the true stratification [51]. Due to the dynamic longitudinal process of AD, health control subjects may convert to mild cognitive impairment (MCI, a prodromal stage of AD) beyond the study follow up. Thus including these marginal subjects into the a specific group will more or less smooth out the features and limit the ultimate predictive power towards various tasks. On the contrary, continuous imaging measurements and cognitive outcomes keep a better track of the disease progressing procedure, and therefore may provide a more informative insight of the underlying mechanism.

Based on that, enormous efforts have been made recently to evaluate the predictive power of sparse learning methods in the neuroimaging field. Earlier studies preferred regression analysis focused on predicting selected cognitive scores one at a time using statistical learning approaches such as stepwise regression [66] and relevance vector regression [58]. Yet for one patient there usually are many cognitive scores grouped by different categories. Scores falling into the same category, like episodic memory, mostly correlate with each other, and ignoring this relationship will inevitably lead to suboptimal results. Also considering multiple scores together with their correlation structures will to some extent help with the noise suppression.

In [69], Wan et al. examined the effect of brain volume and thickness measures on cognitive changes measured by three different sets of scores using sparse Bayesian learning, which implicitly considered the correlation among brain imaging measures. And their recent work extended this model to consider the non-linearity correlation [68]. Other similar studies include [72, 73]. Besides structural imaging, functional imaging is also quite popular as potential predictive measures. Unlike structural imaging, functional imaging focuses on capturing physiological activities by employing

medical image modalities that very often use tracers or probes to reflect the real intensity of certain chemical compounds, like amyloid plaques (one of the hallmarks of AD). Example studies include [34] using functional imaging measures of visual tasks to predict the object sizes through incorporating the spatial information. In [27], Grosenick et al. developed a novel model based on elastic net and graphnet [86] to estimate the predictive power of functional imaging toward the behaviors, with both spatial and temporal correlation considered. Other similar studies include [42, 63].

In the following we briefly go through some typical regression models that have been developed or applied in previous studies, and discuss their prediction capabilities as well as limitations. To investigate the correlation between imaging measures and cognitive outcomes, linear and ridge regression models are two standard methods. Linear regression (Eq. (2.1)), also known as least square fitting, seeks to find a straight line through all data points in such a way with the sum of squared residuals as small as possible. Despite its easy and fast implementation, the inflation and general instability, when features outnumber observations, become a big concern with application to AD research. Millions of imaging voxels as well as their highly correlated structure will inevitably lead to the overfitting problems and very unstable patterns [30, 36].

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2. \tag{2.1}$$

To address this issue, ridge regression [30] was proposed to perform extra shrinkage where large coefficients will be penalized. One more regularization term, the Frobenius norm of trained weights, is applied to achieve that goal and helps ascertain

the numerical stability simultaneously (Eq. (2.2)).

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma||\mathbf{W}||_2^2 \ . \tag{2.2}$$

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

Yet regression weights brought by the Frobenius norm are typically non-sparse, which makes the results hard to interpret and unsuitable for biomarker discovery. To produce sparse solutions, the following traditional Lasso model (Eq. (2.3)) [41] can be used:

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma||\mathbf{W}||_1 \ . \tag{2.3}$$

[22] proposed least angle regression selection (LARS) and its solution paths indicate that the lasso are piecewise linear, which gives the lasso tremendous computational advantages when compared with other methods. However, it is mostly considered to make selections based on the strength of individual features, rather than the strength of general groups of features. Due to its intrinsic sparsity property, only one out of the many correlated subset features will be selected (mostly at random) and reparameterization will lead to a different set of features selected. This unstable pattern identification is quite undesirable, which makes the final results confusing for interpretation. Group lasso [82] was one of the two approaches that were proposed to address this concern. Unlike traditional Lasso model, group lasso applies ℓ_2-norm to features within each group (as a prior) to ascertain within-group similarities, accompanied by the ℓ_1-norm imposed across the groups to achieve group sparsity

(Eq. (2.4)). In particular, its performance in application to small sample size with large input variables is also guaranteed [82], which is extremely desirable in disease research.

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F + \gamma \sum_{i=1}^{k} \sqrt{\sum_{j \in G(i)} w_j^2}. \tag{2.4}$$

Another well-known approach is elastic net (Eq. (2.5)) [86], in which less sparse pattern was pursued implicitly through seeking a tradeoff between group and individual sparsity. ℓ_1-norm and ℓ_2-norm of Lasso and Ridge regression are linearly combined to form the new regularized regularization term, with attempt to find a balance point in between, where the highly correlated features will be taken out together and global sparsity is still guaranteed. With this balancing strategy, it has been proved that elastic net can keep consistent prediction performance as well as stable feature selection ability even when the number of features increases [19].

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1||\mathbf{W}||_1 \quad + \gamma_2||\mathbf{W}||_2. \tag{2.5}$$

However, it is worth noticing that, in all the above methods, multi-task learning is equivalent to applying them to each outcome variable independently. The correlation and colinearity among the outcome variables are generally not taken into account. As a result, despite the fact that the outcome variables are highly correlated, the features selected by the above Lasso based models could only be relevant to some outcomes (i.e., regression weights $\neq 0$) but not to others (i.e., regression weights $= 0$).

$\ell_{2,1}$-norm (Eq. (2.6)), motivated by ridge and Lasso, is a multivariate regression model proposed as follows.

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma||\mathbf{W}||_{2,1}. \tag{2.6}$$

where $||\mathbf{W}||_{2,1} = \sum_{i=1}^{d} ||\mathbf{w}^i||_2$. $\ell_{2,1}$-norm is one of the advanced techniques that address both the outcome correlation and sparsity issues, by enforcing the ℓ_2-norm across the tasks and the ℓ_1-norm across the features. While the ℓ_2-norm ascertains the similarity pattern across the correlated tasks, the ℓ_1-norm guarantees the sparsity across the features. Although early studies proposed $\ell_{2,0}$-norm [41] to achieve similar function, $\ell_{2,1}$-norm has global solution with faster convergency [43] and also yields comparable results as the $\ell_{2,0}$-norm. Thus, in this work we will focus on the $\ell_{2,1}$-norm based sparse representations.

In many cases, various brain structures may be responsible for different functions. Therefore it would be much more meaningful to include the structural information in the regression models. Also direct application of conventional sparse models such as $\ell_{2,1}$-norm are more likely to yield scattered patterns, due to the lack of proper handling of spatial correlation and prior anatomical knowledge. With the same motivation as group lasso and elastic net, a recent research introduced an extension of $\ell_{2,1}$-norm to incorporate the prior group structures of input variables. This new extension, G-SMuRFS (Group-Sparse Multi-task Regression and Feature Selection) [71] takes into account the group information in the regression procedure. The new regularization term was applied in G-SMuRFS to consider both the group sparsity through the $\mathbf{G}_{2,1}$-norm and the individual biomarker sparsity for joint learning via an $\ell_{2,1}$-norm regularization [46]. In the objective function Eq. (2.7), the second term helps to couple all the regression coefficients of grouped features across all tasks together and

the third term penalizes all regression coefficients of each individual feature as a whole to select features shared by multiple learning tasks.

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F + \gamma_1||\mathbf{W}||_{G_{2,1}} + \gamma_2||\mathbf{W}||_{2,1} \ . \tag{2.7}$$

where $||\mathbf{W}||_{G_{2,1}} = \sum_{i=1}^{k} \sqrt{\sum_{j \in G(i)} ||\mathbf{w}^j||_2^2}$ is the $\mathbf{G}_{2,1}$-norm.

Since G-SMuRFS has demonstrated very promising performances in a previous study [71] using genotypes to predict AD related imaging measures, in Chapter 3 we will examine its power through a multi-resolution imaging study and compare its performance with existing models. In Chapter 4, we will extend our efforts to develop a more general sparse regression model NG-L21, with prior knowledge incorporated in the network format, and then G-SMuRFS can be viewed as a special case of NG-L21. This new model will be evaluated though examining the prediction power of multimodal measures toward various correlated cognitive outcomes.

2.2 GENETIC MECHANISM EXPLORATION

While biomarker discovery depends on regression analysis and usually attempts to explore the predictive biomarkers toward a specific task/task set, genetic mechanism exploration is usually a more symmetric strategy, where both datasets are treated equally. It alternatively seeks the sophisticated multiple-to-multiple relationship rather than multiple-to-one in regression analysis. One example is brain imaging genetics study, an emerging field in pursuit of system level explanation of the disease through integrating imaging and genetics data to reveal the effect of genetics variations on diverse imaging phenotypes. Such analysis can not only discover AD

biomarkers to help with the early diagnosis, but also hold great promise to unveil the disease-modifying genetic mechanisms.

Early imaging genetics research examined the pair-wise correlation and treated all phenotype and genotype features separately [55]. Recently more complicated association models have been gradually introduced into the imaging genetics field [14], particularly the sparse canonical correlation analysis (SCCA), which is originally derived from traditional canonical correlation analysis (CCA) framework. It is a bi-multivariate model that was designed to find the linear transformation of imaging and genetic data, which can yield the highest correlation coefficients between transformed vectors. Compared to CCA, SCCA usually has an extra Lasso penalty term for both canonical loadings of imaging and genetic features, and can solve the non-sparsity problem and the curse of high-dimensional features [76].

However, the highly correlated brain imaging and genetic data imposes great challenges to association study, just as it does to regression analysis. It has become an essential concern to enable the extraction of grouped features, rather than random selection out of them. Many advanced learning strategies have been dedicated to address this issue, like joint multi-task [72, 73, 84, 85] and structured sparse learning [69], but most of them still ended up with over simplified structures. While the group structure, spatial and temporal correlation have been proved to be important prior knowledge, it is also worth noticing that human brain is a very complicated organ, with regions wiring together to perform certain functions. With the emerging interest in human connectome, it will be of more interest to examine the effect of multiple brain networks on the AD biomarker discovery and imaging genetics associations. Another issue exists that most solutions of SCCA-based algorithms assume

the independence over all features [38, 40], which will make their covariance matrix an identity. Obviously, this assumption could hardly hold for imaging and genetic data, due to the regional structure of brain and the linkage disequilibrium (LD) block structure in the genome. Therefore new solutions are needed before they could be applied to the imaging genetics study.

Another limitation worth noticing is the limited scalability shared by existing methods. Due to rapid advances of the data generation technology in both imaging and sequencing areas, researchers are experiencing a data explosion like never before, with millions of voxels, trillions of gene variations and more on their way. Accompanied with the escalating data volume is the unprecedented computational challenge. Besides the extremely expensive computational cost, the highly inter-correlated structures among features should as well deserve our attention.

Similarly as in regression analysis, many advanced association models are derived from very basic concepts. In the following we briefly go through those typical association models and discuss their capabilities as well as limitations.

CCA is a classical bi-multivariate method that has been applied to imaging genetics fields. Compared to regression analysis, CCA provides a more "symmetric" solution by seeking linear combinations of variables in $\mathbf{X}$ and variables in $\mathbf{Y}$ so that $\mathbf{Xu}$ and $\mathbf{Yv}$ can achieve maximal correlation, and can be formulated as:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \quad s.t. \;\; \mathbf{u}^T\mathbf{X}^T\mathbf{Xu} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Yv} = 1 \tag{2.8}$$

Here u and v are canonical loadings or weights of imaging and genetic measures respectively. However, CCA requires the number of observations n to exceed the

combined dimension of $\mathbf{X}$ and $\mathbf{Y}$ variables. And like many machine learning algorithms, overfitting problems could arise in CCA when the features outnumber the participants. In addition, CCA outcome usually spread nontrivial effects across all the features, due to the lack of proper regularization, rather than a few significant ones and makes the final pattern difficult to interpret.

To address these issues, SCCA was proposed in [77] by introducing penalty terms, $P_1(\mathbf{u}) \leq c_1$ and $P_2(\mathbf{v}) \leq c_2$, to regularize the weights, as shown in (Eq. (2.9)).

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \\ &s.t. \quad ||\mathbf{X}\mathbf{u}||_2^2 = 1, ||\mathbf{Y}\mathbf{v}||_2^2 = 1, P_1(\mathbf{u}) \leq c_1, P_2(\mathbf{v}) \leq c_2 \end{aligned} \tag{2.9}$$

Here the objective function is bi-linear in $\mathbf{u}$ and $\mathbf{v}$: when $\mathbf{u}$ is fixed, it is linear in $\mathbf{v}$ and vice versa. But due to the ℓ_2 equality, with $\mathbf{u}$ or $\mathbf{v}$ fixed, the constraints are not convex. This can be solved by reformulating the ℓ_2 equality into inequality as $||\mathbf{X}\mathbf{u}||_2^2 \leq 1$ and $||\mathbf{Y}\mathbf{v}||_2^2 \leq 1$. For easy computation, (Eq. (2.9)) is commonly rewritten in its Lagrangian form.

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\gamma_1}{2}||\mathbf{X}\mathbf{u}||_2^2 - \frac{\gamma_2}{2}||\mathbf{Y}\mathbf{v}||_2^2 - \beta_1 P_1(\mathbf{u}) - \beta_2 P_2(\mathbf{v}) \tag{2.10}$$

[77] and [78] explored two penalty forms, ℓ_1 penalty and the chain structured fused Lasso penalty. ℓ_1 penalty imposes sparsity on both $\mathbf{u}$ and $\mathbf{v}$, and assumes that each canonical correlation involves only a few features from $\mathbf{X}$ and $\mathbf{Y}$. The fused Lasso penalty promotes the smoothness of weight vectors, and encourages neighboring features to be selected together.

For imaging and genetics data, neighboring information is usually not as informa-

tive as the group and network structure. Due to the crucial role of prior structure information in the high dimensional data, other groups later come up with more SCCA extensions to incorporate more realistic imaging structures with group- and network-guided penalties [10,12]. Based on the assumption that a real imaging genetic signal typically involves a small number of SNPs and brain regions, SCCA has also been applied in several imaging genetic studies by imposing the Lasso regularization term to yield sparse results [15, 38, 67]. Yet the popular solution of existing SCCA-based algorithms are designed using the soft thresholding technique, which requires the input data $\mathbf{X}$ and $\mathbf{Y}$ to have an orthonormal design $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ and $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}$ (see Section 10 in [61]), indicating that all the features should be independent from each other. In this case, ℓ_2 penalty term will be greatly simplified and then the solution could be sought in the unit ball. But unfortunately this assumption can hardly hold for either the imaging or genetic data (e.g., functional brain networks and LD blocks in the genome). And such failure to acknowledge the intrinsic covariance structure in the input data will inevitably limit the capability of yielding optimal results.

One possible solution is to orthogonalize the input data through principal component analysis (PCA) before running SCCA. However, we aim to identify relevant imaging and genetic markers, and thus prefer a sparse model. The combined PCA and SCCA strategy cannot achieve this goal, since PCA loadings on the original imaging and genetic markers are non-sparse. In this work, we will develop more advanced association models on top of SCCA, with additional penalty terms included to help incorporate the prior knowledge into the learning procedure. Their performance will be evaluated through associating imaging and genetics data with guidance of various brain structures and transcriptome knowledge derived from independent datasets.

Chapter 3

LOCALIZED CORTICAL BIOMARKERS FOR PREDICTING COGNITIVE OUTCOMES USING GROUP $\ell_{2,1}$-NORM

Despite the two well-known hallmarks of AD: beta-amyloid plaques and neurofibrillary tangles, various cognitive tests remain the most common clinical routine for diagnosis. Compared with the binary disease status, AD-relevant cognitive outcomes, a continuous quantitative trait, may be more informative for revealing the underlying disease mechanisms. And biomarkers identified from cognitive outcomes should be more representative than those from binary status. In this chapter, we present a framework, which combines clustering strategy and brain structure guided sparse regression model, to help identify localized cortical regions responsible for the changes of correlated cognitive outcomes.

3.1 BACKGROUND

In most existing studies, summary statistics of each region of interest (ROI) (e.g., average intensity) have been widely used as input features. Although voxel-based image measures or vertex-based surface measures could provide more detailed morphometric information than ROI summary statistics, direct application of conventional regression models to these measures may be inadequate to yield biologically meaningful results. For example, standard linear or ridge regression model typically produces non-sparse results that are not ideal for biomarker discovery. Conventional sparse models such as Lasso [61] are likely to yield scattered patterns hard to interpret, due to the lacking of proper handling of the spatial correlation and prior

anatomical knowledge in these models. To address this issue, we propose to employ a new sparse multi-task learning model called G-SMuRFS [71] for identifying effective surface biomarkers that can predict cognitive outcomes. We demonstrate its effectiveness by examining the predictive power of detailed cortical thickness measures towards three types of cognitive scores (ADAS, MMSE and RAVLT) in the ADNI cohort.

Enormous efforts have been made to evaluate the power of sparse learning methods in the neuroimaging field, such as identifying structural [3, 8, 52, 68, 69, 72] or functional [27, 34, 42, 63] imaging biomarkers associated with other imaging modality [3], cognitive scores [34, 63, 68, 69, 72], behavior [27, 42], as well as diagnostic conditions [8, 52]. However, using detailed cortical surface measures to predict cognitive outcomes is an under-explored area. In this study, we attempt to explore a novel application of G-SMuRFS to the identification of detailed surface-based cortical biomarkers that are relevant to cognitive outcomes. G-SMuRFS proposes a group-level $\ell_{2,1}$-norm strategy to achieve three goals: (1) group relevant surface features together in an anatomically meaningful manner (i.e., ROI information is incorporated) and use this prior knowledge to guide the learning process (i.e., spatial correlation within each ROI is addressed); (2) take into account the correlation among cognitive outcomes for building a more appropriate predictive model (i.e., multiple correlated cognitive scores are predicted together); and (3) optimize the selection of cognition-relevant surface biomarkers while maintaining high prediction accuracy. The high dimensionality of the vertex-based cortical surface data (e.g., 327,684 vertices in our study) introduces major computational challenges. To address this issue, we introduce a down-sampling technique to merge neighboring vertices into small surface patches to

reduce the computational cost while preserving detailed surface information. Our overarching goal is to examine and validate the predictive power of these detailed cortical thickness measures towards cognitive outcomes while considering the group structures defined by anatomically meaningful ROIs. The results may provide important information about potential surrogate biomarkers for early detection and/or therapeutic trials in AD.

3.2 MATERIALS AND METHODS

3.2.1 NEUROIMAGING AND COGNITION DATA

All the data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) [75]. 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.

We downloaded the baseline 1.5 T magnetic resonance imaging (MRI) scans, demographic information, and baseline diagnosis for all the ADNI Phase 1 (ADNI-1) participants. We also downloaded three types of baseline cognitive scores: Alzheimer's Disease Assessment Scale (ADAS), Mini-Mental State Examination (MMSE), and Rey Auditory Verbal Learning Test (RAVLT). For each participant, FreeSurfer V4, an automatic brain segmentation and cortical parcellation tool, was applied to automatically label cortical and subcortical tissue classes [18, 24] and to extract surface-based

Table 3.1: Thickness measures at surface locations from the following 34 pairs of bilateral FreeSurfer cortical regions (68 ROIs in total) were analyzed in this study.

ID	ROI Name	ID	ROI Name
1	banks of the superior temporal sulcus	18	pars opercularis
2	caudal anterior cingulate	19	pars orbitalis
3	caudal middle frontal gyri	20	pars triangularis
4	corpus collosum	21	pericalcarine gyri
5	cuneus	22	postcentral gyri
6	entorhinal cortex	23	posterior cingulate
7	fusiform gyri	24	precentral gyri
8	inferior parietal gyri	25	precuneus
9	inferior temporal gyri	26	rostral anterior cingulate
10	isthmus cingulate	27	rostral middle frontal gyri
11	lateral occipital gyri	28	superior frontal gyri
12	lateral orbitofrontal gyri	29	superior parietal gyri
13	lingual gyri	30	superior temporal gyri
14	medial orbitofrontal	31	supramarginal gyri
15	middle temporal gyri	32	frontal pole
16	parahippocampal gyri	33	temporal pole
17	paracentral lobule	34	transverse temporal pole

thickness measures. We focused our study on examining the thickness measures from surface locations labeled with any of the 34 FreeSurfer cortical ROIs (shown in Table 3.1) in both hemispheres. The measures from surface locations labeled with "unknown" were excluded in this study. Following a previous imaging genetics study [55], in this work, we concentrated our analyses on the Caucasian subjects determined by population stratification analysis using the ADNI genetics data [53]. 718 out of 745 Caucasian participants with no missing MRI morphometric and the cognitive outcome information were included in the study. The 718 participants were categorized by three baseline diagnostic groups: healthy control (HC, n=197), mild cognitive impairments (MCI, n=349), and Alzheimer's disease patients (AD, n=172). Demographics information of these subjects can be found in Table 6.1. All the imaging and cognitive outcome measurements were adjusted for age, gender, education and handedness, while intracranial volume was applied as an extra covariate for imaging measurements.

Table 3.2: Participant characteristics

Category	HC	MCI	AD
Number of Subjects	197	349	172
Gender (M/F)	107/90	224/125	94/78
Handedness (R/L)	183/14	316/33	160/12
Baseline Age (years, mean±SD)	76.2±5.0	75.0±7.3	75.6±7.5
Education (years, mean±SD)	16.2±2.7	15.7±3.0	14.9±3.1

3.2.2 G-SMURFS

As mentioned in Chapter 2, G-SMuRFS is a multivariate sparse regression model with the capability of incorporating group structure as prior knowledge. Initially motivated by sparse learning, such as Lasso [61] and group Lasso [82], G-SMuRFS has a new regularization term to consider both the group sparsity through the $G_{2,1}$-norm and the individual biomarker sparsity for joint learning via an $\ell_{2,1}$-norm regularization [46]. In the objective function Eq. (3.1), while the second term couples all the regression coefficients of a group of features across all the c tasks together, the third term penalizes all c regression coefficient of each individual feature as whole to select features across multiple learning tasks.

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F + \gamma_1||\mathbf{W}||_{G_{2,1}} + \gamma_2||\mathbf{W}||_{2,1} \ . \tag{3.1}$$

Solution of the objective function Eq. (3.1) can be obtained through an iterative optimization procedure. By setting the derivative with respect to $\mathbf{W}$ to zero, $\mathbf{W}$ can be solved as in Eq. (3.2).

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D} + \gamma_2\tilde{\mathbf{D}})^{-1}\mathbf{X}\mathbf{Y}^T, \tag{3.2}$$

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, m_k is the total feature numbers included in

group k, $\tilde{\mathbf{D}}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{w}^i\|_2}$. Detailed optimization procedure and algorithm can be found in [71].

Generally, the advantage of this model is three-fold: (1) It addresses the highly correlated nature of the cortical vertices within each surface ROI. (2) It takes into account the correlation of multiple scores of the same cognitive function test. (3) It achieves both the global biomarker sparsity as well as ROI group sparsity.

3.3 EXPERIMENTAL RESULTS AND DISCUSSION

3.3.1 EXPERIMENTAL SETTING

In this study, we examined all the cortical thickness measures across 34 pairs of bilateral cortical surface ROIs (68 ROIs in total) (Table 3.1) regarding their power for predicting the ADAS, MMSE and RAVLT cognitive scores. Our cortical surface data, generated by FreeSurfer, contains 327,684 vertices per surface. For the efficiency purpose, we completed a preprocessing step to down-sample 327,684 vertex-based thickness measures to 3,133 surface-patch-based measures using the following approach. First, we randomly selected cortical surface data from 50 HC participants. Second, for each ROI (say, with m vertices), we performed the k-mean clustering using this pre-selected HC subset to partition the ROI into roughly m/100 surface patches, where each patch was formed by a set of neighboring vertices with similar thickness. As a result, 3,133 patches were defined on the cortical surface. Third, excluding 320 patches from the region labeled as "unknown", we got 2813 patches from the ROIs shown in Table 3.1. Finally, we applied this patch scheme to the entire data set. The cortical thickness measures of all vertices within one patch were averaged to

represent the patch-level thickness measure. These 2813 patch-level measures were used as predictors in our regression analysis.

The response variables in the multivariate multiple regression analysis included the following five cognitive scores: ADAS-cog total score (ADAS), MMSE score (MMSE), RAVLT total score (TOTAL), RAVLT 30 minutes delay score (T30), and RAVLT recognition score (RECOG). To provide an unbiased estimate of the prediction performance of each method tested in the experiments, we employed five-fold cross-validation, where each fold contained a similar portion of AD, MCI and HC participants. We calculated the following two metrics to compare the prediction performance across different methods: (1) Root Mean Square Error (RMSE) between the actual and predicted scores of all the test subjects; and (2) Pearson Correlation Coefficient (CC) between the actual and predicted scores of all the test subjects.

In our experiments, we compared G-SMuRFS with four competing multivariate regression methods: (1) $\ell_{2,1}$-norm, (2) Partial Least Square (PLS), (3) ridge, and (4) linear regression. Parameters for these models were optimally tuned using a nested cross-validation strategy on the training data, with search grid in the range of $[5 \times 10^{-3}, 5 \times 10^{3}]$. For these regression analyses, the input data included 2,813 surface patch-level thickness measures as predictors and cognitive scores as response variables. We also performed univariate surface-based analysis using SurfStat [16] to cross check whether univariate and multivariate methods could yield similar patterns.

3.3.2 RESULTS AND DISCUSSION

Prediction performance, measured by RMSE and CC, of the cortical thickness measurement under five different regression models is shown in Table 3.3. The prediction

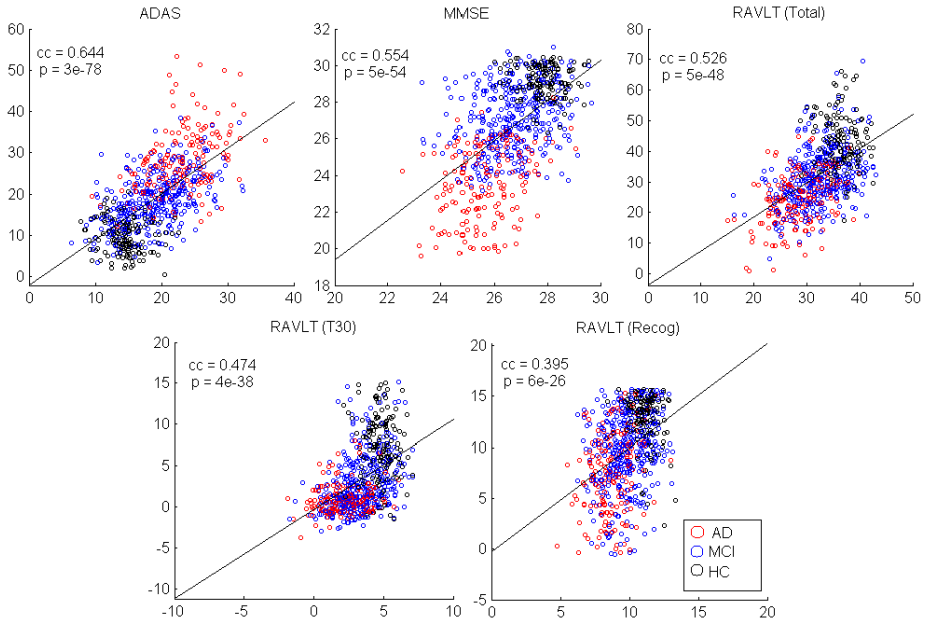


Figure 3.1: Scatter plots of actual (on y-axis) and predicted (by G-SMuRFS, on x-axis) cognitive scores. Note that the actual cognitive scores are pre-adjusted and thus may have negative values. The testing samples across five cross-validation trials were pulled together to calculate the correlation coefficients (CC) and the p values. Thus, the CCs shown here are slightly different from the CCs shown in Table 3.3 that were calculated separately for each cross-validation trial.

performances using those features selected by G-SMuRFS and $\ell_{2,1}$-norm are higher (i.e., lower RMSE and higher CC) than those of linear, ridge and PLS regression models. In particular, G-SMuRFS significantly outperforms PLS and linear regression on predicting all five scores, and ridge regression on predicting MMSE and RAVLT-RECOG. Prediction performances of G-SMuRFS and $\ell_{2,1}$-norm are similar. Fig. 3.1 shows scatter plots of actual and predicted (by G-SMuRFS) cognitive scores.

Fig. 3.2 shows the histogram of regression weights associated with all the cortical measures for each method, in an example cross-validation trial. While, in ridge and linear regression models, most surface measures share relatively similar impact on

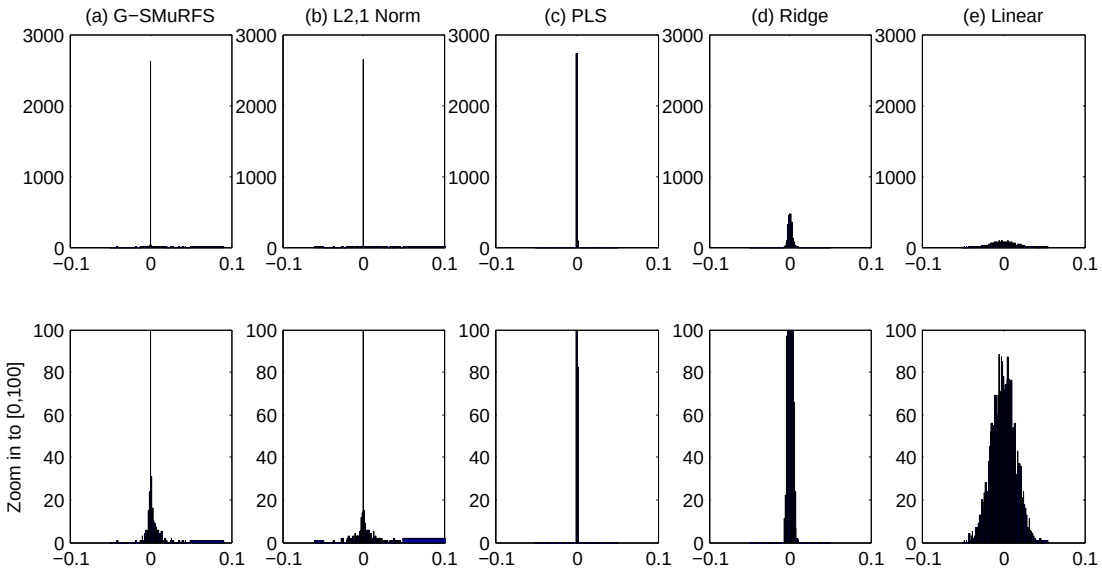


Figure 3.2: Histogram of regression weights of all cortical measures for predicting the RAVLT TOTAL score in an example cross-validation trial. Shown from left to right are the results of (a) G-SMuRFS, (b) $\ell_{2,1}$-norm, (c) PLS, (d) ridge regression, and (e) linear regression. The top row shows the complete histograms, and the bottom row shows the zoom in view of the partial histograms for $y \in [0, 100]$.

the prediction performance, G-SMuRFS and $\ell_{2,1}$-norm present a much better sparsity across all the cortical measures. Besides the sparsity at the cortical patch level, we also examined the group sparsity of all five models at the ROI level. Fig. 3.3 shows the histogram of "high impact" (i.e., top 50) cortical markers against each of the 34 pairs of bilateral ROIs for (a) G-SMuRFS, (b) $\ell_{2,1}$-norm, (c) PLS, (d) ridge regression, and (e) linear regression respectively. It is shown that high impact biomarkers identified through $\ell_{2,1}$-norm, ridge regression, and linear regression scattered across a large portion of cortical surface regions, making the result hard to interpret. G-SMuRFS yielded sparse patterns at the ROI level that have the potential for identifying relevant biomarkers. Although PLS also yielded sparse patterns, the predictive power of its top 50 markers (RMSE=1.045, CC=0.377) is lower than that of G-SMuRFS's (RMSE=0.938, CC=0.46).

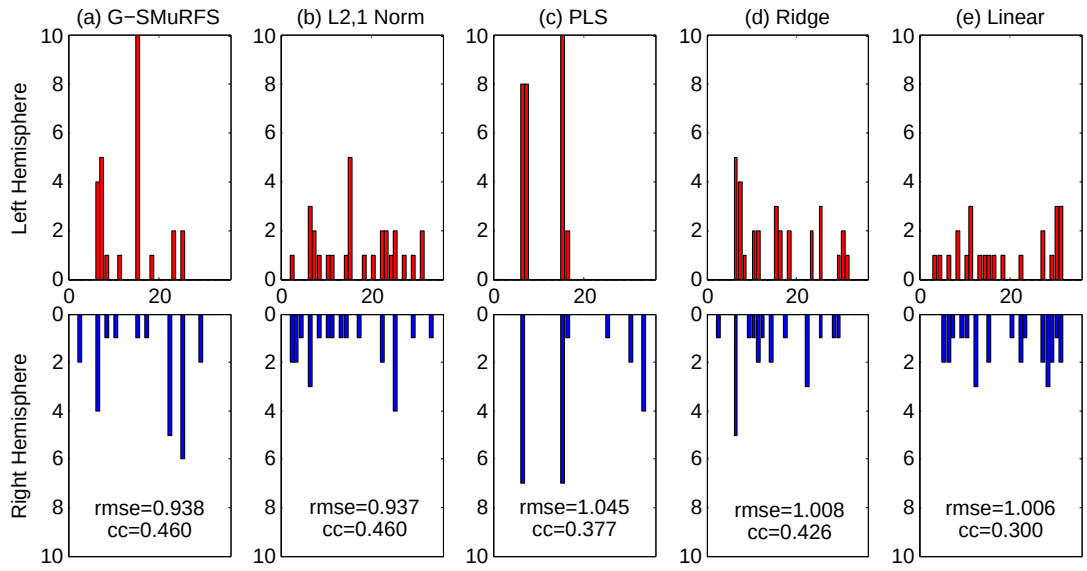


Figure 3.3: Number of "high impact" (i.e., top 50) cortical markers for predicting the RAVLT TOTAL score, in an example cross-validation trial, is plotted against the corresponding ROI (34 ROIs in total). The x axis shows the ROI IDs (see Table 3.1 for the corresponding ROI names). The y axis shows the number of top markers in the left hemisphere ROI (top row) or the right hemisphere ROI (bottom row). Shown from left to right are the results of (a) G-SMuRFS, (b) $\ell_{2,1}$-norm, (c) PLS, (d) ridge regression, and (e) linear regression. The cross-validation performance using these top 50 markers, measured by root mean square error (RMSE) and correlation coefficient (CC) between the actual and predicted RAVLT TOTAL scores of all the test subjects, is shown in each panel.

Fig. 3.4 shows example G-SMuRFS regression weights that were averaged over the five cross-validation trials and were then mapped back onto the cortical surface. Our multi-task regression experiment was performed to identify thickness measures for jointly predicting ADAS, MMSE, RAVLT TOTAL, RAVLT RECOG, and RAVLT T30 scores. The weight maps for ADAS (Fig. 3.4a), MMSE (Fig. 3.4b), TOTAL (Fig. 3.4c), RECOG (Fig. 3.4d), and T30 (not shown) are very similar to one another except that the ADAS pattern is in the opposition direction. Thickness measures from left and right entorihnal cortex, left middle temporal gyri, left inferior parietal gyri, right medial orbitofrontal gyri, and right precunes are positively correlated to the

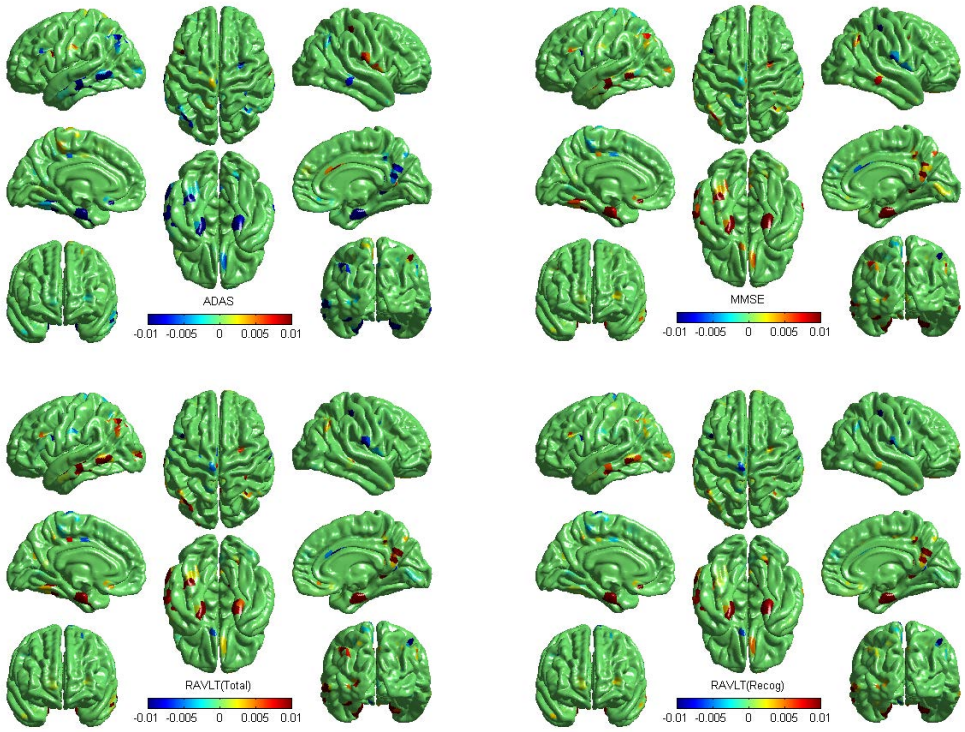


Figure 3.4: Example G-SMuRFS regression weights are color-coded and mapped onto the cortical surface. The red color indicates regions where the thickness is positively correlated with the corresponding cognitive score ((a) ADAS, (b) MMSE, (c) RAVLT-TOTAL, or (d) RAVLT-RECOG), and the blue color indicates regions where the thickness is negatively correlated with the score.

MMSE and RAVLT scores, and negatively correlated to ADAS. The measures from left fusiform are correlated to ADAS, MMSE, TOTAL and T30, and the measures from right middle temporal gyri are correlated to ADAS, MMSE, and RECOG. These patterns identified by our multivariate G-SMuRFS regression analysis match well with the weight map patterns computed by the univariate SurfStat analysis shown in Fig. 3.5.

The ROIs identified in this work are either related to AD or in accordance with findings in similar prior studies. For example, entorihnal cortex (part of medial temporal cortex) and precuneus are among the cortical signature of AD studied in

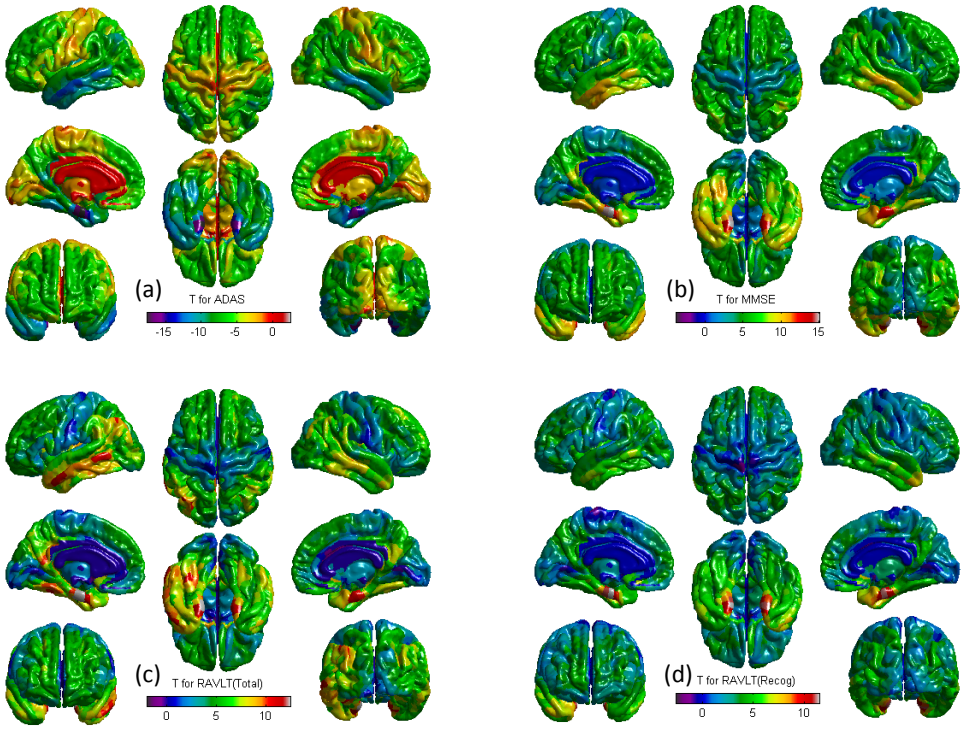


Figure 3.5: Example SurfStat t-statistic map: t-statistics are color-coded and mapped onto the cortical surface.

[4,20,68,69,72] performed similar regression studies for predicting cognitive outcomes using MRI measures. However, they examined only summary statistics (volume, thickness, or gray matter density) of both cortical and subcortical ROIs instead of detailed cortical thickness measures. The mean thickness of entorihnal cortex was found to be correlated with ADAS [**?**]wan2014), MMSE [68] and RAVLT [68, 72] scores. The mean thickness of inferior parietal gyri was found to be correlated with ADAS [68]and RAVLT [72] scores. The mean thickness of middle temporal gyri was found to be correlated with RAVLT scores [69]. Partly due to the detailed cortical analysis, this work identified some additional ROIs associated with the studied cognitive scores. Replication of these results in independent samples will remain of critical importance for confirmation. The computational cost of G-SMuRFS was

similar to that of the $\ell_{2,1}$-norm model but more expensive than linear, ridge and PLS regressions. We implemented all the regression models using Matlab. For one cross-validation trial in our experiments, G-SMuRFS and $\ell_{2,1}$-norm took 75-77 seconds while linear took 48 seconds, and ridge and PLS took ≤ 2 seconds. One interesting future direction is to develop more efficient implementation of G-SMuRFS and make it applicable to the analysis of larger scale data sets. To sum up, our empirical results are very encouraging and have demonstrated the promise of the G-SMuRFS method in the application of relating cortical morphology to cognitive outcomes: (1) G-SMuRFS regression model outperformed linear, ridge and PLS regression models, and performed similarly to the multi-task $\ell_{2,1}$-norm model in terms of overall RMSE and CC (Table 3.3). (2) The biomarkers identified by the G-SMuRFS method were sparser at the patch level than linear regression and ridge regression, and yielded a more stable performance for predicting cognitive scores. (3) Both G-SMuRFS and $\ell_{2,1}$-norm methods yielded sparse results at the vertex level (Fig. 3.2), however the G-SMuRFS model presented a sparser pattern at the ROI level (Fig. 3.3) than the $\ell_{2,1}$-norm model. Taking into account the spatial information makes the best use of the detailed surface information, yet leading to a clustered group level result instead, which is more visible and interpretable.

3.4 CONCLUSION

We have investigated the power of detailed cortical thickness measurements for predicting ADAS, MMSE and RAVLT cognitive scores using the data from the ADNI cohort. We have proposed to employ a newly developed sparse multi-task learning algorithm called G-SMuRFS, and have observed the following strengths of this approach

Table 3.3: Cross validation performance comparison of linear regression, ridge regression, PLS, $\ell_{2,1}$-norm, and G-SMuRFS: performance is measured by the root mean squared error (RMSE) and correlation coefficient (CC) between the actual and predicted scores of the test subjects. The p values, calculated from the paired sample t test between two sets of RMSEs or CCs, are shown for comparing G-SMuRFS with each competing method.

(a) Performance comparison using RMSE	ADAS	MMSE	TOTAL	RAVLT T30	RECOG
G-SMuRFS	0.7663±0.0375	0.8325±0.0399	0.8490±0.0654	0.8776±0.0701	0.9167±0.0471
$\ell_{2,1}$-norm	0.7631±0.0346	0.8322±0.0411	0.8464±0.0697	0.8807±0.0657	0.9215±0.0440
PLS	0.8844±0.0425	0.8999±0.0453	0.9246±0.0533	0.9515±0.0578	0.9706±0.0512
Ridge	0.7859±0.0327	0.8579±0.0332	0.8453±0.0697	0.8977±0.0553	0.9315±0.0430
Linear	1.0524±0.0559	1.1342±0.0739	1.1726±0.0956	1.3042±0.0803	1.2775±0.1168
P values (RMSE comparison): G-SMuRFS vs others					
$\ell_{2,1}$-norm	0.33752	0.67577	0.48045	0.35188	0.30053
PLS	0.00023	0.00136	0.00082	0.00099	0.01615
Ridge	0.07175	0.01939	0.56285	0.13234	0.03379
Linear	0.00172	0.00278	0.00034	0.00086	0.00434

Continued.

(b) Performance comparison using CC	ADAS	MMSE	TOTAL	RAVLT T30	RECOG
G-SMuRFS	0.6438±0.0258	0.5552±0.0078	0.5277±0.0539	0.4753±0.0591	0.3985±0.0533
$\ell_{2,1}$-norm	0.6468±0.0175	0.5548±0.0058	0.5301±0.0544	0.4692±0.0522	0.3889±0.0588
Partial Least Square	0.4608±0.0551	0.4339±0.0406	0.3782±0.0591	0.3037±0.0565	0.2390±0.0382
Ridge	0.6167±0.0171	0.5127±0.0272	0.5296±0.0540	0.4364±0.0416	0.3632±0.0681
Linear	0.4902±0.0544	0.3779±0.0246	0.3685±0.1005	0.2958±0.1359	0.2533±0.0726
P values (CC comparison): G-SMuRFS vs others					
$\ell_{2,1}$-norm	0.52828	0.86331	0.69516	0.36234	0.24498
PLS	0.00080	0.00225	0.00118	0.00045	0.01150
Ridge	0.06132	0.01676	0.87913	0.12379	0.02943
Linear	0.01378	0.00005	0.02002	0.05008	0.01715

that could greatly improve the prediction performance: (1) seamless integration of anatomical knowledge in the learning process by coupling cortical measures from the same ROI together; (2) sparsity at both patch level and ROI level; and (3) multitask learning scheme for addressing correlation among response variables.

Compared to Linear, ridge, PLS, or $\ell_{2,1}$-norm regression, combining the group $\ell_{2,1}$-norm in the regularization term has not only helped select the potential biomarkers in a few ROIs, but also improved overall predictive power. Its application to multimodal imaging data would be promising future directions for biomarker discovery and better mechanistic understanding in AD research. Exploration of other imaging modalities as well as the combination of multiple modalities warrants further investigation. Further effort may be made to include more complicated prior structure, like multiple layer groups or networks, to guide the learning procedure. Another possible future topic could be to investigate whether nonlinear models can help improve the prediction rates as well as derive biologically meaningful results.

Chapter 4

JOINT IDENTIFICATION OF IMAGING AND PROTEOMICS BIOMARKERS USING NG-L21

While G-SMuRFS does present a better performance with its capability of incorporating prior brain structures, it only considers spatial correlation within brain regions. But human brain functions more as a complex network rather than simple grouping of ROIs. In this chapter, we propose a novel sparse regression model to take the brain network as prior and evaluate its performance by examining the prediction power of multimodal measurements toward cognitive outcomes.

4.1 BACKGROUND

The study of Alzheimer's disease (AD) is experiencing an important shift from disease categories to AD-relevant outcomes for better identification of biomarkers for early detection. One particular example is the extension from disease status to cognitive outcomes. Although beta-amyloid plaques and neurofibrillary tangles are two major hallmarks of AD [74], various cognitive tests remain the most common way to help with clinical diagnosis. Exploring the relationship between multimodal biomarker measurements and cognitive outcomes has become an important research topic.

Earlier studies have been performed on single modality data, such as magnetic resonance imaging (MRI), positron emission tomography (PET), or cerebrospinal fluid (CSF) biomarkers. Recent efforts have turned to multimodal data and reported improved performance (e.g., MRI, FDG-PET and CSF biomarkers were jointly studied in [65]). Since multiple modalities are not completely isolated, there exist inter-

correlations among them. Existing multimodal methods typically employ a simple strategy to bundle these data together, and thus ignore or oversimplify their relationships(e.g. [79]).

To address this issue, we propose a new network-guided sparse learning model embracing both the complementary information and inter-relationships between modalities. The proposed model is applied to evaluate the predictive power of MRI and CSF proteomic measurements towards cognitive outcomes. The empirical results demonstrate significant improvements over the state-of-the-arts competing models, and also yield stable multimodal biomarkers across cross-validation trials.

4.2 METHODS

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.

We focus on multi-task learning paradigm, where MRI and CSF measures are used to predict one or more cognitive outcomes. Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subseteq \Re^d$ be MRI and CSF measures and $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\} \subseteq \Re^c$ cognitive outcomes, where n is the number of samples, d is the number of predictors (feature dimensionality) and c is the number of response variables (tasks). Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$.

The $\ell_{2,1}$ norm [44] is a multi-task version of traditional lasso. While lasso only focuses on the feature level sparsity, The $\ell_{2,1}$ norm is proposed to couple multiple

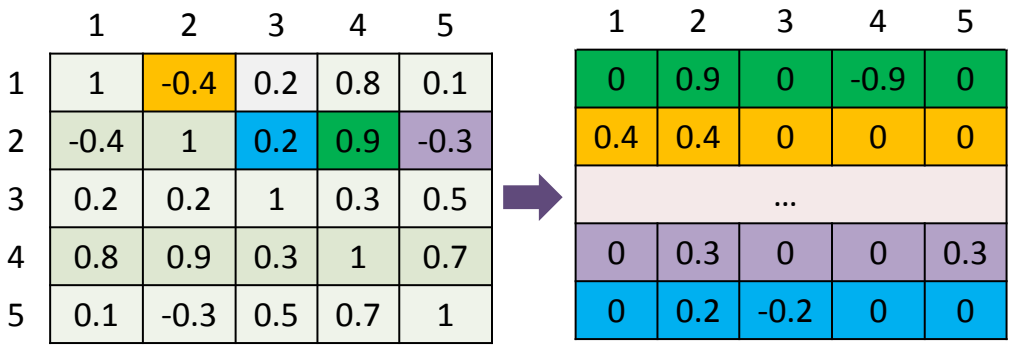


Figure 4.1: Each row in network matrix $\mathbf{N}$ (Right) corresponds to an element in correlation matrix $\mathbf{R}$ (Left).

tasks together in addition to the original sparsity property:

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma||\mathbf{W}||_{2,1} \ . \tag{4.1}$$

Yet in this model the rows of $\mathbf{W}$ are equally treated, which ignores the structures among predictors. Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) method [70] was proposed to exploit the structures within and between the predictors and response variables. It assumes 1) a partition scheme exists among predictors, and 2) predictors within one partition should have similar weights. G-SMuRFS can be thought of as a multi-task version of group lasso.

In practice, the relationship among predictors may not be as simple as a straightforward partition as used in G-SMuRFS. Many studies have shown that the human brain can be modeled as a complex network. To model these complicated structures among predictors, we propose a new *Network-Guided* $\ell_{2,1}$ *Sparse Learning (NG-L21)* model as follows.

Let $\mathbf{R}$ be the predictor correlation matrix with $\mathbf{R_{ij}}$ indicating correlation between

predictors i and j. Given a threshold t, we can construct a network by connecting predictors with high positive and negative correlations (i.e., $|\mathbf{R_{ij}}| \geq t$). We hypothesize that positively (negatively) correlated predictors share positively (negatively) similar weights across tasks.

Therefore, we try to minimize

$$\sum_{1 \leq i < j \leq d \ \& \ |\mathbf{R_{ij}}| \geq t} ||\mathbf{W}_i - sign(\mathbf{R_{ij}})\mathbf{W}_j||_F^2. \tag{4.2}$$

Since the selection of a proper threshold may be trivial, we also propose a threshold-free or weighted method as follows:

$$\sum_{1 \leq i < j \leq d} \mathbf{R_{ij}} ||\mathbf{W}_i - sign(\mathbf{R_{ij}})\mathbf{W}_j||_F^2. \tag{4.3}$$

Here $sign(\mathbf{R_{ij}})$ deals with positive and negative correlations while $\mathbf{R_{ij}}$ itself is applied to reduce the constraint impact on the less correlated pairs. Thus, the whole regularization term can be reformatted as $||\mathbf{NW}||_F^2$, in which $\mathbf{N}$ is a neighboring matrix transformed from the symmetric correlation matrix $\mathbf{R}$. In $\mathbf{N}$, each row corresponds to one element in the correlation matrix $\mathbf{R}$ (Fig. 4.1). For each pair of predictors i, j, we create a row in $\mathbf{N}$ with i-th entry as $-\mathbf{R_{ij}}$, j-th entry as $-\mathbf{R_{ij}}\mathbf{sign}(\mathbf{R_{ij}})$ and all the other entries as zeros. The intuition is that the weight difference between two correlated predictors should be minimized, which is reflected by the new regularization term of $||\mathbf{NW}||_F^2$. Also by including the weight, the more correlated a feature pair is, the more constraint the pair is imposed by. We call this model *NG-L21*. Although this model is similar to graph-guided fusion [35] and group weighted fusion [25], neither

of the prior models explored the correlation of predictors and multi-task responses together. The final objective function is formulated as:

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1||\mathbf{NW}||_F^2 + \gamma_2||\mathbf{W}||_{2,1} \ , \tag{4.4}$$

where $\mathbf{N}$ is a sparse matrix where each row indicates a neighborhood relationship (i.e., edge) in the predictor network.

Well known to be non-derivable, ℓ_1-norm can be easily solved by approximate lasso where an extremely small value is added to enable the smoothness. Eq. (4.4) can then be solved by simply taking the derivative w.r.t $\mathbf{W}$ and setting it to 0:

$$\mathbf{X}\mathbf{X}^T\mathbf{W} - \mathbf{X}\mathbf{Y}^T + \gamma_1\mathbf{D_1}\mathbf{W} + \gamma_2\mathbf{D_2}\mathbf{W} = 0, \tag{4.5}$$

where $\mathbf{D_1} = \mathbf{N^T N}$, a matrix with each row integrating all neighboring relations. For the i-th row, it is the sum of all the rows in $\mathbf{N}$ whose i-th element is nonzero. $\mathbf{D_2}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{w}^i\|_2}$. We have

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D_1} + \gamma_2\mathbf{D_2})^{-1}\mathbf{X}\mathbf{Y}^T, \tag{4.6}$$

where $\mathbf{W}$ can be efficiently obtained by solving $(\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D}_1 + \gamma_2\mathbf{D}_2)\mathbf{W} = \mathbf{X}\mathbf{Y}^T$. Following [70], an efficient iterative algorithm based on Eq. (4.6) can be developed as follows (Algorithm 1), and can be shown to converge to the global optimum.

Input: $\mathbf{X}$, $\mathbf{Y}$
Initialize $\mathbf{W}^1 \in \mathbb{R}^{d \times c}$, $t = 1$;
while *not converge* **do**
1. Calculate the diagonal matrices $\mathbf{D}_2^{(t)}$, where the i-th diagonal element of $\mathbf{D}_2^{(t)}$ is $\frac{1}{2\|\mathbf{w}_t^i\|_2}$;
2. $\mathbf{W}^{(t+1)} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D}_1 + \gamma_2\mathbf{D}_2^{(t)})^{-1}\mathbf{X}\mathbf{Y}^T$;
3. $t = t + 1$;
end
Output: $\mathbf{W}^{(t)} \in \mathbb{R}^{d \times c}$.

Algorithm 1: NG-L21 algorithm

4.3 RESULTS

4.3.1 DATA AND EXPERIMENTAL SETTING

The MRI, CSF proteomic, and cognitive data were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. One goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD [1]. For up-to-date information, see www.adni-info.org.

This study (N=204) included 66 AD, 57 MCI and 81 healthy control (HC) participants (Table 6.1). For each baseline MRI scan, FreeSurfer (FS) V4 was employed to extract 73 cortical thickness measures and 26 volume measures. 82 CSF proteomic analytes, evaluated by Rules Based Medicine, Inc. (RBM) proteomic panel [48] and surviving quality control process, were also included in this work. The 99 imaging measures and 82 proteomic analytes were used to predict a set of cognitive scores [1]: Rey Auditory Verbal Learning Test (RAVLT, 5 scores shown in Table 4.2 as joint outcomes). Using the regression weights from HC participants, all the MRI, CSF,

Table 4.1: Participant characteristics (all from ADNI-1).

Category	AD	MCI	HC
Number	66	57	81
Gender(M/F)	37/29	36/21	42/39
Handness(R/L)	63/3	55/2	77/4
Age(mean±std)	75.14±7.66	74.59±7.42	76.24±5.35
Education	15.04±2.97	15.79±2.88	15.86±2.85

Table 4.2: RAVLT scores.

Score 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 number of words recalled
RECOG	30 minute delay recognition score

and cognitive measures were pre-adjusted for the baseline age, gender, education, and handedness, with intracranial volume as an additional covariate for MRI only.

4.3.2 EXPERIMENTAL RESULTS

We denote the weighted network model as NG-L21$_{\mathbf{w}}$, and the thresholded one as NG-L21$_{\mathbf{t}}$. For comparing performances between these two models and competing methods (i.e., Linear, Ridge, elastic net and L21), regression analysis was conducted jointly on all five RAVLT scores. Based on the assumption that FS and CSF measures could provide complementary information, we performed 18 experiments based on six different methods and three datasets (FS, CSF, FS+CSF). In each experiment, Pearson's correlation coefficients (CCs) between the actual and predicted cognitive scores were computed to measure the prediction performances. Using 10-fold cross-validation, parameters were estimated and average CCs over 10 trials were reported.

In our experiments, CSF proteomic analytes were found to have limited prediction power by itself (typically CC$<$ 0.4). But combining CSF and FS yielded improved results than using FS alone (Table 4.3), indicating possible complementary information

Table 4.3: Average correlation coefficient between predicted and actual scores over 10 cross-validation trials: FS results (top panel) and FS+CSF results (bottom panel) are shown.

	TOTAL	T30	RECOG	TOT6	TOTB
NG-L21$_{\mathbf{w}}$	0.6084	0.5395	0.55	0.5337	0.4888
NG-L21$_{\mathbf{t}}$	0.5879	0.5173	0.5497	0.5052	0.4775
L21	0.574	0.5007	0.5227	0.4915	0.4771
ElasticNet	0.6032	0.4971	0.5462	0.5061	0.4953
Ridge	0.5996	0.5365	0.5316	0.5335	0.4828
Linear	0.4015	0.3505	0.3778	0.3141	0.2041
NG-L21$_{\mathbf{w}}$	0.6314	0.523	0.5885	0.5872	0.4991
NG-L21$_{\mathbf{t}}$	0.6312	0.5223	0.5908	0.5883	0.4954
L21	0.6207	0.5178	0.575	0.5744	0.4858
ElasticNet	0.6293	0.5154	0.5691	0.5397	0.5002
Ridge	0.6428	0.5445	0.5595	0.5935	0.4936
Linear	0.1935	0.0826	0.0944	0.1632	0.0374

provided by the two modalities. Both NG-L21 models outperformed the other methods in most cases. Ridge obtained comparable and sometimes better performances than NG-L21; but Ridge's root mean square error (not shown due to space limit) tended to be higher than NG-L21.

Fig. 4.2(a) and Fig. 4.2(b) show the regression weights in heatmap and brain respectively. Ridge produced non-sparse patterns, and made the results less interpretable. Both NG-L21 and L21 identified a small number of imaging markers, including AmygVol, EntCtx, and HippVol, which were known to be related to RAVLT scores. NG-L21$_{\mathbf{w}}$ achieved similar or slightly better performance than NG-L21$_{\mathbf{t}}$. This indicates that the NG-L21 performance is mainly determined by the correlations of high values and small weights (those were included in NG-L21$_{\mathbf{w}}$ but excluded in NG-L21$_{\mathbf{t}}$) have just modest effect on improving the performance. In addition, we also compared NG-L21 with G-SMuRFS using symmetric information as grouping strategy. Generally they achieved similar performance, but tuned parameters of $\ell_{2,1}$-norm in G-SMuRFS shrunk to almost 0, and led to non-sparse results.

Fig. 4.2(c) shows regression coefficients of CSF proteomic markers for all RAVLT scores. Due to their highly correlated structure, the coefficients across these scores are similar for the top proteomic markers. Some known AD-risk proteins (e.g., APOE, ApoC-III, CD40, FRTN) are detected for all these scores. For example, FRTN (ferettin) is the main iron-storage protein capable of containing thousands of iron atoms. Recently it has been reported that ferritin from AD patients has higher aluminum than that of controls [57]. And the irregular iron accumulative and disrupted iron metabolism have also been previously identified to be related with brain disorders. Detailed analysis of these identified proteomic markers warrants further investigation.

4.4 CONCLUSION

We proposed a new network-guided sparse learning framework, NG-L21, aiming to flexibly incorporate and model structure among predictors. Unlike traditional methods, this model could provide advantages in several folds: 1) explicitly incorporating the relationships among predictors in a more general way, 2) using data-driven patterns without any predefined parameters, 3) effectively identifying biomarkers influencing multiple responses, and 4) selection of correlated markers together rather than picking only one of them to improve the stability. With the application to the ADNI multimodal data (predicting memory scores from MRI and CSF proteomic measures), NG-L21 demonstrated improved prediction performance over the state-of-the-art competing methods, and identified stable and meaningful multimodal biomarkers. Combining MRI and CSF proteomic data yielded enhanced prediction performance than each single modality.

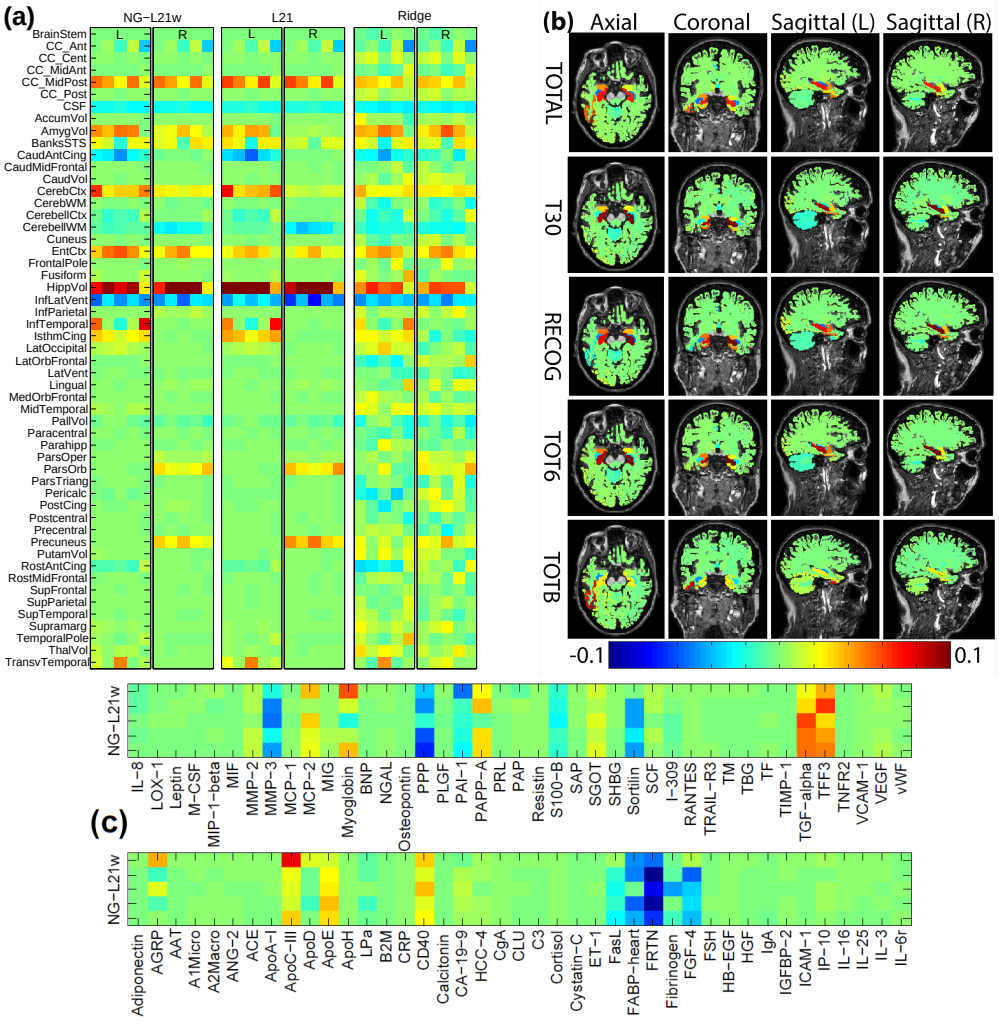


Figure 4.2: Heat maps of regression weights (average over 10-fold cross-validation) for predicting RAVLT scores using FS + CSF measures. (a) FS weights from NG-L21$_{\mathbf{w}}$, L21, and Ridge respectively. Results from left (L) and right (R) sides are shown in a pair of panels. For each panel, 5 columns show results for TOTAL, T30, RECOG, TOT6, and TOTB respectively. The measures shown in the first seven row are unilateral, and the remaining ones are bilateral. (b) FS weights from NG-L21$_{\mathbf{w}}$ mapped onto brain. (c) CSF weights from NG-L21$_{\mathbf{w}}$: 5 columns correspond to TOTAL, T30, RECOG, TOT6, and TOTB respectively.

Chapter 5

STRUCTURE-AWARE SCCA IN IMAGING GENETICS ASSOCIATION ANALYSIS

While regression analysis helps to discover the predictive biomarkers toward a specific task/task set, association study treats both datasets more symmetrically and alternatively seeks the sophisticated multivariate relationship among multimodal measurements. Brain imaging genetics is one typical example, which aims to identify associations between genetic factors such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from neuroimaging data. This chapter and the following Chapter 6 will discuss the association modeling based on various prior knowledge.

5.1 BACKGROUND

Unlike traditional univariate analyses [55] that aim to discover single-SNP-single-QT associations or regression analyses [29] for joint effect of multiple SNPs on one or a few QTs, bi-multivariate analyses [14, 39, 56, 64] have been recently applied to examine complex multi-SNP-multi-QT associations. SCCA is one typical bi-multivariate association model which can help yield sparse patterns in both imaging and genetic features for easy interpretation. But as we discussed earlier in Chapter 2, most existing SCCA algorithms are dependent on the soft threshold strategy for final solution, and assume the independence among data features. This assumption clearly does not hold in either imaging or genetic data, and therefore will inevitably limit the performance of the algorithm.

To overcome this limitation and still preserve the sparse patterns, we propose a novel structure-aware SCCA (denoted as S2CCA) algorithm for brain imaging genetics applications to achieve the following two goals: (1) our algorithm is not based on the soft threshold framework and eliminates the independence assumption for the input data; (2) our model can incorporate group-like structure (e.g., voxels in an ROI, or SNPs in an LD block) to yield more stable and biologically more meaningful results than conventional SCCA model. We perform an empirical comparison between the proposed S2CCA algorithm and a widely used SCCA implementation in the PMD software package (http://cran.r-project.org/web/packages/PMA/) [77] using both synthetical and real imaging genetic data. The empirical results demonstrate that the proposed S2CCA algorithm can yield improved prediction performances and biologically meaningful findings.

5.2 STRUCTURE-AWARE SCCA (S2CCA)

We denote vectors as boldface lowercase letters and matrices as boldface uppercase ones. For a given matrix $\mathbf{M} = (m_{ij})$, we denote its i-th row and j-th column to $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}^T \subseteq \Re^p$ be the SNP data and $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}^T \subseteq \Re^q$ be the imaging QT data, where n is the number of participants, p and q are the numbers of SNPs and QTs, respectively. CCA seeks linear combinations of variables in $\mathbf{X}$ and variables in $\mathbf{Y}$, which are maximally correlated between $\mathbf{Xu}$ and $\mathbf{Yv}$, that is:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \qquad s.t. \;\; \mathbf{u}^T\mathbf{X}^T\mathbf{Xu} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Yv} = 1 \tag{5.1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are canonical vectors or weights. Two major weaknesses of CCA are that it requires the number of observations n to exceed the combined dimension of $\mathbf{X}$ and $\mathbf{Y}$ and that it produces nonsparse $\mathbf{u}$ and $\mathbf{v}$ which are difficult to interpret. SCCA removes these weaknesses by maximizing the correlation between $\mathbf{Xu}$ and $\mathbf{Yv}$ subject to the weight vector constraints $P_1(\mathbf{u}) \leq c_1$ and $P_2(\mathbf{v}) \leq c_2$. The penalized matrix decomposition (PMD) toolkit [77] provided a widely used SCCA implementation, where the L_1 penalty $P(A) = \sum_{k=1}^{p} |A(k)|$ was used for both P_1 and P_2. As mentioned earlier, similar to most SCCA methods, PMD employed the soft threshold strategy for solving the L1 penalty term, which required the input data to have an orthonormal design $\mathbf{X^T X} = \mathbf{I}$ and $\mathbf{Y^T Y} = \mathbf{I}$ (see Section 10 in [62]). This independence assumption usually does not hold in imaging genetic data (e.g., correlated voxels in an ROI, correlated SNPs in an LD block), and thus inevitably limits the capability of identifying meaningful imaging genetic associations.

To overcome this limitation, we propose a novel structure-aware SCCA (denoted as S2CCA) algorithm to not only eliminate the independence assumption for the input data, but also incorporate group-like structure in the model. Instead of using L_1, we define a group L_1 constraint on P_1 and P_2 as follows:

$$\begin{aligned} P_1 &= ||\mathbf{u}||_G = \gamma_1 \sum_{k_1=1}^{K_1} \sqrt{\sum_{i \in \pi_{k_1}} u_i^2} = \gamma_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2, \\ P_2 &= ||\mathbf{v}||_G = \gamma_2 \sum_{k_2=1}^{K_2} \sqrt{\sum_{i \in \pi_{k_2}} v_i^2} = \gamma_2 \sum_{k_2=1}^{K_2} ||\mathbf{v}^{k_2}||_2. \end{aligned} \tag{5.2}$$

In Eq. (2), SNPs are partitioned into K_1 groups $\Pi_1 = \{\pi_{k_1}\}_{k_1=1}^{K_1}$, such that $\{u_i\}_{i=1}^{m_{k_1}} \in \pi_{k_1}$, and m_{k_1} is the number of SNPs in π_{k_1}; and imaging QTs are partitioned into K_2 groups $\Pi_2 = \{\pi_{k_2}\}_{k_2=1}^{K_2}$, such that $\{v_i\}_{i=1}^{m_{k_2}} \in \pi_{k_2}$, and m_{k_2} is the number of QTs in

π_{k_2}. $||\cdot||_G$ is the constraint for the group structure. In this work, we partition voxels using AAL ROIs and SNPs using LD blocks.

Now the S2CCA objective function can be formally written as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \gamma_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2 - \gamma_2 \sum_{k_2=1}^{K_2} ||\mathbf{v}^{k_2}||_2 \tag{5.3}$$
$$s.t.\ \ \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v} = 1,$$

Using Lagrange multipliers, Eq. (5.3) can be transformed as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \gamma_1||\mathbf{u}||_G - \gamma_2||\mathbf{v}||_G - \beta_1||\mathbf{X}\mathbf{u}||_2^2 - \beta_2||\mathbf{Y}\mathbf{v}||_2^2 \tag{5.4}$$

Taking the derivative about $\mathbf{u}$ and $\mathbf{v}$ and setting them to zero, we have

$$\mathbf{X}^T\mathbf{Y}\mathbf{v} - \gamma_1\mathbf{D}_1\mathbf{u} - \beta_1\mathbf{X}^T\mathbf{X}\mathbf{u} = 0, \tag{5.5}$$

$$\mathbf{Y}^T\mathbf{X}\mathbf{u} - \gamma_2\mathbf{D}_2\mathbf{v} - \beta_2\mathbf{Y}^T\mathbf{Y}\mathbf{v} = 0, \tag{5.6}$$

where $\mathbf{D}_1$ is the block diagonal matrix of the k_1-th diagonal block as $\frac{1}{2||\mathbf{u}^{k_1}||_2}$, and $\mathbf{D}_2$ is the block diagonal matrix of the k_2-th diagonal block as $\frac{1}{2||\mathbf{v}^{k_2}||_2}$.

With $\mathbf{v}$ fixed, we can use an approach similar to G-SMuRFS [70] to solve for $\mathbf{u}$. With $\mathbf{u}$ fixed, we can do the same to solve for $\mathbf{v}$. We propose Algorithm 2 to alternatively compute $\mathbf{u}$ and $\mathbf{v}$ until the result converges. We use $\max\{|\delta| \mid \delta \in (\mathbf{u}_{t+1} - \mathbf{u}_t)\} < 10^{-5}$ and $\max\{|\delta| \mid \delta \in (\mathbf{v}_{t+1} - \mathbf{v}_t)\} < 10^{-5}$ as stop criterion, and nested cross-validation to automatically tune parameters γ_1, γ_2, β_1 and β_2.

Input: $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}^T$, $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}^T$
$t = 1$, Initialize $\mathbf{u}_t \in \Re^{p\times 1}$, $\mathbf{v}_t \in \Re^{q\times 1}$;
while *not converge* **do**
1. Calculate the block diagonal matrix $\mathbf{D}_{1_t}$, where the k_1-th diagonal is $\frac{1}{2||\mathbf{u}_t^{k_1}||_2}$;
2. $\mathbf{u}_{t+1} = (\beta_1 \mathbf{X}^T\mathbf{X} + \gamma_1 \mathbf{D}_{1_t})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{v}_t$;
Scale $\mathbf{u}_{t+1}$ so that $\mathbf{u}_{t+1}^T\mathbf{X}^T\mathbf{X}\mathbf{u}_{t+1} = 1$;
3. Calculate the block diagonal matrix $\mathbf{D}_{2_t}$, where the k_2-th diagonal is $\frac{1}{2||\mathbf{v}_t^{k_2}||_2}$;
4. $\mathbf{v}_{t+1} = (\beta_2 \mathbf{Y}^T\mathbf{Y} + \gamma_2 \mathbf{D}_{2_t})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{u}_{t+1}$;
Scale $\mathbf{v}_{t+1}$ so that $\mathbf{v}_{t+1}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v}_{t+1} = 1$;
5. $t = t + 1$.
end
Output: $\mathbf{u}_t \in \Re^{p\times 1}$, $\mathbf{v}_t \in \Re^{q\times 1}$.

Algorithm 2: S2CCA algorithm

5.3 EXPERIMENTAL RESULTS

5.3.1 RESULTS ON SIMULATION DATA

We first performed a comparative study between S2CCA and PMD using simulated data. The following procedure is used to generate two synthetic datasets $\mathbf{X}$ and $\mathbf{Y}$, both with $n = 1000$ and $p = q = 50$: (1) We created a random positive definite non-overlapping group structured covariance matrix $\mathbf{M}$. (2) Data set $\mathbf{Y}$ with covariance structure $\mathbf{M}$ was calculated through Cholesky decomposition. (3) We repeated the above two steps to generate another data set $\mathbf{X}$. (4) Canonical loadings $\mathbf{u}$ and $\mathbf{v}$ were set based on the group structures of $\mathbf{X}$ and $\mathbf{Y}$ respectively, where all the variables within the group share the same weight. For simplicity, we selected only one group in $\mathbf{Y}$ to be associated with 4 groups in $\mathbf{X}$. (5) The portion of the specified group in $\mathbf{Y}$ were replaced based on the $\mathbf{u}$, $\mathbf{v}$, $\mathbf{X}$ and the assigned correlation. We generated 7 pairs of $\mathbf{X}$ and $\mathbf{Y}$ with correlations ranging from 0.45 to 0.99. The canonical loadings and group structure remained the same across all the synthetic data sets.

Table 5.1: Five-fold cross-validation performance on synthetic data: mean±std is shown for estimated correlation coefficients and AUC of the test data using the trained model. P-value of paired t-test between S2CCA and PMD results is also shown.

True	Correlation Coefficient (CC)			Area under ROC (AUC)					
CC	S2CCA	PMD	p	S2CCA:**u**	PMD:**u**	p	S2CCA:**v**	PMD:**v**	p
0.445	0.42±0.05	0.27±0.08	7E-4	1.00±0	0.68±0.02	4E-6	1.00±0	0.84±0.02	4E-5
0.526	0.48±0.04	0.32±0.11	4E-3	1.00±0	0.66±0.01	3E-7	1.00±0	0.87±0.06	3E-3
0.594	0.56±0.07	0.39±0.12	2E-3	1.00±0	0.64±0.01	3E-7	1.00±0	0.81±0.05	7E-4
0.697	0.67±0.01	0.47±0.07	2E-3	0.94±0.02	0.66±0.03	6E-5	1.00±0	0.85±0.04	3E-4
0.814	0.80±0.04	0.49±0.06	7E-5	0.98±0.02	0.63±0.01	1E-6	1.00±0	0.83±0.04	5E-4
0.906	0.90±0.01	0.56±0.06	9E-5	1.00±0	0.66±0.01	4E-7	1.00±0	0.82±0.04	4E-4
1.000	0.99±0.00	0.65±0.04	2E-5	1.00±0	0.66±0.01	3E-7	1.00±0	0.86±0.07	4E-3

We applied S2CCA and PMD to all seven data sets. The regularization parameters were optimally tuned using a grid search from 10^{-5} to 10^{5} through nested 5-fold cross-validation. The true and estimated $\mathbf{u}$ and $\mathbf{v}$ values are shown in Fig. 5.1. Due to different normalization strategies, the weights yielded through S2CCA and PMD showed different scales. Yet the overall profile of the estimated $\mathbf{u}$ and $\mathbf{v}$ values from S2CCA kept consistent with the ground truth across the entire range of tested correlation strengths (from 0.45 to 0.99), while PMD only identified an incomplete portion of all the signals. Furthermore, we also examined the correlation in the test set computed using the learned CCA models from the training data for both methods. The left part of Table 5.1 demonstrated that S2CCA outperformed PMD consistently and significantly, and it could accurately reveal the embedded true correlation even in the test data. The right part of Table 5.1 demonstrated the sensitivity and specificity performance using area under ROC (AUC), where S2CCA also significantly outperformed PMD no matter whether the correlation was weak or strong. From the above results, it can also be observed that S2CCA could identify the correlations and signal locations not only more accurately but also more stably.

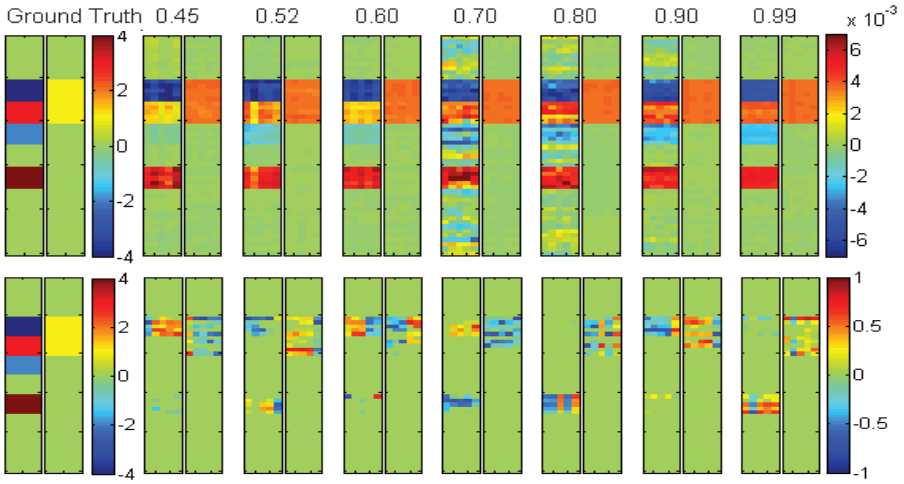


Figure 5.1: 5-fold trained weights of **u** and **v**. Ground truth of **u** and **v** are shown in the most left two panels. S2CCA results (top row) and PMD results (bottom row) are shown in the remaining panels, corresponding to true correlation coefficients (CCs) ranging from 0.45 to 0.99. For each panel pair, the five estimated **u** values are shown on the left panel, and the five estimated **v** values are shown on the right panel.

5.3.2 RESULTS ON REAL NEUROIMAGING GENETICS DATA

S2CCA and PMD were also compared using real neuroimaging and SNP data. The magnetic resonance imaging (MRI) and SNP data were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. One goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

This study included 176 AD, 363 MCI and 304 healthy control (HC) non-Hispanic Caucasian participants (Table 5.2). Structural MRI scans were processed with voxel-based morphometry (VBM) in SPM8 [2, 49]. Briefly, scans were aligned to a T1-weighted template image, segmented into gray matter (GM), white matter (WM)

Table 5.2: Participant characteristics.

	HC	MCI	AD
Num	304	363	176
Gender(M/F)	111/193	235/128	95/81
Handedness(R/L)	190/14	329/34	166/10
Age (mean±std)	76.07±4.99	74.88±7.37	75.60±7.50
Education (mean±std)	16.15±2.73	15.72±2.30	14.84±3.12

and cerebrospinal fluid (CSF) maps, normalized to MNI space, and smoothed with an 8mm FWHM kernel. Rather than using ROI summary statistics, in this study we subsampled the whole brain and examined correlations between the voxels (GM density measures) and SNPs. Totally 465 voxels spanning all brain ROIs were extracted. All SNPs within LD block of APOE e4 were extracted from an imputed genetic data set containing only SNPs in Illumina 610Q and/or OmniExpress arrays after basic quality control. As a result, four SNPs (rs429358, rs439401, rs445925, rs534007) from this LD block were included in this study. Using the regression weights derived from the healthy control participants, VBM and genetic measures are pre-adjusted for removing the effects of the baseline age, gender, education, and handedness.

Both S2CCA and PMD were evaluated on the normalized VBM and SNP measurements. Five-fold nested cross-validation were applied to search optimal parameters. Table 5.3 shows 5-fold cross-validation results, indicating that S2CCA significantly and consistently outperformed PMD in terms of identifying high correlations from the training data and replicating those in the testing data. Shown in Fig. 5.2(a) is the canonical loadings trained from 5-fold cross-validation, suggesting relevant imaging and genetic markers. Although the S2CCA model did not explicitly impose sparsity on individual voxels, it was still able to discover a very small number of relevant ROIs for easy interpretation due to the imposed group sparsity. The strongest imaging signals came from right hippocampus, which were inversely correlated with APOE e4

allele rs429358. In contrast, PMD identified many more ROIs than S2CCA (Fig. 5.2 (a-b)), making results hard to interpret. In addition, comparing the results from 5 cross-validation trials, S2CCA yielded a more stable and consistent pattern than PMD. It is reassuring that S2CCA identified a well-known correlation between hippocampal morphometry and APOE in an AD cohort, which shows the promise of S2CCA to correctly identify biologically meaningful imaging genetic associations.

Table 5.3: Five-fold cross validation results on real data: the CCA models learned from the training data were used to estimate the correlation coefficients between canonical components for both training and testing sets. P-values of paired t-tests were obtained for comparing S2CCA and PMD results.

Correlation coefficients	S2CCA					PMD					p-value
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	
Training	0.28	0.27	0.27	0.27	0.27	0.26	0.26	0.26	0.26	0.24	0.016
Testing	0.21	0.24	0.28	0.23	0.26	0.20	0.21	0.21	0.20	0.24	0.017

5.4 CONCLUSION

Most existing SCCA algorithms (e.g., [14, 39, 45, 64, 77]) are designed using the soft threshold strategy, which assumes that the features in the data are independent from each other. This independence assumption usually does not hold in imaging genetic data, and thus limits the capability of yielding optimal results. We have proposed a novel structure-aware sparse canonical correlation analysis (S2CCA) algorithm, which not only removes the above independence assumption, but also can take into consideration the group-like structure in the data. We have compared S2CCA with PMD (a widely used SCCA implementation) on both synthetic data and real imaging genetic data. The promising empirical results demonstrated that S2CCA significantly outperformed PMD in both cases. In addition, S2CCA could accurately recover the

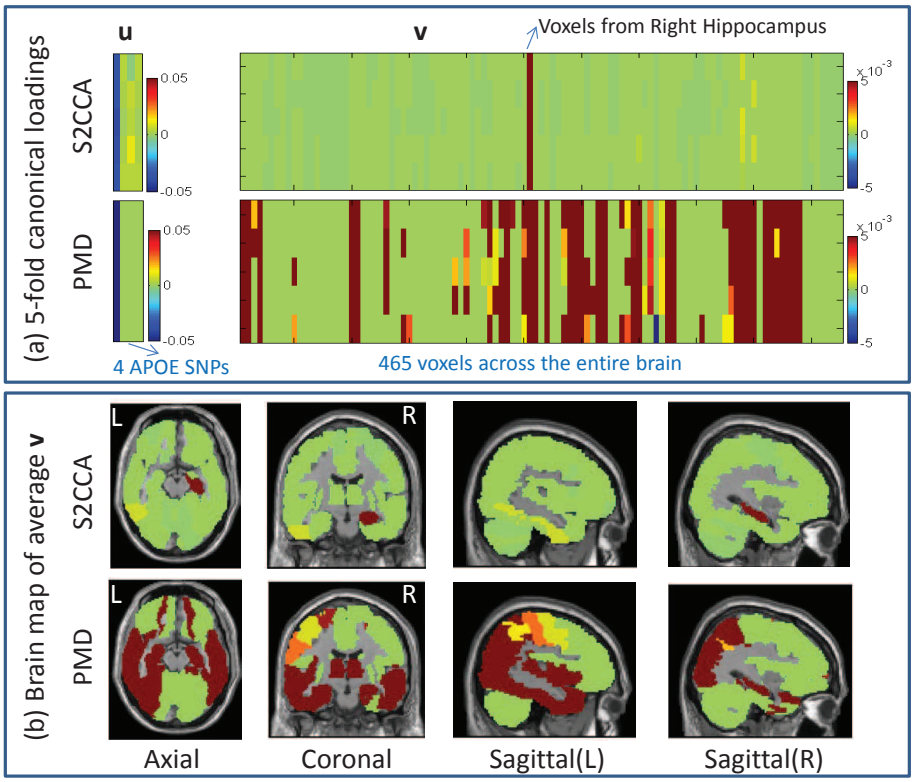


Figure 5.2: Comparison between S2CCA and PMD on identified canonical vectors in cross-validation trials: (a) 5-fold canonical loadings of **u** and **v** on 4 APOE SNPs and 465 VBM measures; (b) mapping the average of imaging canonical loadings **v** of 5 cross-validation trials onto the brain.

true signals from the synthetic data, as well as yield improved canonical correlation performances and biologically meaningful findings from real data. This study is an initial attempt to remove the feature independence assumption many existing SCCA methods have. Since joint multivariate modeling of imaging genetic data is computationally and statistically challenging, we have downsampled our data via a targeted APOE analysis to reduce computational burden and overfitting risk. The S2CCA sparsity has been designed to reduce model complexity and further overcome overfitting. Future directions include evaluating S2CCA using more realistic settings and expanding S2CCA to address efficiency and scalability.

Chapter 6

TRANSCRIPTOME-GUIDED AMYLOID IMAGING GENETICS VIA KG-SCCA

While S2CCA does proved itself with a better performance with its capability of incorporating prior brain structures, its power is still quite limited since its prior is only limited to group structure. Human brain is well known to function more like a complex network system rather than a simple grouping of ROIs. In this chapter, we propose a new knowledge-guided SCCA algorithm (KG-SCCA) to overcome this limitation by incorporating valuable prior knowledge in a more flexible network format. The proposed KG-SCCA method is able to model two types of prior knowledge: one as a group structure (e.g., LD blocks among SNPs) and the other as a network structure (e.g., gene co-expression network among brain regions). The new model incorporates these prior structures by introducing new regularization terms to encourage similarity between grouped or connected features. A new algorithm is designed to solve the KG-SCCA model without imposing the independence constraint on the input features. We demonstrate the effectiveness of our algorithm with both synthetic and real data. For real data, using an Alzheimer's disease (AD) cohort, we examine the imaging genetic associations between all SNPs in the *APOE* gene (i.e., top AD gene) and amyloid deposition measures among cortical regions (i.e., a major AD hallmark). In comparison with a widely used SCCA implementation in the PMA software package (http://cran.r-project.org/web/packages/PMA/) [77], KG-SCCA produces improved cross-validation performances as well as biologically meaningful results.

Table 6.1: Participant characteristics.

Subjects	AD	MCI	HC
Number	28	343	196
Gender(M/F)	18/10	203/140	102/94
Handedness(R/L)	23/5	309/34	178/18
Age(mean±std)	75.23±10.66	71.92±7.47	74.77±5.39
Education(mean±std)	15.61±2.74	15.99±2.75	16.46±2.65

6.1 MATERIALS AND DATA SOURCES

To demonstrate the proposed KG-SCCA algorithm, we apply it to an amyloid imaging genetic analysis in the study of AD. Deposition of amyloid-β in the cerebral cortex is a major hallmark in AD pathogenesis. Our prior studies [47,59] performed univariate genetic association analyses of amyloid measures in a few candidate cortical regions of interest (ROIs), and identified several promising hits including rs429358 in *APOE*, rs509208 in *BCHE*, and rs7551288 in *DHCR24*. In this work, using the proposed KG-SCCA algorithm, we perform a bi-multivariate analysis to examine the association between all the available SNPs (58 in total) in the *APOE* gene (i.e., the top genetic risk factor for late onset AD) and 78 ROIs across the entire cortex. We employ two types of prior knowledge in this analysis: (1) a group structure is imposed to the SNP data using the LD block information (see Fig. 6.4), and (2) a network structure is imposed to the amyloid imaging data by computing an amyloid pathway-based gene co-expression network in the brain using Allen Human Brain Atlas [83]. Below, we first describe our amyloid imaging and genotyping data, and then discuss our method for creating the amyloid pathway-based gene co-expression network in the brain.

6.1.1 IMAGING AND GENOTYPING DATA

The proposed algorithm, KG-SCCA, was empirically evaluated using the amyloid imaging and genotyping data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.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 AD. For up-to-date information, see www.adni-info.org. Pre-processed [18F]Florbetapir PET scans (i.e., amyloid imaging data) were downloaded from LONI (adni.loni.usc.edu). Before downloading, images were averaged, aligned to a standard space, re-sampled to a standard image and voxel size, smoothed to a uniform resolution and normalized to a cerebellar gray matter (GM) reference region resulting in standardized uptake value ratio (SUVR) images as previously described [33]. After downloading, the images were aligned to each participant's same visit MRI scan and normalized to the Montreal Neurological Institute (MNI) space as 2x2x2mm voxels using parameters from the MRI segmentation. ROI level amyloid measurements were further extracted based on the MarsBaR AAL atlas. Genotype data of both ADNI-1 and ADNI-2/GO phases were also obtained from LONI (adni.loni.usc.edu). All the *APOE* SNPs were extracted based on the quality controlled and imputed data combining two phases together. Only SNPs available in Illumina 610Quad and/or OmniExpress arrays were included in the analysis. As a result, we had 58 SNPs located within 10 LD blocks (see Fig. 6.4) computed using HaploView [5]. 568 non-Hispanic Caucasian participants with both complete amyloid measurements and *APOE* SNPs

were studied, including 28 AD, 343 MCI and 196 healthy control (HC) subjects (Table 6.1). Using the regression weights derived from the HC participants, amyloid and SNP measures were preadjusted for removing the effects of the baseline age, gender, education, and handedness.

6.1.2 AMYLOID PATHWAY-BASED GENE CO-EXPRESSION NETWORK IN THE BRAIN

Since we examine cortical amyloid deposition in relation to genetic variation, we hypothesize that amyloid pathway-based gene co-expression profiles among cortical ROIs may provide valuable information in search for *APOE*-related amyloid distribution pattern in the cortex. Thus, we employed the brain transcriptome data from the Allen Human Brain Atlas (AHBA) [83], coupled with 15 candidate genes from amyloid pathways studied in [59], to create such a brain network.

Gene expression profiles across the whole human brain were downloaded from Allen Institute for Brain Science. One of their goals is to advance the research and knowledge about neurobiological conditions, with extensive mapping of whole-genome gene expression throughout the brain. Among various organisms, AHBA is one of the projects seeking to combine the genomics with the neuroanatomy to better understand the connection between genes and brain functioning. Gene expression profiles in 8 health human brains have been released, including 2 full brains and 6 right hemispheres. Details can be found in www.brain-map.org.

Brain-wide expression data of all 15 amyloid-related candidate genes, reported in [59], were extracted from AHBA to construct the brain network. Since an early report indicated that individuals share as much as 95 percent gene expression pro-

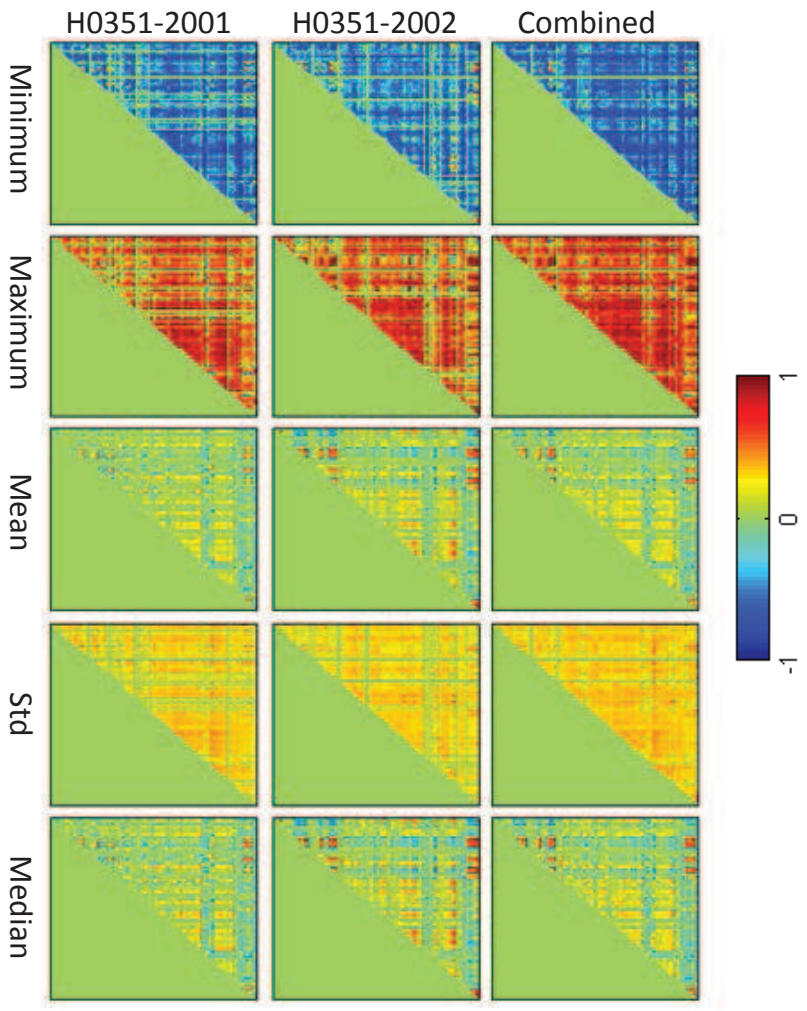


Figure 6.1: Amyloid pathway-based gene co-expression networks among 78 AAL cortical ROIs constructed from AHBA using different statistics (see different rows) for two individuals and their combination.

file [83], in this study, we only included two full brains (H0351-2201 and H0351-2002) to construct the co-expression network. First all the brain samples (∼900) in AHBA were mapped to MarSBAR AAL atlas which included 116 brain ROIs. According to [47], cortical ROIs are typically believed to hold the amyloid signals whereas other ROIs hold similar amyloid measures across individuals. Thus 39 pairs of bilateral cortical ROIs (78 in total), from frontal lobe, cingulate, parietal lobe, temporal lobe, occiptal lobe, insula and sensory-motor cortex, were included in our analysis. Corre-

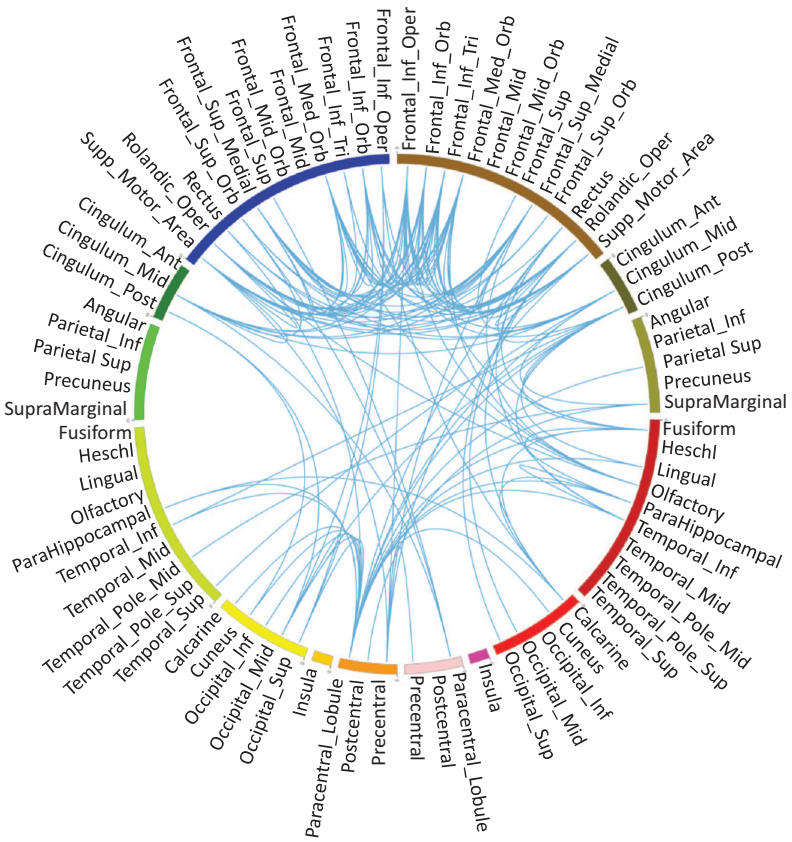


Figure 6.2: Network visualization by thresholding the connectivity matrix shown in the lower right corner of Fig. 6.1, where edges correspond to matrix entries with values ≥ 0.5 or ≤ -0.5. The circle is symmetric (left measures on left and right measures on right), from top to bottom are frontal lobe, cingulate, parietal lobe, temporal lobe, occiptal lobe, insula, and sensory-motor cortex.

lation among ∼900 brain locations were first calculated based on the gene expression profile of 15 amyloid candidate genes. Due to many-to-one mapping from the brain locations to AAL ROIs, for each ROI, there are more than one connections, represented by correlations between two brain locations. Therefore we calculated ROI-level correlations of two individuals in five ways: minimum, maximum, mean, standard deviation and median. In addition, the ROI correlation structure based on the combination of both individuals was also generated in the same way for comparison (see Figure. 6.1). Clearly, for all five statistics, the pattern remains highly consistent across individu-

als and their combination. For simplicity, in the subsequent analysis, we adopt the brain connectivity matrix generated from the combination sample using the median statistics (i.e., the panel in the lower right corner of Figure. 6.1). Figure. 6.2 shows a network visualization of this matrix, where edges correspond to matrix entries with values ≥ 0.5 or ≤ -0.5.

6.2 METHODS

Now we present our KG-SCCA algorithm. We denote vectors as boldface lowercase letters and matrices as boldface uppercase ones. For a given matrix $\mathbf{M} = (m_{ij})$, we denote its i-th row and j-th column as $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. Let $\mathbf{X} = \{\mathrm{x}_1, ..., \mathrm{x}_n\} \subseteq \mathfrak{R}^p$ be the genotype data (SNP) and $\mathbf{Y} = \{\mathrm{y}_1, ..., \mathrm{y}_n\} \subseteq \mathfrak{R}^q$ be the imaging QT data, where n is the number of participants, p and q are the numbers of SNPs and QTs, respectively.

CCA seeks linear transformations of variables $\mathbf{X}$ and $\mathbf{Y}$ to achieve the maximal correlation between $\mathbf{Xu}$ and $\mathbf{Yv}$, which can be formulated as:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \quad s.t. \;\; \mathbf{u}^T\mathbf{X}^T\mathbf{Xu} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Yv} = 1 \tag{6.1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are canonical loadings or weights, reflecting the significance of each feature in the identified canonical correlation.

Similar to many machine learning algorithms, overfitting could arise in CCA when the features outnumber the participants. In addition, CCA outcomes could spread nontrivial effects across all the features rather than only a few significant ones, making the results difficult to interpret. To address these issues, SCCA was proposed in [77]

by introducing penalty terms, $P_1(\mathbf{u}) \leq c_1$ and $P_2(\mathbf{v}) \leq c_2$, to regularize the weights, as shown in **Eq.** (6.2).

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \\ &s.t. \quad ||\mathbf{X}\mathbf{u}||_2^2 = 1, ||\mathbf{Y}\mathbf{v}||_2^2 = 1, P_1(\mathbf{u}) \leq c_1, P_2(\mathbf{v}) \leq c_2 \end{aligned} \tag{6.2}$$

Here the objective function is bi-linear in $\mathbf{u}$ and $\mathbf{v}$: when $\mathbf{u}$ is fixed, it is linear in $\mathbf{v}$ and vice versa. But due to the L_2 equality, with $\mathbf{u}$ or $\mathbf{v}$ fixed, the constraints are not convex. This can be solved by reformulating the L_2 equality into inequality as $||\mathbf{X}\mathbf{u}||_2^2 \leq 1$ and $||\mathbf{Y}\mathbf{v}||_2^2 \leq 1$. For easy computation, **Eq.** (6.2) is commonly rewritten in its Lagrangian form.

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\gamma_1}{2}||\mathbf{X}\mathbf{u}||_2^2 - \frac{\gamma_2}{2}||\mathbf{Y}\mathbf{v}||_2^2 - \beta_1 P_1(\mathbf{u}) - \beta_2 P_2(\mathbf{v}) \tag{6.3}$$

[77] and [78] explored two penalty forms, L_1 penalty and the chain structured fused Lasso penalty. L_1 penalty imposes sparsity on both $\mathbf{u}$ and $\mathbf{v}$, and assumes that each canonical correlation involves only a few features from $\mathbf{X}$ and $\mathbf{Y}$. The fused Lasso penalty promotes the smoothness of weight vectors, and encourages neighboring features to be selected together. To incorporate other structures, group- and network-guided penalties were introduced [11,13]. As mentioned earlier, most of these methods were designed using the soft thresholding technique, which was first proposed to solve Lasso problem when the features were independent from each other [62]. This condition does not hold in imaging genetics data. Thus direct application of those methods into imaging genetics studies limits the capability of yielding optimal solutions. Below, we first present our KG-SCCA model and then present an effective KG-SCCA

algorithm without using the soft thresholding strategy.

Brain has been studied as a complicated network. The SNP data have structures like LD blocks. Given these prior knowledge, we propose the following KG-SCCA model by introducing two penalty terms for genetic loadings $\mathbf{u}$ and imaging loading $\mathbf{v}$ respectively.

$$\begin{aligned} P_1 = ||\mathbf{u}||_G &= \beta_1 \sum_{k_1=1}^{K_1} \sqrt{\sum_{i \in \pi_{k_1}} u_i^2} + \theta_1 ||\mathbf{u}||_1 \\ &= \beta_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2 + \theta_1 ||\mathbf{u}||_1, \end{aligned} \tag{6.4}$$

$$P_2 = ||\mathbf{v}||_N = \beta_2 \sum_{\substack{(i,j) \in E \\ i<j}} \tau(w_{ij}) ||v_i - sign(w_{ij}) v_j||_2^2 + \theta_2 ||\mathbf{v}||_1$$

$$= \beta_2 ||\mathbf{C}\mathbf{v}||_2^2 + \theta_2 ||\mathbf{v}||_1.$$

In penalty $P_1(\mathbf{u})$, SNPs are partitioned into K_1 groups $\Pi_1 = \{\pi_{k_1}\}_{k_1=1}^{K_1}$, such that $\{u_i\}_{i=1}^{m_{k_1}} \in \pi_{k_1}$, and m_{k_1} is the number of SNPs in π_{k_1}. While the group term $\beta_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2$ helps select all the SNPs in relevant LD blocks, L_1 penalty manages to suppress those non-signals within selected LD blocks. The $P_1(\mathbf{u})$ penalty is essentially the group Lasso penalty applied to the CCA framework.

Penalty $P_2(\mathbf{v})$ applies the network-guided constraint to encourage the joint selection of "connected" features (i.e., their connectivity matrix entry having a high weight) as well as uses L_1 to impose global sparsity. E is the set of all possible imaging QT pairs and $|E|$ is the total number of QT pairs. $\mathbf{C} \in \Re^{|\mathbf{E}| \times q}$ is defined as follows. The row of $\mathbf{C}$ is indexed by all pairs $(i,j) \in \{(i,j) | i \in \{1,...,q\}, j \in \{1,...q\}, i<j\}$, $\mathbf{C}_{(i,j),i} = w_{ij}$ and $\mathbf{C}_{(i,j),j} = sign(w_{ij}) w_{ij}$. $\tau(w_{ij})$ provide the fusion effect that promotes similarity between v_i and v_j of related features. In this paper we use $\tau(w_{ij}) = w_{ij}^2$. With $sign(w_{ij})$ we can have positively related features being pulled together and on

the other hand the negatively related features being fused with opposite direction. Thus, for strongly connected features with a large fusion effect, they tend to be jointly selected or jointly not selected.

In this work, as mentioned earlier, we formed the group structure for the SNP data by partitioning them using LD blocks generated by HaploView [5]. We formed the network structure for the amyloid imaging data by constructing amyloid pathway-based gene co-expression network using AHBA. Since the model could be easily extended to estimate multiple canonical variables, we only focus on creating the first pair of canonical variables in this paper.

We now present our algorithm to solve this model without using soft thresholding approach. By fixing $\mathbf{u}$ and $\mathbf{v}$ respectively, we will have two convex problems shown in **Eq.** (6.5).

$$\begin{aligned} &\max_{\mathbf{u}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\gamma_1}{2}||\mathbf{X}\mathbf{u}||_2^2 - \beta_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2 - \theta_1||\mathbf{u}||_1 \\ &\max_{\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\gamma_2}{2}||\mathbf{Y}\mathbf{v}||_2^2 - \frac{\beta_2}{2}||\mathbf{C}\mathbf{v}||_2^2 - \theta_2||\mathbf{v}||_1 \end{aligned} \tag{6.5}$$

Let $\mathbf{B}_1 = \frac{1}{\gamma_1}\mathbf{Y}\mathbf{v}$ and $\mathbf{B}_2 = \frac{1}{\gamma_2}\mathbf{X}\mathbf{u}$, the above problems can be reformulated to **Eq.** (6.6):

$$\begin{aligned} &\min_{\mathbf{u}} \frac{1}{2}||\mathbf{X}\mathbf{u} - \mathbf{B}_1||_2^2 + \frac{\beta_1}{\gamma_1} \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2 + \frac{\theta_1}{\gamma_1}||\mathbf{u}||_1 \\ &\min_{\mathbf{v}} \frac{1}{2}||\mathbf{Y}\mathbf{v} - \mathbf{B}_2||_2^2 + \frac{\beta_2}{2\gamma_2}||\mathbf{C}\mathbf{v}||_2^2 + \frac{\theta_2}{\gamma_2}||\mathbf{v}||_1 \end{aligned} \tag{6.6}$$

Here, while $\mathbf{u}$ can be solved by the G-SMuRFS method proposed in [70], optimization of $\mathbf{v}$ can be achieved by the network-guided $L_{2,1}$ regression method proposed in [80]. In both solutions, a smooth approximation has been estimated for group $L_{2,1}$

Input: $\mathbf{X}$, $\mathbf{Y}$
$\mathbf{X} = \{\mathrm{x}_1, ..., \mathrm{x}_n\}$, $\mathbf{Y} = \{\mathrm{y}_1, ..., \mathrm{y}_n\}$, group and network structures
$t = 1$, Initialize $\mathbf{u}_t \in \Re^{p\times 1}$, $\mathbf{v}_t \in \Re^{q\times 1}$;
while *not converge* **do**
1. Calculate $\mathbf{B}_{1_t} = \frac{1}{\gamma_1}\mathbf{Y}\mathbf{v}_t$;
2. Calculate the block diagonal matrix $\mathbf{D}_{1t}$ and $\mathbf{D}_{2t}$;
3. $\mathbf{u}_{t+1} = (\mathbf{X}^T\mathbf{X} + \frac{\beta_1}{\gamma_1}\mathbf{D}_{1t} + \frac{\theta_1}{\gamma_1}\mathbf{D}_{2t})^{-1}\mathbf{X}^T\mathbf{B}_{1t}$;
4. Scale $\mathbf{u}_{t+1}$ so that $\mathbf{u}_{t+1}^T\mathbf{X}^T\mathbf{X}\mathbf{u}_{t+1} = 1$;
5. Calculate $\mathbf{B}_{2t} = \frac{1}{\gamma_2}\mathbf{X}\mathbf{u}_{t+1}$;
6. Calculate the block diagonal matrix $\mathbf{D}_{4t}$;
7. $\mathbf{v}_{t+1} = (\mathbf{Y}^T\mathbf{Y} + \frac{\beta_2}{\gamma_2}\mathbf{D}_3 + \frac{\theta_2}{\gamma_2}\mathbf{D}_{4_t})^{-1}\mathbf{Y}^T\mathbf{B}_{2t}$;
8. Scale $\mathbf{v}_{t+1}$ so that $\mathbf{v}_{t+1}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v}_{t+1} = 1$;
9. $t = t + 1$.

end
Output: $\mathbf{u}_t \in \Re^{p\times 1}$, $\mathbf{v}_t \in \Re^{q\times 1}$.

Algorithm 3: KG-SCCA algorithm

and L_1 terms by including an extremely small value. The solution for $\mathbf{u}$ and $\mathbf{v}$ in each iteration step is as follows:

$$\begin{aligned}\mathbf{u} &= (\mathbf{X}^T\mathbf{X} + \frac{\beta_1}{\gamma_1}\mathbf{D}_1 + \frac{\theta_1}{\gamma_1}\mathbf{D}_2)^{-1}\mathbf{X}^T\mathbf{B}_1,\\ \mathbf{v} &= (\mathbf{Y}^T\mathbf{Y} + \frac{\beta_2}{\gamma_2}\mathbf{D}_3 + \frac{\theta_2}{\gamma_2}\mathbf{D_4})^{-1}\mathbf{Y}^T\mathbf{B}_2,\end{aligned} \tag{6.7}$$

where $\mathbf{D_1}$ is a block diagonal matrix with the k-th diagonal block as $\frac{1}{\|\mathbf{u}^k\|_F}\mathbf{I}_k$; $\mathbf{I}_k$ is an identity matrix with size of m_k; m_k is the total feature number in group k; $\mathbf{D}_2$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{\|\mathbf{u}^i\|_2}$; $\mathbf{D}_3 = \mathbf{C}^T\mathbf{C}$ is a matrix in which each row integrates all the neighboring relationships (e.g., for the i-th row, it is the sum of all the rows in α whose i-th element is not zero); and $\mathbf{D_4}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{\|\mathbf{v}^i\|_2}$. Algorithm 3 summarizes the KG-SCCA optimization procedure. Further details on how to solve for two objectives in **Eq.** (6.6) are available in [70] and [80] respectively.

In this algorithm, six parameters $\gamma_1, \gamma_2, \beta_1, \beta_2, \theta_1, \theta_2$ need to be tuned to control

Table 6.2: Five-fold cross-validation performance on synthetic data: mean±std is shown for estimated correlation coefficients and AUC of the test data using the trained model. P-value of paired t-test between KG-SCCA and PMA results are also shown.

True	Correlation Coefficients (CC)			Area under ROC (AUC)				
CC	KG-SCCA	PMA	p	KG-SCCA:u	PMA:u	p	KG-SCCA:v	PMA:v
0.60	0.56±0.12	0.31±0.14	2.19E-03	0.83±0.08	0.64±0.02	3.36E-03	1.0±0.00	1.0±0.00
0.64	0.56±0.1	0.51±0.12	2.32E-02	0.96±0.04	0.65±0.01	2.20E-05	1.0±0.00	1.0±0.00
0.70	0.64±0.1	0.53±0.1	1.27E-05	0.99±0.01	0.62±0.	6.21E-08	1.0±0.00	1.0±0.00
0.77	0.7±0.14	0.6±0.14	6.62E-03	0.99±0.01	0.62±0.	9.67E-09	1.0±0.00	1.0±0.00
0.85	0.76±0.08	0.65±0.1	1.02E-04	0.98±0.03	0.63±0.01	4.57E-06	1.0±0.00	1.0±0.00
0.95	0.87±0.04	0.67±0.09	1.19E-03	1.00±0.00	0.63±0.01	1.39E-08	1.0±0.00	1.0±0.00
1.00	0.92±0.04	0.71±0.06	2.46E-04	1.00±0.00	0.64±0.01	4.02E-08	1.0±0.00	1.0±0.00

both the global sparsity and structured group or network constraints. [13] studied a similar problem and found that their results were quite insensitive to γ_1, γ_2 settings. Following their observation, we set γ_1 and γ_2 to 1 for simplicity. Nested cross-validation can be used for parameter selection but will be extremely time consuming for the remaining 4 parameters. Thus, we followed the strategy proposed in [39]: Parameters β_1, β_2 controlling structural constraints were first tuned without considering sparsity constraints. Then based on the obtained optimal β_1, β_2, another nested cross-validation was performed to acquire the optimal θ_1, θ_2.

6.3 EXPERIMENTAL RESULTS AND DISCUSSIONS

We performed comparative studies between the proposed KG-SCCA algorithm and a widely used SCCA implementation in the PMA package (http://cran.r-project.org/web/packages/PI [77]. For PMA experiments, the SCCA parameters were automatically tuned using a permutation scheme provided in PMA. Below we report our empirical results using both synthetic data and real imaging genetics data.

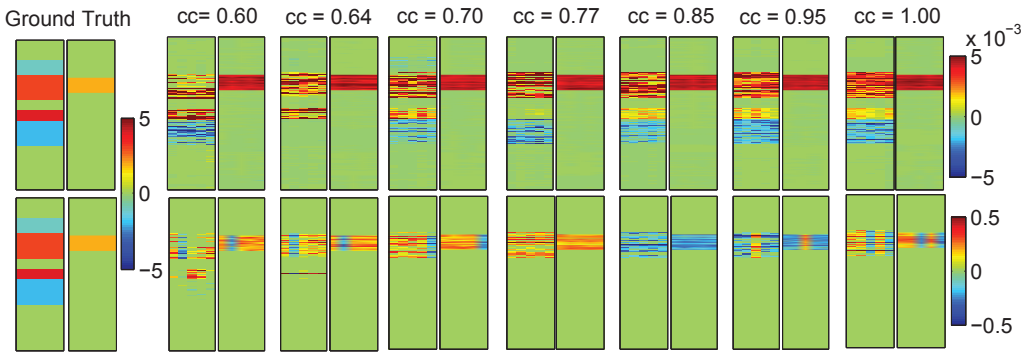


Figure 6.3: Five-fold trained weights of **u** and **v**. Ground truth of **u** and **v** are shown in the most left two panels. KG-SCCA results (top row) and PMA results (bottom row) are shown in the remaining panels, corresponding to true correlation coefficients (CCs) ranging from 0.6 to 1.0. For each panel pair, the five estimated **u** values are shown on the left panel, and the five estimated **v** values are shown on the right panel.

6.3.1 RESULTS ON SIMULATION DATA

Since it is hard to manually construct a data set with a network structure, we simulated group structures for both datasets and then converted them into network structures for one dataset by connecting all the pairs within each group. Synthetic data ($\mathbf{n} = 200, \mathbf{p} = 200, \mathbf{q} = 150$) with diagonal block structure was generated with the following procedure: 1) Random positive definite covariance matrix $\mathbf{M}$ with non-overlapping group structure were created, where correlations range from 0.6 to 1 within group and are set to 0 between groups. 2) Data set $\mathbf{X}$ with covariance structure $\mathbf{M}$ was calculated through Cholesky decomposition. 3) Repeat Steps 1 and 2 to generate another dataset $\mathbf{Y}$. 4) With assigned canonical loadings of $\mathbf{X}$, we calculated the first component $\mathbf{Xu}$. 5) Given a desired correlation between components, we calculated the second component $\mathbf{Yv}$. 6) For simplicity, in this paper, only one group in $\mathbf{Y}$ was assigned to have signals. Therefore, based on predefined canonical loadings of $\mathbf{Y}$ and component $\mathbf{Yv}$, final obtained group signals, added with some white noises

(Signal to Noise Ratio (SNR) =0.5), will replace the data in original dataset $\mathbf{Y}$. By repeating this procedure we generated 7 datasets with correlation levels from 0.6 to 1. The canonical loadings and group structure remained the same across all the datasets.

KG-SCCA and PMA have been both tested on all 7 datasets. All the regularization parameters were optimally tuned using a grid search from 10^{-2} to 10^{2} through nested 5-fold cross-validation. The true and estimated canonical loadings for both $\mathbf{X}$ and $\mathbf{Y}$ were shown in Fig. 6.3. Due to the difference in normalization and optimization procedure, the weights yielded by KG-SCCA and PMA showed different scales. Yet the overall profile of the estimated $\mathbf{u}$ and $\mathbf{v}$ values from KG-SCCA kept consistent with the ground truth across the entire range of tested correlation strengths (from 0.6 to 1.0), whereas PMA was only capable of identifying an incomplete portion of all the signals. Furthermore, we also examined the correlation in the test set computed using the learned models from the training data for both methods. The left part of Table 6.2 demonstrated that KG-SCCA outperformed PMA consistently and significantly, and it could accurately reveal the embedded true correlation even in the test data. The right part of Table 6.2 demonstrated the sensitivity and specificity performance using area under ROC (AUC), where KG-SCCA also significantly outperformed PMA no matter whether the correlation was weak or strong in $\mathbf{u}$. Since $\mathbf{v}$ is relatively simple structured, both KG-SCCA and PMA can restore the signals without any loss. From the above results, it is also observed that KG-SCCA could identify the correlations and signal locations not only more accurately but also more stably.

Table 6.3: Five-fold cross validation results on real data: the models learned from the training data were used to estimate the correlation coefficients between canonical components for both training and testing sets. P-values of paired t-tests were obtained for comparing KG-SCCA and PMA results.

Method		Train						Test					
		f1	f2	f3	f4	f5	mean	f1	f2	f3	f4	f5	mean
KG-SCCA	exp1	0.471	0.448	0.475	0.451	0.46	0.461	0.431	0.515	0.401	0.417	0.459	0.445
	exp2	0.476	0.453	0.454	0.476	0.461	0.464	0.402	0.505	0.503	0.401	0.458	0.454
	exp3	0.476	0.474	0.474	0.468	0.402	0.459	0.408	0.393	0.413	0.435	0.565	0.443
	exp4	0.468	0.466	0.459	0.46	0.466	0.464	0.441	0.409	0.47	0.476	0.445	0.448
	exp5	0.49	0.502	0.434	0.449	0.447	0.464	0.35	0.297	0.584	0.527	0.528	0.457
PMA	exp1	0.439	0.418	0.438	0.438	0.426	0.432	0.368	0.45	0.398	0.379	0.439	0.407
	exp2	0.444	0.416	0.425	0.436	0.432	0.431	0.354	0.463	0.449	0.399	0.416	0.416
	exp3	0.442	0.445	0.439	0.427	0.398	0.43	0.382	0.341	0.382	0.432	0.544	0.416
	exp4	0.434	0.44	0.425	0.427	0.431	0.432	0.414	0.363	0.445	0.438	0.415	0.415
	exp5	0.459	0.462	0.406	0.416	0.411	0.431	0.288	0.287	0.517	0.486	0.501	0.416
						pvalue	3.08E-6					pvalue	8.07E-5

6.3.2 RESULTS ON REAL IMAGING GENETIC DATA

Both KG-SCCA and PMA have been performed on real amyloid imaging and *APOE* genetics data. Similar to previous analysis, 5-fold nested cross-validation was applied to optimally tune the parameters. Five experiments were performed with five different partitions to eliminate the bias. For each single experiment, the same partition was used for both KG-SCCA and PMA. Table 6.3 shows both the training and test performances of KG-SCCA and PMA in all 5 folds of 5 experiments. Both methods demonstrated stable results across five trials. KG-SCCA was observed to outperform the PMA in every single experiment on both training and test performance. Paired t-test was performed to compare the performance across five experiments, and KG-SCCA outperformed PMA significantly in both training (p=3.08E-6) and test cases (p=8.07E-5). We also tested two simplified KG-SCCA models: one with only the penalty term for the LD structure and the other with only the penalty term for the network structure. Interestingly, both performed similarly to the original KG-SCCA, and significantly outperformed PMA.

Fig. 6.4 demonstrates the canonical loadings trained from 5-fold cross-validation in

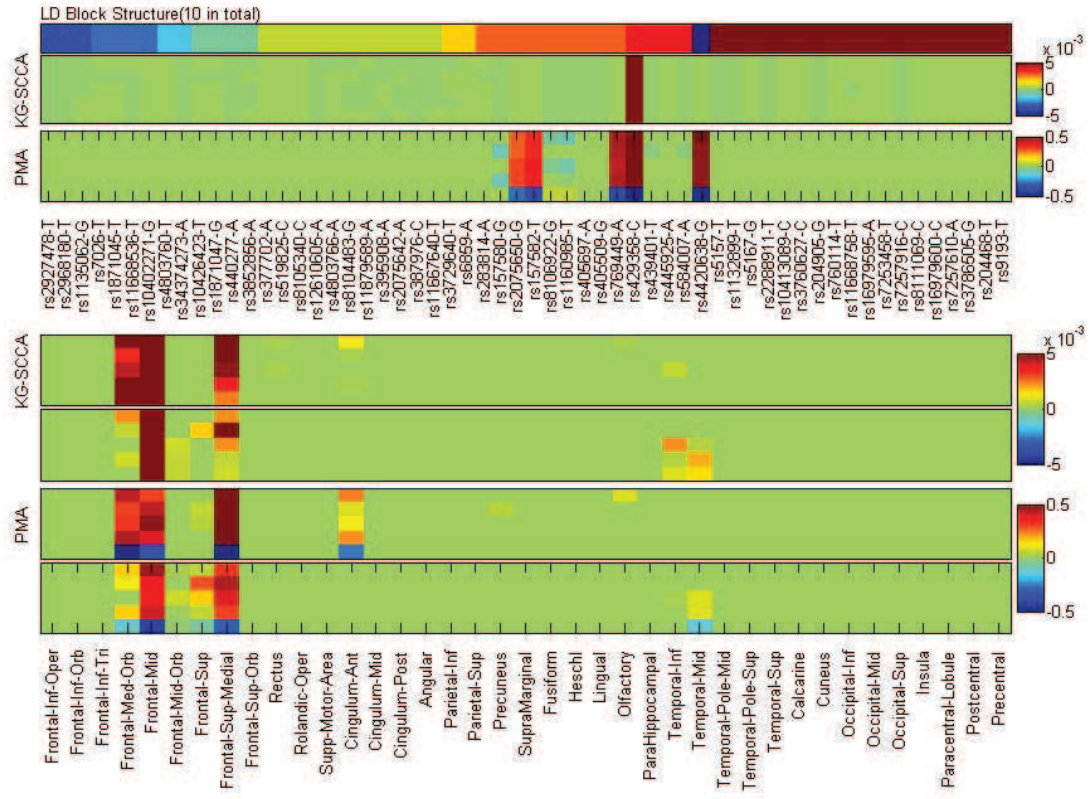


Figure 6.4: Five-fold trained weights of $\mathbf{u}$ (top panel) and $\mathbf{v}$ (bottom panel). KG-SCCA results and PMA results are shown for each panel. For each of KG-SCCA and PMA imaging results (i.e., the bottom panel), the top and bottom rows correspond to left and right hemispheres respectively.

one experiment, suggesting relevant genetic (top panel) and imaging (bottom panel) markers. Although LD block constraints were imposed on relevant SNP markers, L_1 penalty managed to exclude irrelevant signals. Only *APOE* e4 SNP (rs429358) was identified to be associated with amyloid accumulations in the brain. PMA also achieved a similar pattern as KG-SCCA, but including a few additional SNPs from multiple LD blocks. The bottom panel of Fig. 6.4 shows the canonical loading for the imaging data. Both methods identified similar imaging patterns, which are in accordance with prior findings [47]. Fig. 6.5 shows a brain map of canonical loadings generated by KG-SCCA.

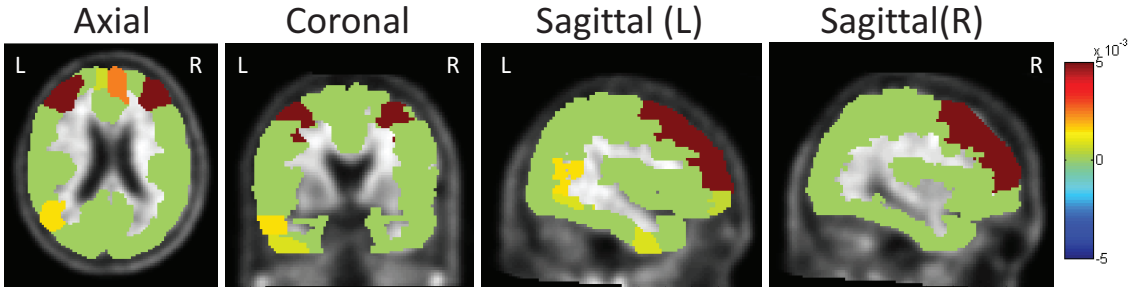


Figure 6.5: Mapping canonical loading generated by KG-SCCA onto the brain.

6.4 CONCLUSION

We have performed a brain imaging genetics study to explore the relationship between brain-wide amyloid accumulation and genetic variations in the *APOE* gene. Since most existing SCCA algorithms are solved using the soft thresholding technique, which assumes independence among data features, direct application of such methods into brain imaging genetics study cannot yield optimal results due to the correlated imaging and genetic features. We proposed a novel knowledge-guided sparse canonical correlation analysis (KG-SCCA) algorithm, which not only removes the above independence assumption, but also can model both the group-like and network-like prior knowledge for improved results. It was compared with a widely used SCCA implementation (PMA) on both synthetic and real data. The empirical results showed that KG-SCCA significantly outperformed PMA in both cases. Furthermore, KG-SCCA accurately recovered the true signals from the synthetic data, and yielded improved performances and biologically meaningful findings from real data. This study is an initial attempt to remove the feature independence assumption many existing SCCA methods have. The empirical studies designed here are targeted to identify relatively clean and simple multi-SNP-multi-QT correlations. Given only 58 SNPs analyzed

here, this work is not a demonstration of a genome-wide analysis. Comparison with other complex SCCA models, building scalable KG-SCCA models, and applications to more complex imaging genetic tasks warrant further investigation.

Chapter 7

DATA INTENSIVE COMPUTING IN BRAIN IMAGING GENETICS STUDY

Although S2CCA and KG-SCCA proposed earlier have yielded promising results, the efficiency and scalability of their implementations still remain as a big concern given the modest sizes of the data sets analyzed in these studies. In particular, a standard practice to evaluate the significance of the results is to use permutation for computing a p-value [77]. This requires to run the same test on permuted data sets many times, and a large N (e.g., $\geq$ 1,000) is often needed to provide a reasonable estimate of the empirical null distribution. For example, in our experiments, it takes more than 1,200 hours to run an N=1,000 SCCA permutation test for analyzing an imaging genetic data set with 1,000 participants, 3,200 SNPs and 10,000 voxels. With priors included, S2CCA and KG-SCCA will take even longer. Therefore it is an urgent need to develop new concepts and enabling tools and offer a more efficient solution.

Recent development of hardware and software enables massive parallelism with existing domain-specific analysis with minimal to no modification. In this chapter we develop and evaluate a set of acceleration strategies to speed up a widely used SCCA implementation, which is provided by the PMD[1] software package [77]. Using several simulated imaging genetics data sets, we perform an empirical comparison between the existing solution and the accelerated one that combines parallel data strategy and the offload model for Intel Many Integrated Core (MIC).The empirical results demonstrate that our approach can achieve 2-fold speedup for SCCA algorithm.

[1] http://cran.r-project.org/web/packages/PMA/

7.1 SCCA FOR IMAGING GENETICS

We denote vectors as boldface lowercase letters and matrices as boldface uppercase ones. For a given matrix $\mathbf{M} = (m_{ij})$, we denote its i-th row and j-th column to $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. Let $\mathbf{X} = \{\mathrm{x}_1, ..., \mathrm{x}_n\} \subseteq \Re^p$ be the SNP data and $\mathbf{Y} = \{\mathrm{y}_1, ..., \mathrm{y}_n\} \subseteq \Re^q$ be the imaging QT data, where n is the number of participants, p and q are the number of SNPs and QTs, respectively.

As mentioned in Chapter 2, CCA seeks linear combinations of variables in $\mathbf{X}$ and variables in $\mathbf{Y}$, which are maximally correlated between $\mathbf{Xu}$ and $\mathbf{Yv}$, that is:

$$\begin{aligned} &\arg\max_{\mathbf{u},\mathbf{v}} \sum_{i=1}^{n} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \\ &\text{subject to} \quad \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v} = 1 \end{aligned} \tag{7.1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are canonical vectors or weights.

Two major weaknesses of CCA are that it requires the number of observations n to exceed the combined dimension of $\mathbf{X}$ and $\mathbf{Y}$ and that it produces nonsparse $\mathbf{u}$ and $\mathbf{v}$ which are difficult to interpret. SCCA removes these weaknesses by maximizing the correlation between $\mathbf{Xu}$ and $\mathbf{Yv}$ subject to the weight vector constraints $P_1(\mathbf{u}) \leq c_1$ and $P_2(\mathbf{v}) \leq c_2$. The penalized matrix decomposition (PMD) toolkit [77] provided a widely used SCCA implementation, where the L_1 penalty $P(A) = \sum_{k=1}^{p} |A(k)|$ was used for both P_1 and P_2.

For simplicity, this algorithm assumed $\mathbf{X^T X} = \mathbf{I}$ and $\mathbf{Y^T Y} = \mathbf{I}$, and implemented

SCCA by alternately performing the following two steps until convergence.

$$1. \quad \mathbf{u} \leftarrow \arg\max_{\mathbf{u}} \mathbf{u^T X^T Y v},$$
$$\text{subject to } \|\mathbf{u}\|^{\mathbf{2}} \leq \mathbf{1}, P_1(\mathbf{u}) \leq c_1$$

$$2. \quad \mathbf{v} \leftarrow \arg\max_{\mathbf{v}} \mathbf{u^T X^T Y v},$$
$$\text{subject to } \|\mathbf{v}\|^{\mathbf{2}} \leq \mathbf{1}, P_2(\mathbf{v}) \leq c_2$$

where P_1 and P_2 are the L_1 penalty functions to yield $\mathbf{u}$ and $\mathbf{v}$ sparse. The first update takes the form

$$\mathbf{u} \leftarrow \frac{S(\mathbf{X^T Y v}, \Delta)}{\|S(\mathbf{X^T Y v}, \Delta)\|_2},$$

where $S(x, \Delta) = sgn(x)(|x| - \Delta)_+$ is the soft thresholding operator and $\Delta \geq 0$ is chosen so that $P_1(\mathbf{u}) = c_1$. The second update takes a similar form by swapping (1) $\mathbf{X}$ and $\mathbf{Y}$, (2) $\mathbf{u}$ and $\mathbf{v}$, (3) P_1 and P_2, and (4) c_1 and c_2.

7.2 ACCELERATING SCCA AT XSEDE

Although R is adopted as a "high productivity" language in SCCA, high performance has not been a development goal of R. Designed as a computing language with high level expressiveness, R lacks much of the fine grained control and basic constructs to support highly efficient code development. For example, most features in the existing SCCA are implemented as single thread processes. As mentioned before, a typical standard practice to evaluate the significance of the SCCA results often require us to run the same SCCA test on permuted data sets many times, and a large

N (e.g., $\geq$1,000) is often needed to provide a reasonable estimate of the empirical null distribution. In addition, cross-validation is often used to optimally tune the two SCCA parameters. For example, using a 10-fold cross-validation coupled with an 11-by-11 grid search strategy (i.e., 11 possible values for each of the two parameters), we need to run SCCA $10 \times 11 \times 11 = 1,210$ times. All these will significantly increase the computational needs and workload of existing SCCA solutions.

In this section, we propose and evaluate approaches to accelerate SCCA implemented in R. The computation of SCCA involve significant amount of linear algebra and matrix computation. Our basic idea of the acceleration is to offload the computationally intensive calls to optimized mathematical library (e.g., the Intel Math Kernel Library, or MKL). While MKL can automatically manages the computing details, we can pursue even better performance improvement by distributing the work across the compute host and the many-integrate-core (MIC).

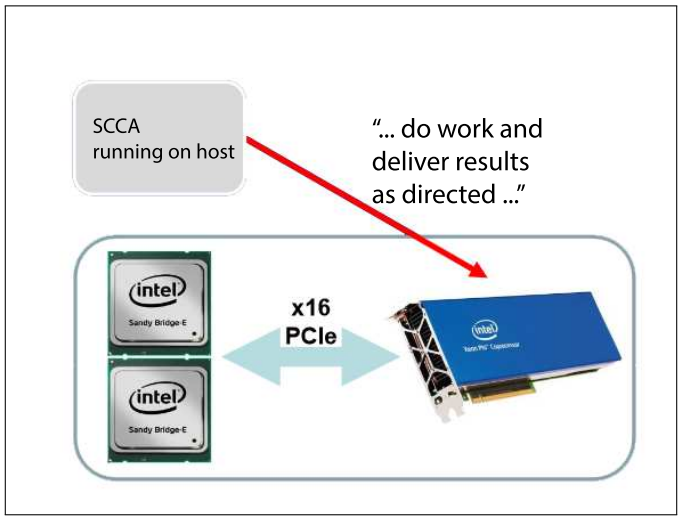


Figure 7.1: Adopting offload model on Stampede cluster at XSEDE: a SCCA program running on the host can "offload" work by directing the MIC to execute a specified block of code. The host also directs the exchange of data between host and MIC. Ideally, the host stays active while the MIC coprocessor does its assigned work.

7.2.1 AVAILABLE ACCELERATION STRATEGIES

LINKING R AND OPTIMIZED MATHEMATICAL LIBRARIES

R can be linked to other shared mathematics libraries to speed up many basic computation tasks. One option for linear algebra computation is to use Intel Math Kernel Library (*MKL*) [32]. *MKL* includes a wealth of routines to accelerate application performance and reduce development time such as highly vectorized and threaded linear algebra, fast fourier transforms (FFT), vector math and statistics functions. Furthermore, the *MKL* has been optimized to utilize multiple processing cores, wider vector units and more varied architectures available in a high end system. Different from using parallel packages, *MKL* can provide parallelism transparently and speed up programs with supported math routines without changing code. It has been reported that the compiling R with *MKL* can provide three times improvements out of box [28].

EXPLOITING ACCELERATOR CARDS

Significant efforts have been made in developing accelerator cards that can easily increase the parallel processing potential in recent years. General purpose graphic processing units (*GPGPU*) extends parallel functions and technologies traditionally embedded in graphic processing units to handle more generic computations. Computational solutions can utilize the parallel features provided by *GPU* through programing interface such as *OPENCL* and *CUDA*. Most recently, the Intel Xeon Phi SE10P Co-processor (Xeon Phi) integrate 60 processing cores and 8GB memory in a single card. A critical advantage of the Xeon Phi co-processor is that, unlike GPU-

based co-processors, the processing cores run the Intel x86 instruction set (with 64-bit extensions), allowing the use of familiar programming models, software, and tools. In addition to allowing the host system to offload computing workload partially to the Xeon Phi, it also can run a compatible program independently.

7.2.2 ACCELERATING SCCA WITH MKL AND OFFLOAD MODEL

Our first objective is to investigate the benefit of using MIC and offload model on MIC at XSEDE resources. We first tested using the R-25 benchmark script[2]. The testing script includes fifteen common computational tasks grouped into three categories: *Matrix Calculation*, *Matrix functionc* and *Programmation*. The twelve tasks are listed in Table 7.1. The test was performed in a high performance compute environment, we used the Texas Advanced Computing Center Stampede cluster. Stampede provides several different techniques for achieving higher performance computations which include using its Xeon Phi accelerators and/or NVIDIA Kepler 20 GPUs for large matrix calculations. In this test, each compute node has two Intel Xeon E5-2680 processors each of which has eight computing cores running @2.7GHz. There is 32GB DDR3 memory in each node for the host CPUs. The Xeon Phi SE10P Coprocessor installed on each compute node has 61 cores with 8GB GDDR5 dedicated memory connected by an x16 PCIe bus. The NVIDIA K20 GPUs on each node have 5GB of on-board GDDR5. All compute nodes are running CentOS 6.3. For this study we used the stock R 3.01 package compiled with the Intel compilers (v.13) and built with Math Kernel Library (MKL v.11).

Based on our observation of significant performance improvement of benchmark

[2]http://r.research.att.com/benchmarks/

Table 7.1: Translation of benchmark number to R-25 benchmark description for all R-25 plots.

#	R25 Benchmark Task Description
1	Creation, transp., deformation of a 2500 × 2500 matrix (sec)
2	2400 × 2400 normal distributed random matrix
3	Sorting of 7,000,000 random values
4	2800 × 2800 cross-product matrix
5	FFT over 2,400,000 random values
6	Eigenvalues of a 640 × 640 random matrix
7	Determinant of a 2500 × 2500 random matrix
8	3,500,000 Fibonacci numbers calculation (vector calc)
9	Creation of a 3000 × 3000 Hilbert matrix (matrix calc)
10	Grand common divisors of 400,000 pairs (recursion)
11	Creation of a 500 × 500 Toeplitz matrix (loops)
12	Escoufier's method on a 45 × 45 matrix (mixed)

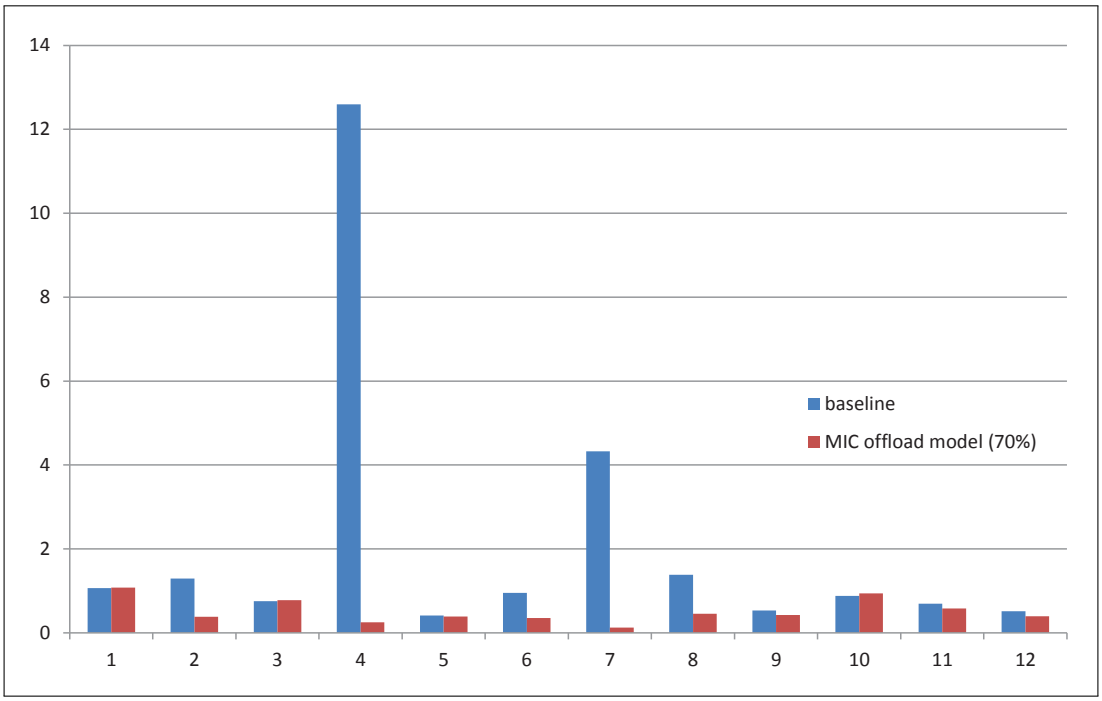


Figure 7.2: Basic vectorized and matrixed operations can be obtained significant speed-up with MLK and offload on MIC.

version of R computation using MKL and offload model (see e.g., Figure 7.2), we compile SCCA to run with the same acceleration strategy and on the same resources. By choosing work-sharing at the 30% host (16 threads) 70% coprocessor (240 threads) sweet spot (see e.g., [23]), we achieve consistent 2-fold speedup when running SCCA algorithm over various input sizes. We note that this acceleration or speedup involves *NO* code changes or further code optimizations for SCCA implementation in R (e.g., vectorization , which could potentially improve the performance even better.)

7.3 SCCA FOR LARGE BRAIN IMAGING GENETICS DATA

In our experiments, we created several simulated imaging genetics data sets of different sizes and used these synthetic data to perform comparative study. Although SCCA's performance can be improved with our aforementioned massively parallelism strategy, the scalability of applying SCCA to large brain imaging genetics data has not been completely explored yet. As mentioned before, an N=1,000 permutation test for evaluating the significance of an SCCA result (i.e., calculating a p-value) requires to run SCCA test on permuted data sets 1,000 times. A 10-fold cross-validation method with an 11-by-11 grid to optimally search for two SCCA parameters requires to run SCCA 1,210 times. If one wants to not only tune the parameters using cross-validation but also calculate a p-value using cross-validation, the total number of SCCA tests could be more than 1 millioin (i.e., $1,000 \times 1,210$). In cases like this, the processing tasks are beyond the computing capability of a local workstation, and explicit parallelization strategies are often desired when we scale to large data sets or complicated computational tasks.

7.3.1 SIMULATED IMAGING GENETICS DATA

To evaluate the performances of the existing and accelerated SCCA implementations, we developed a method to create realistic imaging genetics data with known underlying correlation structures. First, the synthetic genotype data was generated through FREGENE genome simulator ([9,31]), which is aimed to simulate sequence-like data over large genomic regions in large diploid populations. In this study, we generated N=1,000 diploid individuals over 20,000 generations with a 10 Mb genome with the average mutation rate as 2.5×10^{-8} per site per generation. Among all acquired SNPs, 3,274 SNPs with minor allele frequency (MAF) greater than 0.05 were extracted and included in our analysis. Figure 7.3 shows the correlation structure of the simulated genotype data. We formed four SNP data sets (i.e., g500, g1000, g2000, and g3274) by taking the first 500, 1,000, 2,000, and 3,274 SNPs from the entire data, respectively. The inset shows an enlarged view of the first four linkage disequilibrium blocks, to which we introduced the imaging genetic association discussed below (see also Figure 7.7(a-b)).

To create simulated imaging data, we assume that the image consists of multiple regions of interest (ROIs) and the image voxel values within each ROI are highly correlated with each other. Thus, we created a random positive definite non-overlapping group structured covariance matrix $\mathbf{M}$, where each group (of voxels) corresponds to an ROI. After that, we employed Cholesky decomposition to obtain the background imaging data (i.e., N=1,000 images) with covariance structure $\mathbf{M}$. Since the background imaging data was randomly drawn from a Gaussian distribution with a specified structure. Therefore it was reasonable to assume there was no relationship

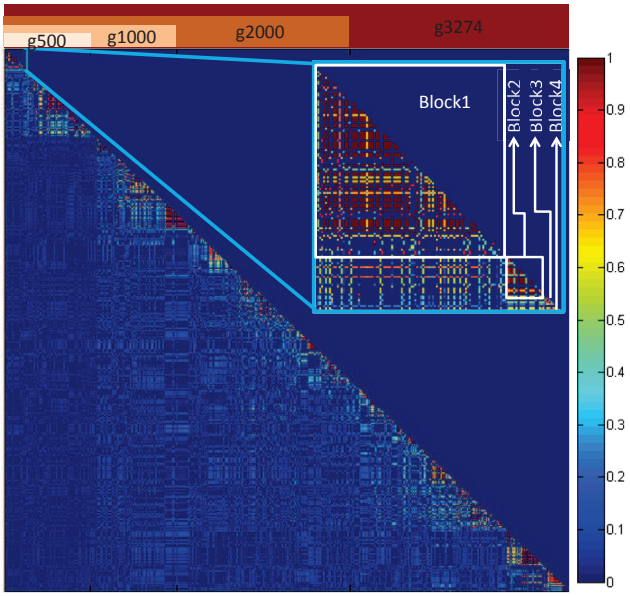


Figure 7.3: Correlation structure of the simulated genotype data. We formed four SNP data sets (i.e., g500, g1000, g2000, and g3274) by taking the first 500, 1,000, 2,000, and 3,274 SNPs from the entire data, respectively. The inset shows an enlarged view of the first four linkage disequilibrium blocks, to which we introduced the imaging genetic association (see also Figure 7.7(a-b)).

between the simulated genotype data and the background imaging data. We created three sets of phenotypic imaging data (i.e., p1000, p5000, and p10000), consisting of 1,000, 5,000 and 10,000 voxels respectively.

The imaging genetic association was introduced using the following steps. (1) The Haploview software ([6,7]) was used to identify the linkage disequilibrium (LD) block information of the simulated SNP data and partition SNPs into groups (i.e., each LD block forms a group, see Figure 7.3 inset for the first four LD blocks). (2) Canonical loadings $\mathbf{u}$ and $\mathbf{v}$ were set based on the group structures of $\mathbf{X}$ and $\mathbf{Y}$ respectively, where all the variables within a group share the same weights. In this initial study, for simplicity, we selected only one group in $\mathbf{Y}$ (i.e., imaging data) to be associated with 4 groups in $\mathbf{X}$ (i.e., SNP data). (3) The portion of the specified group in $\mathbf{Y}$ were replaced based on the $\mathbf{u}$, $\mathbf{v}$, $\mathbf{X}$ and the assigned correlation.

Figure 7.4 shows six example images. Each consists of 5 ROIs. The intensity values within each ROI are the same. All the image voxels except those in these ROIs are white noises. The imaging genetic association (with correlation coefficient equal to 1) exists only between ROI1 and the first four LD blocks in the SNP data (see also Figure 7.3 and Figure 7.7(a-b)). Figure 7.7(a-b) shows the details of this imaging genetic association by plotting the canonical vectors for both imaging and genetic data.

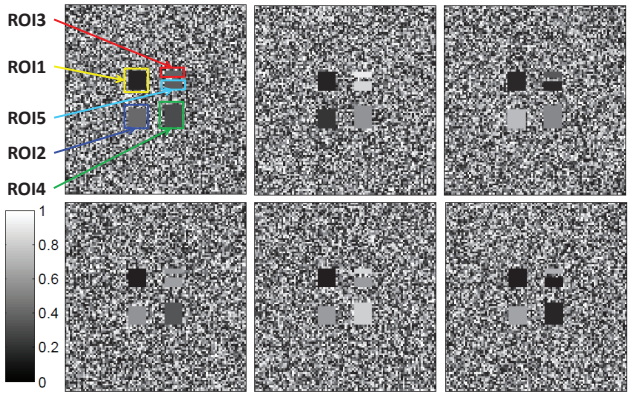
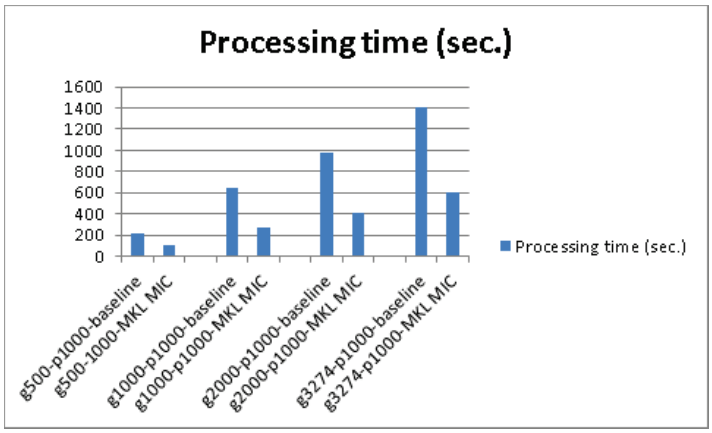


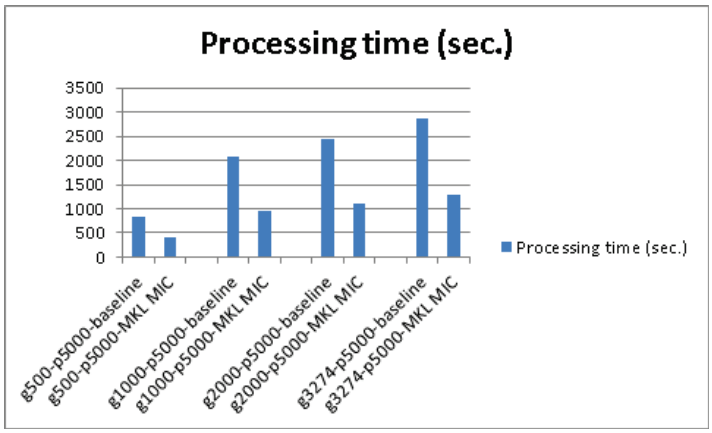
Figure 7.4: Example image data. Each image consists of 5 ROIs. The intensity values within each ROI are the same. All the image voxels except those in these ROIs are white noises. The imaging genetic association exists only between ROI1 and the first four LD blocks in the SNP data (see also Figure 7.3 and Figure 7.7(a-b)).

7.3.2 SCALING TO LARGE DATASETS WITH DATA PARALLEL STRATEGY

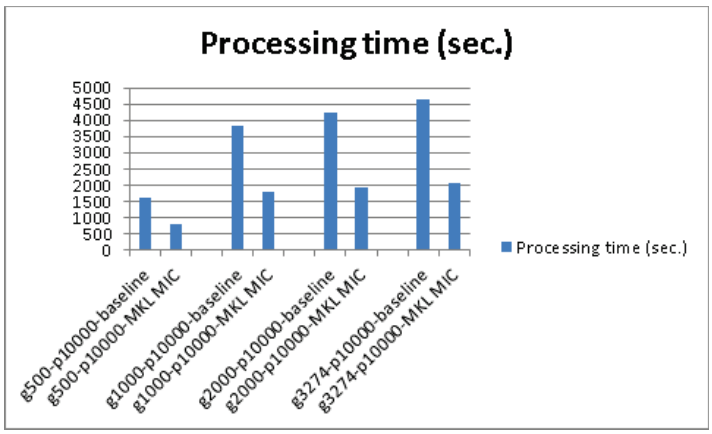
There are nearly 30 packages that are related in enabling parallelism listed in CRAN task view for high performance computing. Among them, some are designed to provide explicit parallelism where users control the parallelization (such as *Rmpi* and *snow*); some are specially designed to provide implicit parallelism so that the system can abstract parallelization away (such as *multicore*); others are high level wrapper



(a)



(b)



(c)

Figure 7.5: Comparison of SCCA speed-up for different combinations of the genotype data sets (g500, g1000, g2000, and g3274) and phenotype data sets (p1000, p5000, and p10000): a consistent 2-fold speedup has been achieved in all the situations.

for other packages and intended to ease the use of parallelism, such as *snowfall* and *foreach*. Here we only reviewed some of the major packages that are directly related to our investigation. *Rmpi* is one of the earliest parallel package developed for R and is still used today and is built upon by other packages [81]. *Rmpi* provides an interface between R and Message Passing Interface and can link to an existing MPI implementation. The users need to link the R package with a MPI library installed separately, then the package enables users to use mpi-like code in R scripts. The package also includes parallel implementations of apply-like functions. The *snow* package utilizes *Rmpi* and several other existing parallel packages to expand the parallel support through a simple interface [50]. There are also several packages for exploiting parallelism within a single compute node. Fork is based on the system processing management interface to generate additional threads for computations [81]. *Pnmath* uses the *OPENMP* to implement many common mathematic functions to run in parallel. R/parallel provides support for running loops in parallel using a master-slave model [37]. *multicore* package has been developed for utilize multiple cores available on the system. In addition, there are projects related with big data but not directly compared here, e.g. *pbdR*, *Rhadoop* etc. For a more comprehensive reviews of the parallel packages, interested reader can refer [21, 54].

Our experiments included running the existing and accelerated SCCA implementations over multiple genotype data sets (i.e., g500, g1000, g2000, and g3274) and phenotypes data sets (i.e., p1000, p5000, and p10000). The experiment can be basically considered a 4×3 grid, with each cell analyzing the relationship between the simulated genotype data and the imaging data. Our basic solution was to couple the SCCA acceleration supported by MKL and Intel Xeon Phi processor with explicit

Table 7.2: A set of 4 × 3 experiments running SCCA analytics over combinations of 4 genotype data sets (g500, g1000, g2000, and g3274) and 3 phenotype data sets (p1000, p5000, and p10000). The table lists the processing time required to run each of the 12 experiments, in a serial fashion that costs a total processing time of 25804.48 seconds (or around 7.5 hours).

	p1000	p5000	p10000
g500	215.026	842.058	1621.942
g1000	644.073	2088.001	3835.77
g2000	947.037	2448.248	4234.645
g3274	1403.307	2878.28	4646.093
Total	25804.48		

Table 7.3: The aforementioned 4 × 3 experiments running improved SCCA analytics with data parallel approach. The table lists the processing time required to run each of the 12 experiments with SCCA employing MKL and offload model on MIC,in an embarrassingly parallel model that reduces the total processing time down to 2232.2 seconds (or around half hour).

	p1000	p5000	p10000
g500	100.408	418.441	799.086
g1000	264.943	968.705	1782.067
g2000	413.804	1109.941	1945.346
g3274	600.598	1275.952	2082.772
Total	2232.2		

parallel packages. We processed the 4 × 3 task grid in parallel using Snowfall package. Table 7.2 shows the individual running time of each cell as well as the estimated total running time as the sum of all 12 cells (7.5 hours). Using Snowfall package (sfLapply) with MKL and offload model on each compute node, we could accelerate our analysis tasks by distributing the 12 task cell to multiple compute nodes. Table 7.3 shows the improvement: each individual cell task obtained its own 2-fold speedup, and more importantly, since these were all executed and process in parallel, we could reduce the total processing time down to 0.5 hour, or a 15-fold speedup. Figure 7.5 shows a graphical representation of the performance comparison.

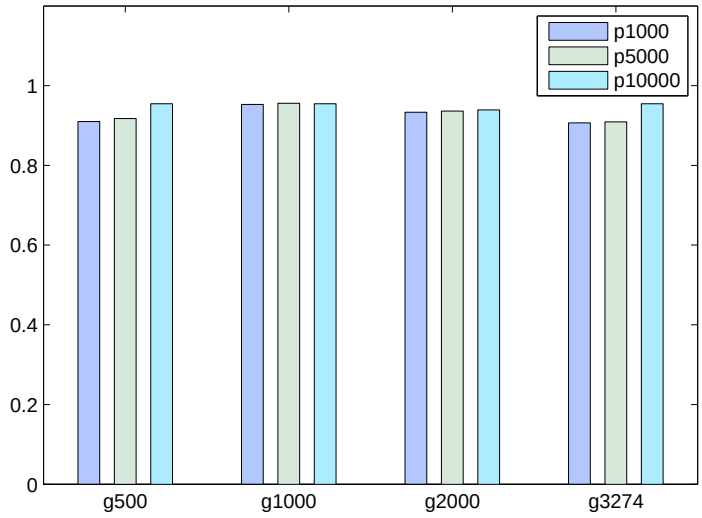


Figure 7.6: Correlation coefficients

The accelerated implementation yielded the same SCCA result as the existing implementation for each combination of the genotype and phenotype data. Figure 7.6 shows the resulting correlation coefficients for all the cases, which are very close to the ground truth value of 1. Figure 7.7(c-d) shows the identified canonical vectors for analyzing g500 and p1000. Compared with the ground truth shown in Figure 7.7(a-b), SCCA identified most of the signals but ignored some. It is currently an active research topic to develop improved SCCA algorithms that can yield more accurate results. Accelerating these new SCCA algorithms is an interesting future direction to pursue.

7.4 CONCLUSION

We have presented a set of massively parallel strategies to accelerate a widely used sparse canonical correlation analysis (SCCA) implementation provided by the PMD software package. In particular, we have exploited parallel packages of R, optimized

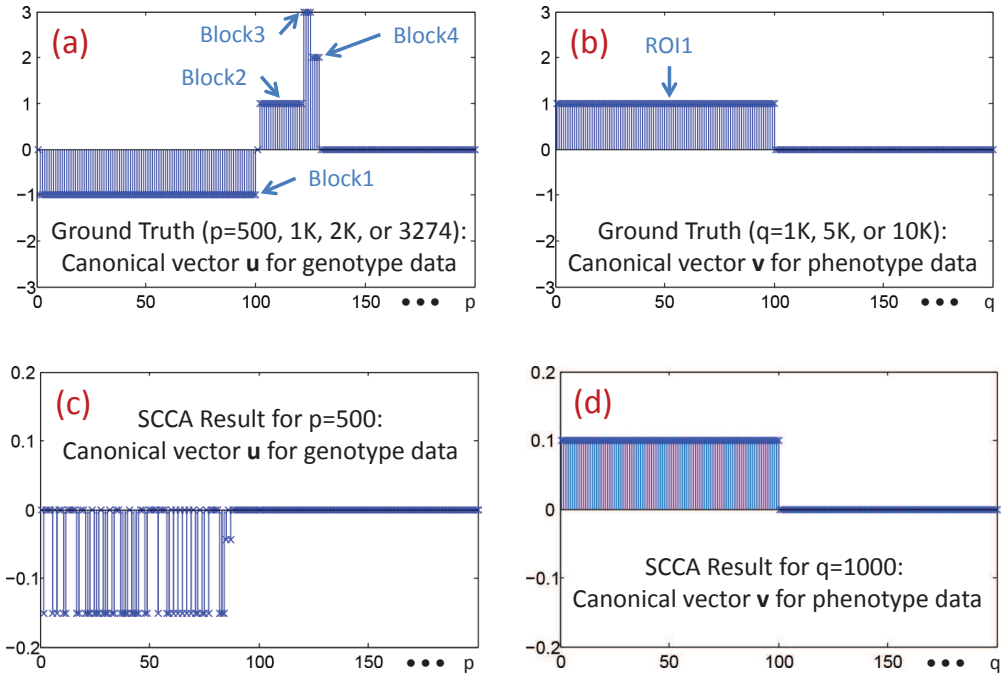


Figure 7.7: (a-b) Ground truth of canonical vectors u for genotype data (p=500, 1000, 2000, or 3274) and v for phenotype data (q=1000, 5000, or 10000). See Figure 7.3 for block1-block4 and Figure 7.4 for ROI1. (c-d) Canonical vectors u and v identified by applying SCCA to g500 and p1000 (i.e., p=500 and q=1000). Note that the canonical vectors discovered by SCCA had a different scale from the ground truth, since SCCA applied a normalization step to the data before performing the actual analysis.

mathematical libraries, and the automatic offload model for Intel Many Integrated Core (MIC) architecture to accelerate SCCA. We have created several simulated imaging genetics data sets of different sizes and used these synthetic data to perform comparative study. Our performance evaluation demonstrates that a 2-fold speedup can be achieved by the proposed acceleration. These preliminary results show that by combining data parallel strategy and the offload model for MIC we could significantly reduce the knowledge discovery timelines involving applying SCCA on large brain imaging genetics data.

Chapter 8

CONCLUSIONS

In this final chapter, we summarize the contributions of this thesis and discuss ideas for future work.

8.1 SUMMARY

In this thesis, we study prior knowledge guided regression and association modeling techniques as well as their application in disease biomarker discovery and genetics mechanism study. While traditional methodologies often ignore the highly correlated nature of current biomedical datasets, the main contributions of this thesis involve extensions and applications of existing predictive and associative models, and are summarized as follows. All the research work conducted has been widely recognized and published in several premier journals and conference proceedings (9 as first author and 9 as co-author), listed as in curriculum vitae section.

Biomarker discovery: Based on a newly developed sparse multi-task learning algorithm called G-SMuRFS, this thesis first investigated the power of intermediate level cortical thickness measures toward the cognitive outcomes. Compared to traditional ROI level measurements, this work performed the k-mean clustering within each ROI to collect the up-scaled yet still computationally affordable measures. And prediction analysis based on these intermediate measures gave us a better understanding of localized brain signals that are related to cognitive outcomes.

After that, considering the complexity of human brain, we further proposed a new network-guided multivariate model NG-L21 to flexibly model prior structure

among predictors. Unlike traditional methods, this model could provide advantages in several folds: (1) explicitly incorporating the relationships among predictors in a more general way, (2) using data-driven patterns without any predefined parameters, (3) effectively identifying biomarkers influencing multiple responses, and (4) selection of correlated markers together rather than picking only one of them to improve the stability. With the application to the ADNI multimodal data (predicting memory scores from MRI and CSF proteomic measures), NG-L21 demonstrated improved prediction performance over the state-of-the-art competing methods, with stable and meaningful multimodal biomarkers.

Genetic mechanism exploration: To investigate the complex imaging genetics associations, we proposed two structure-aware sparse association models: (1) group structure guided SCCA (S2CCA), and (2) knowledge guided SCCA (KG-SCCA).

In S2CCA, we not only removed the independence assumption, but also took into consideration the group-like structure in the data features. With comparison to traditional SCCA, S2CCA was validated with significantly better performance than SCCA on both synthetic and real datasets. In addition, S2CCA could accurately recover the true signals from the synthetic data with improved canonical correlation performances and biologically meaningful findings from real data. This study is the first attempt to remove the feature independence assumption in many existing SCCA based methods.

KG-SCCA is generally an extension of S2CCA, where prior knowledge of brain can be included as a more flexible network format, and S2CCA becomes a special case of KG-SCCA. Based on this model, we performed a brain imaging genetics study to explore the relationship between brain-wide amyloid accumulation and genetic

variations in *APOE* gene, where transcriptome information was utilized to build the prior co-expression brain network. On both synthetic and real datasets, KG-SCCA significantly outperformed traditional methods. Similarly as in S2CCA, we found that network prior knowledge can also better recover the true signals from the synthetic data, with more biologically meaningful findings identified from real data.

Finally, to better promote the application of these advanced models in high dimensional datasets, we made our initial efforts to introduce the parallelization framework. In particular, we exploited parallel packages of R, optimized mathematical libraries, and the automatic offload model for Intel Many Integrated Core (MIC) architecture to accelerate the traditional SCCA. Based on several simulated imaging genetics data sets of different sizes, we observed at least 2-fold speedup for most tasks through the proposed acceleration. And together with parallel strategy, we could significantly reduce the knowledge discovery timeline for application of advanced models to large brain imaging genetics data.

8.2 FUTURE DIRECTIONS OF RESEARCH

This thesis provides a basis for us to continue to pursue research in the area of predictive modeling and association analysis, which, we believe, has a host of fundamental problems yet to be solved, especially for large scale and highly correlated biomedical datasets like brain imaging and genome sequencing data. In this section, we describe a few promising future directions of research that are enabled by the approach presented in this thesis.

Sparse learning models with dynamic data structures: This thesis has well exploited the structured sparse modeling in both predictive and associative analysis.

But one problem exists that there are usually more than one prior structures for even one single data modality. For example, brain network can be constructed either based on correlation or gene co-expression patterns, or many other ways. It still remains unknown how these prior structures make the impact on the prediction procedure and therefore explicit incorporation of one specific structure may not be the optimal solution. Future efforts will be made to further explore and compare more complicated data structures hidden in large-scale, dynamic, longitudinal and heterogeneous data, and to examine their potential role in guiding the learning procedure. Also, it is also of great interest to examine whether the predictive power of one type of measures would be boosted by specific prior knowledge. Another direction worth our effort is to develop some new sparse learning models that can automatically learn the hidden structures rather than taking them as priors.

High-order imaging and genetic features: Despite substantial effort that has been made in AD study, there is very limited progress in discovery of novel biomarkers and underlying mechanisms. we believe that additional signals may not be directly represented by these raw measures, but rather reside within them. With the expanding data pool with growing varieties, we are now able to capture novel features and seek another path to the underlying disease mechanisms. Instead of using those imaging and genomic data as it is, it would be of more interest to perform feature extraction analyses through existing tools. These high-order features will not only provide more deep insights but also enable and promote the expansion of many other research areas, such as previously mentioned predictive, associative study and GWAS of quantitative-traits.

Big data analytics: As shown in Chapter 7, big data framework holds great

promise to promote the application of machine learning in various biomedical problems. While many existing studies are still subject to the curse of dimensionality when attempting to expand the search space to the genome wide and brain wide scales, another future work worth our efforts will be big data analytics. While Map/Reduce framework can facilitate the parallelization of codes and large scale genetic interaction analyses, Giraph framework, designed specifically for graph and network analysis, will help promote our high-order feature exploration from high dimensional brain networks with millions of edges. Since data-intensive, computation-intensive and large network analysis challenges generally exist in most computational biology and neuroscience problems, such pipeline will be generally applicable and thus such effort will be appreciated.

REFERENCES

[1] P. S. Aisen, R. C. Petersen, M. C. Donohue, A. Gamst, R. Raman, R. G. Thomas, S. Walter, J. Q. Trojanowski, L. M. Shaw, L. A. Beckett, J. Jack, C. R., W. Jagust, A. W. Toga, A. J. Saykin, J. C. Morris, R. C. Green, and M. W. Weiner. Clinical core of the alzheimer's disease neuroimaging initiative: progress and plans. *Alzheimers Dement*, 6(3):239–46, 2010.

[2] J. Ashburner and K. J. Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11(6 Pt 1):805–21, 2000.

[3] B. B. Avants, P. A. Cook, L. Ungar, J. C. Gee, and M. Grossman. Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. *Neuroimage*, 50(3):1004–16, 2010.

[4] A. Bakkour, J. C. Morris, and B. C. Dickerson. The cortical signature of prodromal ad: regional thinning predicts mild ad dementia. *Neurology*, 72(12):1048–55, 2009.

[5] J. C. Barrett. Haploview: Visualization and analysis of snp genotype data. *Cold Spring Harb Protoc*, 2009(10):pdb ip71, 2009.

[6] J. C. Barrett. Haploview: Visualization and analysis of snp genotype data. *Cold Spring Harb Protoc*, 2009(10):pdb ip71, 2009.

[7] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–5, 2005.

[8] N. K. Batmanghelich, B. Taskar, and C. Davatzikos. Generative-discriminative basis learning for medical imaging. *IEEE Trans Med Imaging*, 31(1):51–69, 2012.

[9] M. Chadeau-Hyam, C. J. Hoggart, P. F. O'Reilly, J. C. Whittaker, M. De Iorio, and D. J. Balding. Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, 9:364, 2008.

[10] J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu, and H. Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2013.

[11] J. Chen et al. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2013.

[12] X. Chen and H. Liu. An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. *Statistics in Biosciences*, 4(1):3–26, 2012.

[13] X. Chen and H. Liu. An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. *Statistics in Bioscience*, 4(1):3–26, 2012.

[14] E. Chi, G. Allen, et al. Imaging genetics via sparse canonical correlation analysis. In *Biomedical Imaging (ISBI), 2013 IEEE 10th Int Sym on*, pages 740–743, 2013.

[15] E. C. Chi, G. I. Allen, H. Zhou, O. Kohannim, K. Lange, and P. M. Thompson. Imaging genetics via sparse canonical correlation analysis. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 740–743. IEEE.

[16] M. K. Chung, K. J. Worsley, B. M. Nacewicz, K. M. Dalton, and R. J. Davidson. General multivariate linear modeling of surface shapes using surfstat. *Neuroimage*, 53(2):491–505, 2010.

[17] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, and O. Colliot. Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *Neuroimage*, 56(2):766–781, 2011.

[18] A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis. i. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–94, 1999.

[19] C. De Mol, E. De Vito, and L. Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009.

[20] B. C. Dickerson, A. Bakkour, D. H. Salat, E. Feczko, J. Pacheco, D. N. Greve, F. Grodstein, C. I. Wright, D. Blacker, H. D. Rosas, R. A. Sperling, A. Atri, J. H. Growdon, B. T. Hyman, J. C. Morris, B. Fischl, and R. L. Buckner. The cortical signature of alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild ad dementia and is detectable in asymptomatic amyloid-positive individuals. *Cereb Cortex*, 19(3):497–510, 2009.

[21] D. Eddelbuettel. Cran task view: High-performance and parallel computing with r. Technical report, Version 2010-12-12, URL http://CRAN. R-project. org/view= HighPerformanceComputing, 2010.

[22] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[23] Y. El-Khamra, N. Gaffney, D. Walling, E. Wernert, W. Xu, and H. Zhang. Performance evaluation of r with intel xeon phi coprocessor. In *Big Data, 2013 IEEE International Conference on*, pages 23–30. IEEE, 2013.

[24] B. Fischl, M. I. Sereno, and A. M. Dale. Cortical surface-based analysis. ii: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999.

[25] C. Gaudes et al. Structured sparse deconvolution for paradigm free mapping of functional MRI data. *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 322–325, 2012.

[26] E. Gerardin, G. Chételat, M. Chupin, R. Cuingnet, B. Desgranges, H.-S. Kim, M. Niethammer, B. Dubois, S. Lehéricy, and L. Garnero. Multidimensional classification of hippocampal shape features discriminates alzheimer's disease and mild cognitive impairment from normal aging. *Neuroimage*, 47(4):1476–1486, 2009.

[27] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor. Interpretable whole-brain prediction analysis with graphnet. *Neuroimage*, 72:304–21, 2013.

[28] J. L. Gustafson and B. S. Greer. Clearspeed whitepaper: Accelerating the intel math kernel library, 2007.

[29] D. P. Hibar, O. Kohannim, et al. Multilocus genetic analysis of brain images. *Front Genet*, 2:73, 2011.

[30] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[31] C. J. Hoggart, M. Chadeau-Hyam, T. G. Clark, R. Lampariello, J. C. Whittaker, M. De Iorio, and D. J. Balding. Sequence-level population simulations over large genomic regions. *Genetics*, 177(3):1725–31, 2007.

[32] M. Intel. Intel math kernel library, 2007.

[33] W. J. Jagust, D. Bandy, et al. The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. *Alzheimers Dement*, 6(3):221–9, 2010.

[34] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion. Multiscale mining of fmri data with hierarchical structured sparsity. *SIAM Journal on Imaging Sciences*, 5(3):835–856, 2012.

[35] S. Kim et al. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25:i204–i212, 2009.

[36] S. Le Cessie and J. Van Houwelingen. Ridge estimators in logistic regression. *Applied statistics*, pages 191–201, 1992.

[37] M. N. Li and A. Rossini. Rpvm: Cluster statistical computing in r. *Porting R to Darwin/X11 and Mac OS X*, page 4, 2001.

[38] D. Lin, V. D. Calhoun, and Y.-P. Wang. Correspondence between fmri and snp data by group sparse canonical correlation analysis. *Medical image analysis*, 18(6):891–902, 2014.

[39] D. Lin et al. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Medical Image Analysis*, 18(6):891 – 902, 2014.

[40] D. Lin, J. Zhang, J. Li, V. D. Calhoun, H. W. Deng, and Y. P. Wang. Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, 14:245, 2013.

[41] D. Luo, C. Ding, and H. Huang. Towards structural sparsity: An explicit l2/l0 approach. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 344–353. IEEE.

[42] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. Total variation regularization for fmri-based prediction of behavior. *IEEE Trans Med Imaging*, 30(7):1328–40, 2011.

[43] F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint l2, 1-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821.

[44] G. Obozinski et al. Multi-task feature selection. *Technical Report, Technical report, Statistics Department, UC Berkeley*, 2006.

[45] E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8:1–34, 2009.

[46] K. Puniyani, S. Kim, and E. P. Xing. Multi-population gwa mapping via multi-task regularized regression. *Bioinformatics*, 26(12):i208–16, 2010.

[47] V. K. Ramanan, S. L. Risacher, et al. APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study. *Mol Psychiatry*, 19(3):351–7, 2014.

[48] C. Rebecca et al. Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer's disease diagnosis and prognosis. *Plos one*, 6(4):e18850, 2011.

[49] S. L. Risacher, A. J. Saykin, et al. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr Alzheimer Res*, 6(4):347–61, 2009.

[50] A. J. Rossini, L. Tierney, and N. Li. Simple parallel statistical computing in r. *Journal of Computational and Graphical Statistics*, 16(2):399–420, 2007.

[51] M. R. Sabuncu, J. L. Bernal-Rusiel, M. Reuter, D. N. Greve, and B. Fischl. Event time analysis of longitudinal neuroimage data. *Neuroimage*, 97:9–18, 2014.

[52] M. R. Sabuncu and K. Van Leemput. The relevance voxel machine (rvoxm): a self-tuning bayesian model for informative image-based prediction. *IEEE Trans Med Imaging*, 31(12):2290–306, 2012.

[53] 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., and M. W. Weiner. Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimers Dement*, 6(3):265–73, 2010.

[54] M. Schmidberger, M. Morgan, D. Eddelbuettel, H. Yu, L. Tierney, and U. Mansmann. State-of-the-art in parallel computing with r. *Journal of Statistical Software*, 47(1), 2009.

[55] L. Shen, S. Kim, 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*, 53(3):1051–63, 2010.

[56] J. Sheng, S. Kim, et al. Data synthesis and method evaluation for brain imaging genetics. In *Biomedical Imaging (ISBI), IEEE Int Sym on*, pages 1202–05, 2014.

[57] P. D. Solea et al. Possible relationship between Al/ferritin complex and Alzheimer's disease. *Clinical Biochemistry*, 46(1-2):89–93, 2013.

[58] C. M. Stonnington, C. Chu, S. Kloppel, J. Jack, C. R., J. Ashburner, and R. S. Frackowiak. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *Neuroimage*, 51(4):1405–13, 2010.

[59] S. Swaminathan, L. Shen, et al. Amyloid pathway-based candidate gene analysis of [(11)C]PiB-PET in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. *Brain Imaging Behav*, 6(1):1–15, 2012.

[60] W. Thies and L. Bleiler. 2013 alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 9(2):208–245, 2013.

[61] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser.*, B 58(267-288), 1996.

[62] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[63] G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In *ICML 2012*.

[64] M. Vounou, T. E. Nichols, and G. Montana. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage*, 53(3):1147–59, 2010.

[65] K. Walhovd et al. Multi-modal imaging predicts memory performance in normal aging and cognitive decline. *neurobiology of Aging*, 2010.

[66] K. B. Walhovd, A. M. Fjell, A. M. Dale, L. K. McEvoy, J. Brewer, D. S. Karow, D. P. Salmon, and C. Fennema-Notestine. Multi-modal imaging predicts memory performance in normal aging and cognitive decline. *Neurobiol Aging*, 31(7):1107–21, 2010.

[67] J. Wan, S. Kim, M. Inlow, K. Nho, S. Swaminathan, S. L. Risacher, S. Fang, M. W. Weiner, M. F. Beg, and L. Wang. *Hippocampal surface mapping of genetic risk factors in AD via sparse learning models*, pages 376–383. Springer, 2011.

[68] J. Wan, Z. Zhang, B. Rao, S. Fang, J. Yan, A. Saykin, and L. Shen. Identifying the neuroanatomical basis of cognitive impairment in alzheimer's disease by correlation- and nonlinearity-aware sparse bayesian learning. *IEEE Trans Med Imaging*, 2014.

[69] J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen. Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 940–947. IEEE.

[70] H. Wang et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics*, 28(2):229–37, 2012.

[71] H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. Risacher, A. Saykin, and L. Shen. Identifying quantitative trait loci via group-sparse multi-task regression and feature selection: An imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237, 2012.

[72] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A. J. Saykin, and L. Shen. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 557–562. IEEE.

[73] H. Wang, F. Nie, H. Huang, S. Risacher, A. J. Saykin, and L. Shen. *Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression*, pages 115–123. Springer, 2011.

[74] M. Weiner et al. The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimer's dimentia*, pages 1–68, 2012.

[75] M. W. Weiner, P. S. Aisen, J. Jack, C. R., W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett, A. Toga, R. Green, S. Walter, H. Soares, P. Snyder, E. Siemers, W. Potter, P. E. Cole, and M. Schmidt. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimers Dement*, 6(3):202–11 e7, 2010.

[76] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[77] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–34, 2009.

[78] D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*, 8:Article28, 2009.

[79] J. Yan et al. Multimodal neuroimaging predictors for cognitive performance using structured sparse learning. *MBIA'12, LNCS*, 7509:1–17, 2012.

[80] J. Yan, H. Huang, S. L. Risacher, S. Kim, M. Inlow, J. H. Moore, A. J. Saykin, and L. Shen. *Network-Guided Sparse Learning for Predicting Cognitive Outcomes from MRI Measures*, pages 202–210. Springer, 2013.

[81] H. Yu. Rmpi: Parallel statistical computing in r. *R News*, 2(2):10–14, 2002.

[82] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser.*, B 68:49–67, 2006.

[83] H. Zeng et al. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell*, 149(2):483–96, 2012.

[84] D. Zhang, D. Shen, and A. D. N. Initiative. Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. *PLoS One*, 7(3):e33182, 2012.

[85] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen. Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.

[86] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(2):301–320, 2005.

CURRICULUM VITAE

Jingwen Yan

Education

2015 Ph.D in Bioinformatics Indiana University

2009 M.S. in Computer Science Huazhong University of Science and Technology

2007 B.S. in Computer Science Nanjing University of Aeronautics & Astronautics

Honors, Awards, Fellowships

Travel fellowship, European Conference on Computational Biology, France, 2014

Travel fellowship, Indiana University, 2014-2015

Travel fellowship, Indiana University, 2013-2014

Travel fellowship, Alzheimer's Association International Conference, Boston, 2013

Travel fellowship, Alzheimer's Imaging Consortium, Boston, 2013

Graduate Assistantship, Indiana University, 2011-2013

Fellowship, Huazhong University of Science and Technology, China, 2007-2009

Outstanding Graduate Student Awards & Bachelor Dissertation Award, China, 2007

Silver in Jiangsu Mathematics Competition, China, 2004

National Scholarship, China, 2004 (0.01%)

Research and Training Experience

Graduate Research Assistant Indiana university School of Medicine 2011-2015

Machine learning and data mining for disease biomarker discovery

Genetic epistasis effect on multimodal imaging phenotypes

Brain network analysis and functional module identification in brain disorders

Big data frameworks in advancing the high-throughput computation

Visualization analytics of brain and genetic data

Publications

Journal

1. Yan J, Kim S, Nho K, Chen R, Risacher SL, Moore JH, Saykin AJ, Shen L, for the ADNI(2015) Hippocampal transcriptome-guided genetic analysis of correlated episodic memory phenotypes in Alzheimer's disease. Front. Genet.

2. Li J, Zhang Q, Chen F, Yan J, Kim S, Wang L, Feng W, Saykin AJ, Liang H, and Shen L (2015) Genetic interactions explain variance in cingulate amyloid burden: An AV-45 PET genome-wide association and interaction study in the ADNI cohort. BioMed Research International, Article ID 647389.

3. Yan J, Li T, Wang H, Huang H, Wan J, Nho K, Kim S, Risacher LS, Saykin AJ, Shen L, for the ADNI (2014). Cortical surface biomarkers for predicting cognitive outcomes using group L2,1 norm, Neurobiology of Aging.

4. Yan J, Du L, Kim S, Risacher SL, Huang H, Moore JH, Saykin AJ, Shen L, for the ADNI (2014) Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics.

5. Wan J, Zhang Z, Rao BD, Fang S, Yan J, Saykin AJ, Shen L, for the ADNI (2014) Identifying the neuroanatomical basis of cognitive impairment in Alzheimer's disease by correlation- and nonlinearity-aware sparse Bayesian learning. IEEE Trans. on Medical Imaging, 33(7):1476-1488.

6.Li T, Xie Z, Wu J, Yan J, Shen L (2013) Interactive object extraction by merging regions with k-global maximal similarity. Neurocomputing, 120: 610-623.

Conference

1. Yan J, Du L, Kim S, Risacher SL, Huang H, Inlow M, Moore JH, Saykin AJ, Shen L, for the ADNI. (2015) BoSCCA: Mining stable imaging and genetic associations

with implicit structure learning. MICGen 2015: MICCAI Workshop on Imaging Genetics, October 9, 2015.

2. Yan J, Huang H, Kim S, Moore JH, Saykin AJ, Shen L, for the ADNI (2014) Joint identification of imaging and proteomics biomarkers of Alzheimer's disease using network-guided sparse learning.ISBI'14: IEEE Int. Sym. on Biomedical Imaging, pp 665-668, Beijing, China, 28 April - 2 May, 2014.

3. Yan J, Zhang H, Du L, Wernert E, Saykin AJ and Shen L (2014) Accelerating sparse canonical correlation analysis for large brain imaging genetics data. XSEDE'14: The Annual Extreme Science and Engineering Discovery Environment Conference, Article No. 4, Atlanta, GA, July 13-18, 2014.

4. Du L*, Yan J *, Kim S, Risacher SL, Huang H, Inlow M, Moore JH, Saykin AJ, Shen L, for the ADNI (2014) A novel structure-aware sparse learning algorithm for brain imaging genetics. MICCAI'14: Med Image Comput Comput Assist Interv, Lecture Notes in Computer Science, 8675:329-336, Boston, MA, September 14-18, 2014. (*equal contribution)

5. Sheng J, Kim S, Yan J, Moore JH, Saykin AJ, Shen L, for the ADNI (2014) Data synthesis and method evaluation for brain imaging genetics. ISBI'14: IEEE Int. Sym. on Biomedical Imaging, pp 1202-1205, Beijing, China, 28 April - 2 May, 2014.

6. Huang H, Yan J, Nie F, Huang J, Cai W, Saykin AJ, Shen L (2013) A new sparse simplex model for brain anatomical and genetic network analysis. MICCAI'13: Med Image Comput Comput Assist Interv, Lecture Notes in Computer Science, 8150:625-632, Nagoya, Japan, September 22-26, 2013.

7. Yan J, Huang H, Risacher SL, Kim S, Inlow M, Moore JH, Saykin AJ, Shen L, for the ADNI (2013) Network-guided sparse learning for predicting cognitive outcomes

from MRI measures. MBIA'13: MICCAI Workshop on Multimodal Brain Image Analysis, Lecture Notes in Computer Science, 8159:202-210, Nagoya, Japan, September 22, 2013.

8. Wan J, Zhang Z, Yan J, Li T, Rao B, Fang S, Kim S, Risacher S, Saykin A, Shen L, for the ADNI (2012) Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease. CVPR'12: IEEE Int. Conf. on Computer Vision and Pattern Recognition, 940-947, Providence, Rhode Island, June 18-20, 2012.

9. Yan J, Risacher SL, Kim S, Simon JC, Li T, Wan J, Wang H, Huang H, Saykin AJ, Shen L, for the ADNI (2012) Multimodal neuroimaging predictors for cognitive performance using structured sparse learning. MBIA'12: MICCAI Workshop on Multimodal Brain Image Analysis, Nice, France, 2012.

10. Li T, Wan J, Zhang Z, Yan J, Kim S, Risacher SL, Fang S, Beg MF, Wang L, Saykin AJ, Shen L, for the ADNI (2012) Hippocampus as a predictor of cognitive performance: Comparative evaluation of analytical methods and morphometric measures. NIBioAD'12: MICCAI Workshop on Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders, Nice, France, October 5, 2012.

11. Yan J, Li T, Wang H, Huang H, Wan J, Nho K, Kim S, Risacher SL, Saykin AJ, Shen L, for the ADNI (2012) Identification of novel cortical surface biomarkers for predicting cognitive outcomes based on group-level L-21 norm. NIBioAD'12: MICCAI Workshop on Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders, Nice, France, October 5, 2012.